Signal attenuation enables scalable decentralized multi-agent reinforcement learning over networks

Wesley A. Suttle
Army Research Laboratory
U.S. Army DEVCOM
Adelphi, MD USA
wesley.a.suttle.ctr@army.mil

Vipul K. Sharma
Industrial Engineering Dept.
Purdue University
West Lafayette, IN USA
sharm697@purdue.edu

Brian M. Sadler
Oden Institute
University of Texas, Austin
Austin, TX USA
brian.sadler@ieee.org

***Abstract*—Multi-agent reinforcement learning (MARL) methods typically require that agents enjoy global state observability, preventing development of decentralized algorithms and limiting scalability. Recent work has shown that, under assumptions on decaying inter-agent influence, global observability can be replaced by local neighborhood observability at each agent, enabling decentralization and scalability. Real-world applications enjoying such decay properties remain underexplored, however, despite the fact that signal power decay, or signal attenuation, due to path loss is an intrinsic feature of many problems in wireless communications and radar networks. In this paper, we show that signal attenuation enables decentralization in MARL by considering the illustrative special case of performing power allocation for target detection in a radar network. To achieve this, we propose two new constrained multi-agent Markov decision process formulations of this power allocation problem, derive local neighborhood approximations for global value function and policy gradient estimates and establish corresponding error bounds, and develop decentralized saddle point policy gradient algorithms for solving the proposed problems. Our approach, though oriented towards the specific radar network problem we consider, provides a useful model for extensions to additional problems in wireless communications and radar networks.**

***Index Terms*—multi-agent reinforcement learning, radar networks, target detection, power allocation**

I. Introduction

Multi-agent reinforcement learning (MARL) has seen immense attention in recent years, from both theoretical [1]–[4] and experimental [5]–[7] perspectives. Due to limitations of the underlying multi-agent Markov decision process (MDP) model, however, standard methods for MARL in networked systems require global state observability at each agent [4]. This inhibits the development of truly decentralized MARL methods where each agent only needs access to information from its local neighborhood, thereby limiting the scalability of such methods and preventing their application to realistic problems. Fortunately, recent works [8]–[11] on scalable, decentralized MARL have shown that, for problems where inter-agent influence decays sufficiently quickly as the distance between agents increases over their communication network, use of only local neighborhood information at each agent suffices to approximately solve the global problem. This enables the development of truly decentralized methods that scale well as the number of agents increases. However, despite the theoretical advantages of these methods, real-world applications where the prerequisite decay properties hold remain largely unexplored.

Signal power decay, or signal attenuation, due to path loss is a well-known property of wireless communications [12] and radar systems [13]. In problems where multiple agents are widely dispersed over a geographic region, such as radar networks [14], path loss naturally leads to decay of inter-agent influence as distance between agents increases. For example, when the performance metric at each agent is a function of the power of received signals, such as the signal-to-interference-plus-noise ratio (SINR), performance measurements at a given agent are largely decoupled from behavior of other agents that are sufficiently far away over the network. Due to this inherent property of these systems, wireless communications and radar networks provide promising candidates for real-world application of scalable, decentralized MARL methods like [8]–[11] that rely on such decay properties for success.

In this paper, we examine the implications of signal attenuation for the development of scalable, decentralized MARL approaches to the specific problem of performing power allocation for target detection in a radar network. Radar networks are attractive for performing target detection and tracking due to advantages arising from their spatial dispersion and potential signal variety [15]. When determining power allocations in radar networks, maximizing power leads to improved signal strength, but this conflicts with the need to achieve low probability of intercept (LPI) and abide by resource constraints [16]. Existing methods for LPI power allocation for target detection in radar networks are centralized in that they require global observability and global coordination between radars [15]–[17], rendering them impractical in large networks for similar reasons to the MARL methods discussed above.

In this work, we propose a MARL approach for performing decentralized power allocation for LPI target detection in radar networks that mitigates these drawbacks. This is achieved by leveraging the signal attenuation inherent in radar networks to replace global observability and coordination with local observability and coordination among neighboring radars. Specifically, our contributions are as follows: (i) we propose two new constrained multi-agent MDP formulations of the problem of power allocation for target detection in radar networks; (ii) we leverage signal attenuation properties inherent in our setting to derive local approximations of the policy gradient expressions used in our algorithms and rigorously establish error bounds on these approximations; (iii) we propose novel decentralized, policy gradient ascent-descent algorithms for approximately solving the proposed problems. Though we focus on radar networks in this work, our approach can likely be extended to a broad range of applications in wireless communication and radar networks.

II. Problem Formulation

In this section, we first describe our system model for a radar network with widely separated radars, a flying target, and extended clutter. To enable LPI target detection and tracking in this setting, we subsequently propose two different problem formulations: (i) maximizing sum-of-SINRs subject to regional power constraints; (ii) minimizing power consumption subject to a minimal SINR threshold and regional power constraints.

A. Radar network model

We consider a radar network composed of a set $\mathcal{N} = \{1, \ldots, n\}$ networked radars. In the presence of a target, the received signal for radar $i \in \mathcal{N}$ is given by [15], [16]

$$x^i = \alpha^i \sqrt{a^i} y^i + \sum_{i \in \mathcal{N} \setminus \{i\}} \beta^{ji} \sqrt{a^j} y^j + \omega^i, \tag{1}$$

where $y^i = \psi^i z^i$ describes the transmitted signal from radar i and $z^i = \begin{bmatrix} 1 & e^{j2\pi f_{D,i}} & \ldots & e^{j2\pi(N-1)f_{D,i}} \end{bmatrix}^T$ is the Doppler steering vector of radar i associated with the desired target, $f_{D,i}$ denotes the normalized Doppler shift as seen by radar i, N is the number of received pulses at each timestep, and ψ^i denotes the predesigned waveform transmitted from radar i. Furthermore, the parameter α^i denotes the desired channel gain in the target direction, a^i denotes the transmission power of radar i, β^{ji} describes the cross-channel gain between radars i and j, and ω^i denotes zero-mean white Gaussian noise.

We assume that $\alpha^i \sim CN(0, h^\tau_{ii})$, $\beta^{ij} \sim CN(0, c_{ij}(h^\tau_{ij} + h^d_{ij}))$, and $\omega^i \sim CN(0, (\sigma^i)^2)$ for an i-dependent $\sigma^i > 0$, where $c_{ij} h^\tau_{ij}$ represents the variance of the channel gain for the radar i-target-radar j path, $c_{ij} h^d_{ij}$ represents the variance of the channel gain for the direct radar i-radar j path, and c_{ij} denotes the cross-correlation coefficient between the ith and jth radars. Note that $c_{ii} = 1$, so the variance of the channel gain for the radar i-target-radar i path is simply h^τ_{ii}. The variances are given by the radar range equations [13]

$$h^\tau_{ij} = \frac{G_t G_r \sigma^{RCS}_{ij} \lambda^2}{(4\pi)^3 R_i^2 R_j^2}, \quad (2) \qquad\qquad h^d_{ij} = \frac{G'_t G'_r \lambda^2}{(4\pi)^2 d^2_{ij}}, \quad (3)$$

where G_t and G_r are the radar main-lobe transmitting and receiving gains, G'_t and G'_r are the radar side-lobe transmitting and receiving gains, σ^{RCS}_{ij} is the radar cross section (RCS) of the target between the ith and jth radars, λ denotes the wavelength, R_i denotes the distance between radar i and the target, and d_{ij} denotes the distance between radars i and j. The signal-to-interference-plus-noise ratio (SINR) of the signal received at radar i is then given by

$$SINR_i = \frac{h^\tau_{ii} a^i}{(\sigma^i)^2 + \sum_{j \in \mathcal{N} \setminus \{i\}} c_{ji} \left(h^d_{ji} + h^\tau_{ji} \right) a^j}, \tag{4}$$

where σ^i corresponds to the noise received at radar i. See Figure 1 for an illustration of the radar network system model.

B. Constrained multi-agent MDP formulations

To capture the underlying system model described above, we use a constrained multi-agent Markov decision process (CMAMDP) $(\mathcal{S}, \mathcal{A}, p, \mathcal{N}, \{r^i\}_{i\in\mathcal{N}}, \{c^i\}_{i\in\mathcal{N}})$, defined below. Let the set of radars $\mathcal{N} = \{1, \ldots, n\}$ correspond to the n radars. Let $\mathcal{S} = \mathcal{S}^1 \times \ldots \times \mathcal{S}^n$ denote the joint state space, where the ith component $\mathcal{S}^i \subset \mathbb{R}^k \times \mathbb{R}^m$ corresponds to set of possible locations and movements of the target and radar i, i.e., $s^i = (s^i_{target}, s^i_{radar})$ with $s^i_{target} \in \mathbb{R}^k$ and $s^i_{radar} \in \mathbb{R}^m$. When the target is moving in $\mathbb{R}^3$ and the radars are moving in $\mathbb{R}^2$, for example, we may take $k = 9$ and $m = 6$ and let s^i_{target} and s^i_{radar} correspond to the positions, velocities, and accelerations of the target and radar, respectively. Notice that, for all $i, j \in \mathcal{N}$, we have $s^i_{target} = s^j_{target}$. Given the joint state $s_t \in \mathcal{S}$ at time t, we assume that the Doppler steering vectors, predesigned waveforms, and channel gain variances corresponding to radar i are provided and that the goal of radar i is to determine an appropriate transmission power level to supply given its local state information s^i_t. To capture this, let $\mathcal{A} = \mathcal{A}^1 \times \ldots \times \mathcal{A}^n$, where $\mathcal{A}^i = [0, a^{max}]$ denotes the set of possible power allocations at radar i, for a predetermined, finite, maximum power allocation $a^{max} > 0$ over all radars.

Let the transition dynamics $p : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ capture the movement of the target and radars over time. We assume in this paper that the movements of the target and radars, and therefore p, are independent of the transmission power allocations applied at the radars. Given joint state $s \in \mathcal{S}$ and joint power allocation $a \in \mathcal{S}$, let $r^i(s, a) = SINR_i(s, a)$ denote the SINR received at radar i obtained from equation (4) by substituting the channel gains, cross-correlation coefficients, and noise corresponding to s and applying allocation a, i.e.,

$$r^i(s, a) = \frac{h^\tau_{ii}(s) a^i}{(\sigma^i(s, a))^2 + \sum_{j \in \mathcal{N} \setminus \{i\}} c_{ji}(s) \left(h^d_{ji}(s) + h^\tau_{ji}(s) \right) a^j}, \tag{5}$$

where we abuse notation to let $h^\tau_{ii}(s), h^\tau_{ji}(s), h^d_{ji}(s), c_{ji}(s), \sigma^i(s, a)$, and $\sigma^i_\kappa(s, a)$ correspond to the expressions appearing in (4) when the system is in state s and joint action a is selected. Similarly, define the local neighborhood reward by

$$r^i_\kappa(s, a) = \frac{h^\tau_{ii}(s) a^i}{(\sigma^i_\kappa(s, a))^2 + \sum_{j \in \mathcal{N}_\kappa(i) \setminus \{i\}} c_{ji}(s) \left(h^d_{ji}(s) + h^\tau_{ji}(s) \right) a^j}, \tag{6}$$

which captures the SINR received at radar i originating within neighborhood $\mathcal{N}_\kappa(i)$, where $\sigma^i_\kappa(s, a)$ denotes the noise originating within $\mathcal{N}_\kappa(i)$. The expression in (6) will be crucial in the theoretical results of Section III. Finally, let the cost $c^i(s, a)$ denote the cost to radar i of applying power level a^i. We might simply take $c^i(s, a) = a^i$, for example, but our approach accommodates general cost structures.

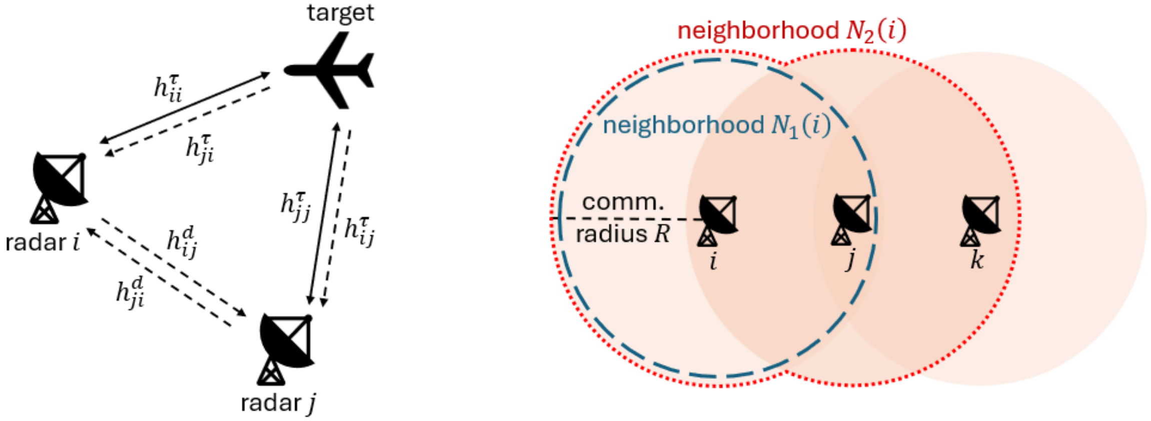

Fig. 1. Illustrations of radar network system model of Section II-A (left) and communication neighborhoods of Section II-B (right).

Let a fixed communication radius $R > 0$ between the radars be given, and denote the undirected communication graph between radars in state $s \in \mathcal{S}$ by $\mathcal{G}(s) = (\mathcal{N}, \mathcal{E}(s))$, where $\mathcal{E}(s) = \{(i,j) \mid d_{ij}(s) \leq R\}$ and $d_{ij}(s)$ denotes the Euclidean distance between radars i and j when the system is in state s. We note that R defines a user-specified communications neighborhood, but that it does not necessarily represent a hard limit on communications within the network. We assume in this paper that the edge set $\mathcal{E} = \mathcal{E}(s)$ remains constant, i.e., that any movement of the radars leaves the topology of the communication network unchanged, and will henceforth suppress the dependence on s and simply write $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. For a given positive integer $\kappa \in \mathbb{N}^+$, let $\mathcal{N}_\kappa(i)$ denote the κ-hop neighborhood of radar i with respect to $\mathcal{G}$. Note that, since $\mathcal{G}$ is undirected, the κ-hop neighborhood relation is symmetric, i.e., $i \in \mathcal{N}_\kappa(j)$ if and only if $j \in \mathcal{N}_\kappa(i)$. Finally, let $\mathcal{N}_\kappa^{-1}(i) = \mathcal{N} \setminus \mathcal{N}_\kappa(i)$ denote the set of all radars outside i's κ-hop neighborhood. See Figure 1 for an illustration of the communication neighborhoods induced by a specified communication radius R on a simple example.

Let $\kappa \in \mathbb{N}^+$ be fixed, and assume each radar i has access to the state information $s^{\mathcal{N}_\kappa(i)} \in \mathcal{S}^{\mathcal{N}_\kappa(i)} = \{s^j \mid j \in \mathcal{N}_\kappa(i)\}$. To each i, let there be associated a parameterized policy class $\{\pi^i_{\theta_i} : \mathcal{S}^{\mathcal{N}_\kappa(i)} \to \Delta(\mathcal{A}^i)\}_{\theta^i \in \Theta^i}$, where $\Theta^i \subseteq \mathbb{R}^d$ is the set of permissible policy parameters, for some positive integer d. Denote the induced joint policy by $\pi_\theta(a|s) = \prod_{i \in \mathcal{N}} \pi^i_{\theta^i}(a^i | s^{\mathcal{N}_\kappa(i)})$, where $\theta = \left[(\theta^1)^T \dots (\theta^n)^T\right]^T \in \Theta = \Theta^1 \times \dots \times \Theta^n$ is the stacked vector of each radar's policy parameters. Given policy π_θ, under this formulation,

$$J_{r^i}(\theta) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{T-1} r^i(s_t, a_t)\Big], \tag{7}$$

$$J_{c^i}(\theta) = \lim_{T \to \infty} \frac{1}{T} \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{T-1} c^i(s_t, a_t)\Big] \tag{8}$$

capture the average expected SINR achieved and average expected cost incurred, respectively, at radar i.

1) Sum-of-SINRs Maximization: Fix $\kappa > 0$. The sum-of-SINRs maximization problem is as follows.

$$\begin{aligned} \max_{\theta \in \Theta} \quad & \sum_{i \in \mathcal{N}} J_{r^i}(\theta) \\ \text{s.t.} \quad & \sum_{j \in \mathcal{N}_\kappa(i)} J_{c^j}(\theta) \leq u^i, \quad \forall i \in \mathcal{N}, \end{aligned} \tag{P_κ^{max}}$$

The objective of this problem is to maximize the sum over the expected average SINRs over all radars while simultaneously ensuring that expected average cost does not exceed "regional" upper bounds at each radar.

2) Power Minimization with SINR Threshold: Fix $\kappa > 0$ and let $\gamma_{min} > 0$ denote the minimum SINR value allowable at each radar, where we assume that γ_{min} is chosen such that the thresholds at each radar in problem (P_κ^{min}) below are achievable. The problem of minimizing overall power consumption while respecting the SINR threshold is

$$\begin{aligned} \min_{\theta \in \Theta} \quad & \sum_{i \in \mathcal{N}} J_{c^i}(\theta) \\ \text{s.t.} \quad & J_{r^i}(\theta) \geq \gamma_{\min}, \quad \forall i \in \mathcal{N}, \\ & \sum_{j \in \mathcal{N}_\kappa(i)} J_{c^j}(\theta) \leq u^i, \quad \forall i \in \mathcal{N}. \end{aligned} \tag{P_κ^{min}}$$

The objective of this problem is to minimize the expected average cost over all radars while simultaneously ensuring that both (a) each radar achieves a minimal SINR value, on average, and (b) expected average cost does not exceed "regional" upper bounds at each radar.

III. Signal Attenuation Enables Decentralization

In this section, we examine how signal decay inherent in the radar problem under consideration leads to natural decentralization in MARL solution approaches. We first recall the MARL global observability issue that prevents true decentralization in the general setting, then establish formal properties of our problem that lead to natural decentralization. We leverage the properties established in this section to develop decentralized MARL algorithms for solving problems (P_κ^{max}) and (P_κ^{min}) in the next section.

We first review a few key concepts from RL. Let $r(s,a) = \sum_{i\in\mathcal{N}} r^i(s,a)$ denote the global reward and $c(s,a) = \sum_{i\in\mathcal{N}} c^i(s,a)$ the global cost. Define the global averages

$$J_r(\theta) = \sum_{i\in\mathcal{N}} J_{r^i}(\theta), \qquad J_c(\theta) = \sum_{i\in\mathcal{N}} J_{c^i}(\theta). \tag{9}$$

Fixing $\theta \in \Theta$, the action-value function corresponding to π_θ and the reward or cost function $f \in \{r, c\}$ is given by

$$\begin{aligned} Q_\theta^f(s,a) &= \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} f(s_t,a_t) - J_f(\theta) \mid s_0 = s, a_0 = a\right] && (10)\\ &= \sum_{i\in\mathcal{N}} \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} f^i(s,a) - J_{f^i}(\theta) \mid s_0 = s, a_0 = a\right] && (11)\\ &= \sum_{i\in\mathcal{N}} Q_\theta^{f^i}(s,a), && (12)\end{aligned}$$

where the Q function $Q_\theta^{f^i}(s,a)$ corresponding to f^i is

$$Q_\theta^{f^i}(s,a) = \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} f^i(s,a) - J_{f^i}(\theta) \mid s_0 = s, a_0 = a\right], \tag{13}$$

for the ith reward or cost $f^i \in \{r^i, c^i\}$.

A. MARL Global Observability Issue

Ideally, we would like each agent i to be able to compute its contribution $\nabla_{\theta^i} J_f(\theta)$ to the overall gradient $\nabla_\theta J_f(\theta)$ using only information from its local neighborhood $\mathcal{N}_\kappa(i)$, as this would make the development of decentralized algorithms easier. However, applying the classic policy gradient theorem [18] to differentiate with respect to θ^i gives

$$\begin{aligned} \nabla_{\theta^i} J_f(\theta) &= \mathbb{E}_{\pi_\theta}\left[Q_\theta^f(s,a)\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right] && (14)\\ &= \mathbb{E}_{\pi_\theta}\left[\sum_{j\in\mathcal{N}} Q_\theta^{f^j}(s,a)\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right]. && (15)\end{aligned}$$

From equation (15), global information is clearly required to estimate $\nabla_{\theta^i} J_f(\theta)$, since each of the $Q_\theta^{f^j}(s,a)$ requires access to the global state s and joint action a, and agent i needs access to the Q functions of all other agents $j \in \mathcal{N}$. This is the crux of the global observability issue in policy gradient methods for MARL. To address this issue, we next identify properties of the radar problem we consider under which the term $\sum_{j\in\mathcal{N}} Q_\theta^{f^j}(s,a)$ can be replaced by an approximation depending only on the information available to agent i within its local neighborhood, $\mathcal{N}_\kappa(i)$.

B. Signal Attenuation and Decentralization

We start this section by stating several assumptions that will be needed in the subsequent analysis. Assumption 1 provides a mechanism for ensuring that the radar network provides adequate coverage of the region under consideration through appropriate choice of g. Assumption 2 stipulates that the local costs at each radar are independent of the costs at all other radars. This is reasonable when costs depend only on local power consumption, for example. Assumption 3 provides minimal conditions ensuring that the radar range equations (2)-(3) lead to well-conditioned local rewards (5). Assumptions 4 and 5 are commonly used in the RL [19]–[21] and MARL [8], [9] literatures to enable analysis of policy gradient-based methods. Assumption 6 bounds the effect that changes in a given radar's policy parameters can have on rewards received by radars outside its κ-hop neighborhood, which is reasonable when rewards decay as distance between agents increases. We suspect that Assumption 6 can be proven to hold for our setting, but leave this to future work.

Assumption 1 (Network Coverage)**.** *There exists a function $g : \mathbb{N}^+ \times \mathbb{R}^+ \to [1,\infty)$, strictly increasing in both entries, such that, for all $i \in \mathcal{N}$, all $\kappa \in \mathbb{N}^+$, and any $j \in \mathcal{N}_\kappa^{-1}(i)$, we have $d_{ij}(s) \geq g(\kappa, R)$, for all $s \in \mathcal{S}$.*

Remark 1. When $\kappa = 1$, taking $g(\kappa, R) = R$ is a natural choice in Assumption 1, since, for any $j \in \mathcal{N}_1^{-1}(i)$, we have $d_{ij}(s) \geq g(1,R) = R$ by definition of the neighbor relation in $\mathcal{G}$ (see Section II-B). More generally, letting $g(\kappa,R) = \kappa R$ satisfies Assumption 1 for a large class of network topologies, such as when the radars are static and arranged in a grid topology with grid cell side lengths R.

Assumption 2 (Pairwise Cost Independence)**.** *For each $i \in \mathcal{N}$, we have the following two conditions: (i) $\nabla_{\theta^i} J_{c^j}(\theta) = 0$, for all $j \in \mathcal{N}\setminus\{i\}$; (ii) $Q_\theta^{c^i}(s,a) = Q_\theta^{c^i}(s^i,a^i)$, i.e., the value of Q^{c^i} depends on purely local state and action information.*

Assumption 3 (Regularity Conditions)**.** *For all $i,j \in \mathcal{N}$, we have $\inf_{s\in\mathcal{S}} R_i(s) \geq 1$, $\bar\sigma^{RCS} = \sup_{s\in\mathcal{S}} \sigma_{ij}^{RCS}(s) < \infty$ and $\bar\sigma^{RCS} > 0$, $\bar c = \sup_{s\in\mathcal{S}} c_{ji}(s) < \infty$ and $\bar c > 0$, $\underline{\sigma} = \min\{\inf_{s,a}\sigma^i(s,a), \inf_{s,a}\sigma^i_\kappa(s,a)\} > 0$, and $\lambda, G_t, G_r, G'_t, G'_r > 0$.*

Assumption 4 (Uniform Ergodicity)**.** *There exist $\rho \in (0,1)$ and $m \in \mathbb{R}^+$ such that every joint policy π_θ satisfies $d_{TV}(d^t_{\pi_\theta}(\cdot|s_0)||d_\theta(\cdot)) \leq m\rho^t$, for any $s_0 \in \mathcal{S}$ and for all $t \geq 0$, where $d_{TV}(q(\cdot)||q'(\cdot)) = \sup_A|\int_A q(x)\,dx - \int_A q'(x)\,dx|$ denotes the total variation distance between densities q, q', and $d^t_{\pi_\theta}(\cdot|s_0)$ denotes the t-step state occupancy measure induced by π_θ over $\mathcal{S}$ given start state s_0.*

Assumption 5 (Lipschitz Score Functions)**.** *For each $i \in \mathcal{N}$, there exists $L^i > 0$ such that $\left\|\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right\| \leq L^i$, for all $s^i \in \mathcal{S}^i, a^i \in \mathcal{A}^i$.*

Assumption 6 (Bounded Inter-agent Gradients)**.** *There exists $\varepsilon_\kappa > 0$ such that, for each $i \in \mathcal{N}$ and all $j \in \mathcal{N}_\kappa^{-1}(i)$, we have $\|\nabla_{\theta^i} J_{r^j}(\theta)\| \leq \varepsilon_\kappa$, for all $\theta \in \Theta$.*

We now establish properties of the radar problem enabling decentralized solution of the CMAMDP problem. Due to the form of the costs and rewards coupled with properties of the SINR (4) and radar range equations (2) and (3), the CMAMDP enjoys the following property.

Theorem 1. *Let Assumptions 1, 2, 3, and 4 hold. For any $\theta \in \Theta$, $i \in \mathcal{N}$, $s^{\mathcal{N}_\kappa(i)} \in \mathcal{S}^{\mathcal{N}_\kappa(i)}$, $a^{\mathcal{N}_\kappa(i)} \in \mathcal{A}^{\mathcal{N}_\kappa(i)}$, and $f^i \in$*

$\{r^i, c^i\}$, we have

$$\Big|Q_\theta^{f^i}\left((s^{\mathcal{N}_\kappa(i)}, s^{\mathcal{N}_\kappa^{-1}(i)}), (a^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa^{-1}(i)})\right) - Q_\theta^{f^i}\left((s^{\mathcal{N}_\kappa(i)}, \bar{s}^{\mathcal{N}_\kappa^{-1}(i)}), (a^{\mathcal{N}_\kappa(i)}, \bar{a}^{\mathcal{N}_\kappa^{-1}(i)})\right)\Big| \leq \frac{M|\mathcal{N}_\kappa^{-1}(i)|}{g^2(\kappa, R)},$$

for all $s^{\mathcal{N}_\kappa^{-1}(i)}, \bar{s}^{\mathcal{N}_\kappa^{-1}(i)} \in \mathcal{S}^{\mathcal{N}_\kappa^{-1}(i)}$ and all $a^{\mathcal{N}_\kappa^{-1}(i)}, \bar{a}^{\mathcal{N}_\kappa^{-1}(i)} \in \mathcal{A}^{\mathcal{N}_\kappa^{-1}(i)}$, where $M > 0$ is a constant depending only on the quantities in Assumptions 3 and 4.

Proof. Assumption 2 renders the $f^i = c^i$ case trivial, so we give the proof only for the $f^i = r^i$ case. Fix $\theta \in \Theta, i \in \mathcal{N}$, and $\kappa \in \mathbb{N}^+$. Let $s^{\mathcal{N}_\kappa(i)} \in \mathcal{S}^{\mathcal{N}_\kappa(i)}$, $a^{\mathcal{N}_\kappa(i)} \in \mathcal{A}^{\mathcal{N}_\kappa(i)}$, $s^{\mathcal{N}_\kappa^{-1}(i)}, \bar{s}^{\mathcal{N}_\kappa^{-1}(i)} \in \mathcal{S}^{\mathcal{N}_\kappa^{-1}(i)}$, and $a^{\mathcal{N}_\kappa^{-1}(i)}, \bar{a}^{\mathcal{N}_\kappa^{-1}(i)} \in \mathcal{A}^{\mathcal{N}_\kappa^{-1}(i)}$. Define $s = (s^{\mathcal{N}_\kappa(i)}, s^{\mathcal{N}_\kappa^{-1}(i)}), \bar{s} = (s^{\mathcal{N}_\kappa(i)}, \bar{s}^{\mathcal{N}_\kappa^{-1}(i)}), a = (a^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa^{-1}(i)})$, and $\bar{a} = (a^{\mathcal{N}_\kappa(i)}, \bar{a}^{\mathcal{N}_\kappa^{-1}(i)})$.

By the definition of $Q_\theta^{r^i}$, we have

$$\Big|Q_\theta^{r^i}(s,a) - Q_\theta^{r^i}(\bar{s},\bar{a})\Big| \tag{16}$$

$$= \Big|\mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{\infty} r^i(s_t, a_t) - J_{r^i}(\theta) \mid s_0 = s, a_0 = a\Big]$$
$$- \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{\infty} r^i(s_t, a_t) - J_{r^i}(\theta) \mid s_0 = \bar{s}, a_0 = \bar{a}\Big]\Big| \tag{17}$$

$$= \Big|\mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{\infty} r^i(s_t, a_t) \mid s_0 = s, a_0 = a\Big]$$
$$- \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{\infty} r^i(s_t, a_t) \mid s_0 = \bar{s}, a_0 = \bar{a}\Big]\Big|, \tag{18}$$

where the last equality holds by the fact that $\mathbb{E}_{\pi_\theta}\left[J_{r^i}(\theta) \mid s_0 = s, a_0 = a\right] = \mathbb{E}_{\pi_\theta}\left[J_{r^i}(\theta) \mid s_0 = \bar{s}, a_0 = \bar{a}\right]$, since Assumption 4 implies ergodicity of the Markov chain over $\mathcal{S}$ induced by π_θ. Recalling (6) and continuing from (18), we have

$$\Big|Q_\theta^{r^i}(s,a) - Q_\theta^{r^i}(\bar{s},\bar{s})\Big| \tag{19}$$

$$= \Big|\sum_{t=0}^{\infty}\Big[\mathbb{E}_{\pi_\theta}\left[r^i(s_t, a_t) - r_\kappa^i(s_t, a_t) + r_\kappa^i(s_t, a_t) \mid s_0 = s, a_0 = a\right]$$
$$- \mathbb{E}_{\pi_\theta}\left[r^i(s_t, a_t) \mid s_0 = \bar{s}, a_0 = \bar{a}\right]\Big]\Big| \tag{20}$$

$$= \Big|\sum_{t=0}^{\infty}\Big[\mathbb{E}_{\pi_\theta}\left[r^i(s_t, a_t) - r_\kappa^i(s_t, a_t) \mid s_0 = s, a_0 = a\right]$$
$$- \mathbb{E}_{\pi_\theta}\left[r^i(s_t, a_t) - r_\kappa^i(s_t, a_t) \mid s_0 = \bar{s}, a_0 = \bar{a}\right]\Big]\Big|, \tag{21}$$

where (21) follows by the independence of a of the transition dynamics p defined in Section II-B and the fact that $r_\kappa^i(s,a)$ is independent of the value of $s^{\mathcal{N}_\kappa^{-1}(i)}, a^{\mathcal{N}_\kappa^{-1}(i)}$ by definition.

Now notice that

$$\sum_{t=0}^{\infty}\mathbb{E}_{\pi_\theta}\left[r^i(s_t, a_t) - r_\kappa^i(s_t, a_t) \mid s_0 = s, a_0 = a\right] \tag{22}$$

$$= \sum_{t=0}^{\infty}\sum_{s' \in \mathcal{S}}\sum_{a' \in \mathcal{A}} G_\kappa^i(s', a')\pi_\theta(a'|s')d_{\pi_\theta}^t(s'|s_0 = s, a_0 = a), \tag{23}$$

where $d_{\pi_\theta}^t$ denotes the t-step state distribution of Assumption 4 and $G_\kappa^i(s', a') = r^i(s', a') - r_\kappa^i(s', a')$, and that an analogous expression holds for $s_0 = \bar{s}, a_0 = \bar{a}$. Continuing from (21) and (23), we have

$$\Big|Q_\theta^{r^i}(s,a) - Q_\theta^{r^i}(\bar{s},\bar{s})\Big| \tag{24}$$

$$= \Big|\sum_{t=0}^{\infty}\sum_{s' \in \mathcal{S}}\sum_{a' \in \mathcal{A}} G_\kappa^i(s', a')\pi_\theta(a'|s')\Big(d_{\pi_\theta}^t(s'|s_0 = s, a_0 = a)$$
$$- d_{\pi_\theta}^t(s'|s_0 = \bar{s}, a_0 = \bar{a})\Big)\Big| \tag{25}$$

$$\overset{\text{(a)}}{\leq} \sum_{t=0}^{\infty}\sum_{s' \in \mathcal{S}}\sum_{a' \in \mathcal{A}}\Big|G_\kappa^i(s', a')\Big| \cdot \Big|d_{\pi_\theta}^t(s'|s_0 = s, a_0 = a)$$
$$- d_{\pi_\theta}^t(s'|s_0 = \bar{s}, a_0 = \bar{a})\Big| \tag{26}$$

$$\overset{\text{(b)}}{\leq} \sum_{t=0}^{\infty}\sum_{s' \in \mathcal{S}}\sum_{a' \in \mathcal{A}}\Big|G_\kappa^i(s', a')\Big|\Big[\Big|d_{\pi_\theta}^t(s'|s_0 = s, a_0 = a) - d_{\pi_\theta}(s')\Big|$$
$$+ \Big|d_{\pi_\theta}^t(s'|s_0 = \bar{s}, a_0 = \bar{a}) - d_{\pi_\theta}(s')\Big|\Big] \tag{27}$$

$$\overset{\text{(c)}}{\leq} \sup_{s', a'}\Big|G_\kappa^i(s', a')\Big|\sum_{t=0}^{\infty}\Big[d_{TV}(d_{\pi_\theta}^t(\cdot|s_0 = s, a_0 = a) \mid\mid d_{\pi_\theta}(\cdot))$$
$$+ d_{TV}(d_{\pi_\theta}^t(\cdot|s_0 = \bar{s}, a_0 = \bar{a}) \mid\mid d_{\pi_\theta}(\cdot))\Big] \tag{28}$$

$$\overset{\text{(d)}}{\leq} \sup_{s', a'}\Big|G_\kappa^i(s', a')\Big|\sum_{t=0}^{\infty} 2m\rho^t = \sup_{s', a'}\Big|G_\kappa^i(s', a')\Big|\frac{2m}{1-\rho}, \tag{29}$$

where inequality (a) follows by the triangle and Cauchy-Schwarz inequalities and the fact that $0 \leq \pi_\theta(a'|s') \leq 1$, inequality (b) follows by adding and subtracting $d_\theta(s')$ and applying the triangle inequality, (c) follows by taking suprema and the definition of total variation distance, and (d) follows by Assumption 4.

All that remains is to bound $\sup_{s', a'}\big|G_\kappa^i(s', a')\big|$. Fix $s \in \mathcal{S}, a \in \mathcal{A}$. Recall that

$$r^i(s,a) = \frac{h_{ii}^\tau(s)a^i}{\sigma^2(s,a) + \alpha}, \quad r_\kappa^i(s,a) = \frac{h_{ii}^\tau(s)a^i}{(\sigma_\kappa(s,a))^2 + \beta}, \tag{30}$$

where

$$\alpha = \sum_{j \in \mathcal{N}\setminus\{i\}} c_{ji}(s)\left[h_{ji}^d(s) + h_{ji}^\tau(s)\right]a^j, \tag{31}$$

$$\beta = \sum_{j \in \mathcal{N}_\kappa(i)\setminus\{i\}} c_{ji}(s)\left[h_{ji}^d(s) + h_{ji}^\tau(s)\right]a^j. \tag{32}$$

In light of this, we can write

$$\big|G_\kappa^i(s,a)\big| = \left|\frac{h_{ii}^\tau(s)a^i(\alpha - \beta)}{(\sigma^2(s,a) + \alpha)((\sigma_\kappa^i(s,a))^2 + \beta)}\right| \tag{33}$$

$$\overset{\text{(a)}}{=} \frac{h_{ii}^\tau(s)a^i|\alpha - \beta|}{|\sigma^2(s,a) + \alpha||(\sigma_\kappa^i(s,a))^2 + \beta|} \tag{34}$$

$$\overset{\text{(b)}}{\leq} \frac{h_{ii}^\tau(s)a^i|\alpha - \beta|}{\underline{\sigma}^4}, \tag{35}$$

where (a) follows by nonnegativity of $h_{ii}^\tau(s)$ and a^i, and (b) follows by nonnegativity of α and β and Assumption 3. By (2) and Assumption 3,

$$h_{ii}^\tau(s)a^i = \frac{G_t G_r \sigma_{ii}^{RCS}(s)\lambda^2 a^i}{(4\pi)^3 R_i^4(s)} \leq \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2 a^{max}}{(4\pi)^3}, \tag{36}$$

whence it follows that

$$\left|G^i_\kappa(s,a)\right| \leq \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2 a^{max}}{(4\pi)^3 \underline{\sigma}^4}|\alpha - \beta|. \tag{37}$$

It remains to bound $|\alpha - \beta|$. Notice that

$$\alpha - \beta = \sum_{j \in \mathcal{N}_\kappa^{-1}(i)} c_{ji}(s)\left[h^d_{ji}(s) + h^\tau_{ji}(s)\right] a^j \tag{38}$$

$$\leq \bar{c} a^{max} \sum_{j \in \mathcal{N}_\kappa^{-1}(i)} \left[h^d_{ji}(s) + h^\tau_{ji}(s)\right] \tag{39}$$

so we need only bound $h^d_{ji}(s)$ and $h^\tau_{ji}(s)$. For $h^d_{ji}(s)$, notice that, for $j \in \mathcal{N}_\kappa^{-1}(i)$, by (3) and Assumption 1 we have

$$h^d_{ji}(s) = \frac{G'_t G'_r \lambda^2}{(4\pi)^2 (d_{ij}(s))^2} \leq \frac{G'_t G'_r \lambda^2}{(4\pi)^2 g^2(\kappa, R)}. \tag{40}$$

For $h^\tau_{ji}(s)$, we consider four possible cases that may arise depending on the physical configuration of the radar network. Fix $j \in \mathcal{N}_\kappa^{-1}(i)$.

Case 1: $R_j(s) \leq d_{ij}(s) \leq R_i(s)$. In this case, we know $R_i(s)R_j(s) \geq d_{ij}(s)$, whence

$$h^\tau_{ji}(s) = \frac{G_t G_r \sigma^{RCS}(s)\lambda^2}{(4\pi)^3 R_i^2(s) R_j^2(s)} \leq \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2}{(4\pi)^3 g^2(\kappa, R)}, \tag{41}$$

where the inequality holds by Assumptions 1 and 3.

Case 2: $R_i(s) \leq d_{ij}(s) \leq R_j(s)$. This case follows by reasoning identical to Case 1.

Case 3: $d_{ij}(s) \leq \min\{R_i(s), R_j(s)\}$. In this case, $R_i(s)R_j(s) \geq d^2_{ij}(s)$, whence

$$h^\tau_{ij}(s) \leq \frac{G_t G_r \bar{\sigma}^R CS\lambda^2}{(4\pi)^3 g^4(\kappa, R)} \overset{\text{(a)}}{\leq} \frac{G_t G_r \bar{\sigma}^R CS\lambda^2}{(4\pi)^3 g^2(\kappa, R)}, \tag{42}$$

where (a) holds since $g(\kappa, R) \geq 1$ by Assumption 1.

Case 4: $d_{ij}(s) \geq \max\{R_i(s), R_j(s)\}$. The triangle inequality gives $R_i(s) + R_j(s) \geq d_{ij}(s)$, whence $R_i(s) \geq g(\kappa, R) - R_j(s)$. If $R_j \geq \frac{1}{2}g(\kappa, R)$, then

$$h^\tau_{ij}(s) = \frac{G_t G_r \sigma^{RCS}(s)\lambda^2}{(4\pi)^3 R_i^2(s) R_j^2(s)} \leq \frac{G_t G_r \sigma^{RCS}(s)\lambda^2}{(4\pi)^3 R_j^2(s)} \tag{43}$$

$$\leq \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2}{(4\pi)^3 \frac{1}{4} g^2(\kappa, R)}. \tag{44}$$

On the other hand, if $R_j(s) < \frac{1}{2}g(\kappa, R)$, then $R_i(s) \geq g(\kappa, R) - R_j(s) > \frac{1}{2}g(\kappa, R)$. Thus, by the same argument as above,

$$h^\tau_{ji}(s) \leq \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2}{(4\pi)^3 \frac{1}{4} g^2(\kappa, R)}. \tag{45}$$

Taking the maximum over all four cases, we obtain

$$h^\tau_{ji}(s) \leq \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2}{4^2 \pi^3 g^2(\kappa, R)}. \tag{46}$$

Combining (39), (40), and (46), we have

$$|\alpha - \beta| \leq \bar{c} a^{max} \left|\mathcal{N}_\kappa^{-1}(i)\right| \left[\frac{G'_t G'_r \lambda^2}{(4\pi)^2 g^2(\kappa, R)} + \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2}{4^2 \pi^3 g^2(\kappa, R)}\right], \tag{47}$$

which, combined with (37), gives us

$$\left|G^i_\kappa(s,a)\right| \leq \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2 (a^{max})^2 \bar{c}}{(4\pi)^3} \cdot \left[\frac{G'_t G'_r \lambda^2}{(4\pi)^2} + \frac{G_t G_r \bar{\sigma}^{RCS}\lambda^2}{4^2 \pi^3}\right] \frac{\left|\mathcal{N}_\kappa^{-1}(i)\right|}{g^2(\kappa, R)}. \tag{48}$$

Finally, combining (29) with (48), we have

$$\left|Q^{r^i}_\theta(s,a) - Q^{r^i}_\theta(\bar{s}, \bar{s})\right| \leq \frac{M\left|\mathcal{N}_\kappa^{-1}(i)\right|}{g^2(\kappa, R)}, \tag{49}$$

where

$$M = \frac{2m G_t G_r \bar{\sigma}^{RCS}\lambda^4 (a^{max})^2 \bar{c}}{(1-\rho)(4\pi)^3} \left[\frac{G'_t G'_r}{(4\pi)^2} + \frac{G_t G_r \bar{\sigma}^{RCS}}{4^2 \pi^3}\right]. \tag{50}$$

□

In Theorem 1, the scalar M intuitively corresponds to the maximum possible expected contribution of radar j to $SINR_i$ before accounting for signal attenuation due to distance. When $\kappa = 1$ and $g(\kappa, R) = R$, for example, $M|\mathcal{N}_1^{-1}(i)|/R^2$ bounds the maximum possible contribution to $SINR_i$ originating outside radar i's immediate communication neighborhood $\mathcal{N}_1(i)$, decayed by the square of the communication radius R. Furthermore, for general g, as $\kappa > 1$ increases $\mathcal{N}_\kappa^{-1}(i)$ will decrease and $g(\kappa, R)$ will increase, resulting in a tighter overall bound $M|\mathcal{N}_\kappa^{-1}(i)|/g^2(\kappa, R)$. This bound can be further tightened by increasing the communication radius R.

We now proceed to the main result of this section, which establishes bounds on the accuracy of gradient estimators constructed using only local neighborhood information. We first define local Q function approximations. Fix $i \in \mathcal{N}$, and let $w^i : \mathcal{S}^{\mathcal{N}_\kappa^{-1}(i)} \times \mathcal{A}^{\mathcal{N}_\kappa^{-1}(i)} \to [0,1]$ be an arbitrary weighting function satisfying $\sum_{\bar{s} \in \mathcal{S}^{\mathcal{N}_\kappa^{-1}(i)}, \bar{a} \in \mathcal{A}^{\mathcal{N}_\kappa^{-1}(i)}} w^i(\bar{s}, \bar{a}) = 1$. Let $\widetilde{Q}^{f^i}_\theta(s^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa(i)})$ denote agent i's local approximation of $Q^{f^i}(s,a)$, for $f^i \in \{r^i, c^i\}$, defined by

$$\widetilde{Q}^{f^i}_\theta(s^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa(i)}) = \sum_{\bar{s} \in \mathcal{S}^{\mathcal{N}_\kappa^{-1}(i)}, \bar{a} \in \mathcal{A}^{\mathcal{N}_\kappa^{-1}(i)}} Q^{f^i}_\theta\left((s^{\mathcal{N}_\kappa(i)}, \bar{s}), (a^{\mathcal{N}_\kappa(i)}, \bar{a})\right) w^i(\bar{s}, \bar{a}). \tag{51}$$

For any such weighting function w^i, the following approximation result is implied by Theorem 1.

Theorem 2. *Let the conditions of Theorem 1 and Assumption 5 hold. Fix $i, j \in \mathcal{N}, f^i \in \{r^i, c^i\}$, define $f(s,a) = \sum_{i \in \mathcal{N}} f^i(s^i, a^i)$, and let M be as in Theorem 1. We have:*

(i) $|\widetilde{Q}^{f^i}_\theta(s^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa(i)}) - Q^{f^i}_\theta(s,a)| \leq \frac{M|\mathcal{N}_\kappa^{-1}(i)|}{g^2(\kappa,R)}$, *for all* $s \in \mathcal{S}, a \in \mathcal{A}$;

(ii) $\left\|\widehat{h^i_{f^j}}(\theta) - \nabla_{\theta^i} J_{f^j}(\theta)\right\| \leq \frac{ML^i|\mathcal{N}_\kappa^{-1}(j)|}{g^2(\kappa,R)}$, *where* $\widehat{h^i_{f^j}}(\theta) = \mathbb{E}_{\pi_\theta}\left[\widetilde{Q}^{f^j}_\theta(s^{\mathcal{N}_\kappa(j)}, a^{\mathcal{N}_\kappa(j)})\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right]$;

(iii) If Assumption 6 also holds, then $\left\|\widehat{h^i_f}(\theta) - \nabla_{\theta^i} J_f(\theta)\right\| \leq \frac{M\bar{n}L^i|\mathcal{N}_\kappa^{-1}(i)|}{g^2(\kappa,R)} + \left|\mathcal{N}_\kappa^{-1}(i)\right|\varepsilon_\kappa$, *where* $\widehat{h^i_f}(\theta) =$

$\mathbb{E}_{\pi_\theta}\left[\sum_{j\in\mathcal{N}_\kappa(i)} \widetilde{Q}_\theta^{f^j}(s^{\mathcal{N}_\kappa(j)}, a^{\mathcal{N}_\kappa(j)})\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right]$ *and* $\bar{n} = \max_{j\in\mathcal{N}}\left|\mathcal{N}_\kappa^{-1}(j)\right|$.

(iv) If Assumption 6 also holds, then $\left\|\sum_{j\in\mathcal{N}_\kappa(i)}\eta^j\widehat{h^i_{f^j}}(\theta) - \sum_{l\in\mathcal{N}}\eta^l\nabla_{\theta^i}J_{f^l}(\theta)\right\|$ $\leq \sum_{j\in\mathcal{N}_\kappa(i)}\left|\eta^j\right|\frac{ML^i\left|\mathcal{N}_\kappa^{-1}(j)\right|}{g^2(\kappa,R)} + \sum_{j\in\mathcal{N}_\kappa^{-1}(i)}\left|\eta^j\right|\varepsilon_\kappa$*, for all* $\eta\in\mathbb{R}^n$.

Proof. As in Theorem 1, the $f^i = c^i$ case is trivial by Assumption 2, so consider only the $f^i = r^i$ case. Fix $w^i, s\in\mathcal{S}$, and $a\in\mathcal{A}$.

Part (i). By the definition of $\widetilde{Q}_\theta^{r^i}$, we have

$$\left|\widetilde{Q}_\theta^{r^i}(s^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa(i)}) - Q_\theta^{r^i}(s,a)\right| \tag{52}$$

$$\leq \sum_{\bar{s}\in\mathcal{S}^{\mathcal{N}_\kappa^{-1}(i)},\bar{a}\in\mathcal{A}^{\mathcal{N}_\kappa^{-1}(i)}} w^i(\bar{s},\bar{a})\left|Q_\theta^{r^i}\left((s^{\mathcal{N}_\kappa(i)},\bar{s}),(a^{\mathcal{N}_\kappa(i)},\bar{a})\right) - Q_\theta^{r^i}(s,a)\right| \tag{53}$$

$$\leq \sum_{\bar{s}\in\mathcal{S}^{\mathcal{N}_\kappa^{-1}(i)},\bar{a}\in\mathcal{A}^{\mathcal{N}_\kappa^{-1}(i)}} w^i(\bar{s},\bar{a})\frac{M\left|\mathcal{N}_\kappa^{-1}(i)\right|}{g^2(\kappa,R)} = \frac{M\left|\mathcal{N}_\kappa^{-1}(i)\right|}{g^2(\kappa,R)}. \tag{54}$$

Part (ii). By the policy gradient theorem [18],

$$\nabla_{\theta^i}J_{r^j}(\theta) = \mathbb{E}_{\pi_\theta}\left[Q_\theta^{r^j}(s,a)\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right]. \tag{55}$$

This means we can write

$$\left\|\widehat{h^i_{r^j}}(\theta) - \nabla_{\theta^i}J_{r^j}(\theta)\right\| \tag{56}$$

$$= \mathbb{E}_{\pi_\theta}\left[\left(\widetilde{Q}_\theta^{r^j}(s^{\mathcal{N}_\kappa(j)}, a^{\mathcal{N}_\kappa(j)}) - Q_\theta^{r^j}(s,a)\right)\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right] \tag{57}$$

$$\overset{\text{(a)}}{\leq} \mathbb{E}_{\pi_\theta}\left[\left|\widetilde{Q}_\theta^{r^j}(s^{\mathcal{N}_\kappa(j)}, a^{\mathcal{N}_\kappa(j)}) - Q_\theta^{r^j}(s,a)\right|\cdot\left\|\nabla_{\theta^i}\log\pi^i_{\theta^i}(a^i|s^i)\right\|\right] \tag{58}$$

$$\overset{\text{(b)}}{\leq} \mathbb{E}_{\pi_\theta}\left[\frac{M\left|\mathcal{N}_\kappa^{-1}(j)\right|}{g^2(\kappa,R)}\cdot L^i\right] = \frac{ML^i\left|\mathcal{N}_\kappa^{-1}(j)\right|}{g^2(\kappa,R)}, \tag{59}$$

where (a) follows by Jensen's inequality and the Cauchy-Schwarz inequality, and (b) follows by Theorem 1 and Assumption 5.

Part (iii). By (15) and the definition of $\widehat{h^i_r}(\theta)$, we have

$$\nabla_{\theta^i}J_r(\theta) = \sum_{j\in\mathcal{N}}\nabla_{\theta^i}J_{r^j}(\theta), \quad \widehat{h^i_r}(\theta) = \sum_{j\in\mathcal{N}_\kappa(i)}\widehat{h^i_{r^j}}(\theta). \tag{60}$$

We thus have

$$\left\|\widehat{h^i_r}(\theta) - \nabla_{\theta^i}J_r(\theta)\right\| \tag{61}$$

$$= \left\|\sum_{j\in\mathcal{N}_\kappa(i)}\left(\widehat{h^i_{r^j}}(\theta) - \nabla_{\theta^i}J_{r^j}(\theta)\right) - \sum_{j\in\mathcal{N}_\kappa^{-1}(i)}\nabla_{\theta^i}J_{r^j}(\theta)\right\| \tag{62}$$

$$\leq \sum_{j\in\mathcal{N}_\kappa(i)}\left\|\widehat{h^i_{r^j}}(\theta) - \nabla_{\theta^i}J_{r^j}(\theta)\right\| + \sum_{j\in\mathcal{N}_\kappa^{-1}(i)}\left\|\nabla_{\theta^i}J_{r^j}(\theta)\right\| \tag{63}$$

$$\overset{\text{(a)}}{\leq} \sum_{j\in\mathcal{N}_\kappa(i)}\frac{ML^i\left|\mathcal{N}_\kappa^{-1}(j)\right|}{g^2(\kappa,R)} + \sum_{j\in\mathcal{N}_\kappa^{-1}}\varepsilon_\kappa \tag{64}$$

$$\overset{\text{(b)}}{\leq} \frac{ML^i\bar{n}\left|\mathcal{N}_\kappa^{-1}(j)\right|}{g^2(\kappa,R)} + \left|\mathcal{N}_\kappa^{-1}(i)\right|\varepsilon_\kappa, \tag{65}$$

where (a) follows by Part (ii) and Assumption 6, and (b) follows by the definition of $\bar{n}$ in the statement of the theorem.

Part (iv). Fix $\eta\in\mathbb{R}^n$. We know that

$$\left\|\sum_{j\in\mathcal{N}_\kappa(i)}\eta^j\widehat{h^i_{r^j}}(\theta) - \sum_{j\in\mathcal{N}}\eta^j\nabla_{\theta^i}J_{r^j}(\theta)\right\| \tag{66}$$

$$= \left\|\sum_{j\in\mathcal{N}_\kappa(i)}\eta^j\left[\widehat{h^i_{r^j}}(\theta) - \nabla_{\theta^i}J_{r^j}(\theta)\right] - \sum_{j\in\mathcal{N}_\kappa^{-1}(i)}\eta^j\nabla_{\theta^i}J_{r^j}(\theta)\right\| \tag{67}$$

$$\overset{\text{(a)}}{\leq} \sum_{j\in\mathcal{N}_\kappa(i)}\left|\eta^j\right|\left\|\widehat{h^i_{r^j}}(\theta) - \nabla_{\theta^i}J_{r^j}(\theta)\right\| + \sum_{j\in\mathcal{N}_\kappa^{-1}(i)}\left|\eta^j\right|\left\|\nabla_{\theta^i}J_{r^j}(\theta)\right\| \tag{68}$$

$$\overset{\text{(b)}}{\leq} \sum_{j\in\mathcal{N}_\kappa(i)}\left|\eta^j\right|\frac{ML^i\left|\mathcal{N}_\kappa^{-1}(j)\right|}{g^2(\kappa,R)} + \sum_{j\in\mathcal{N}_\kappa^{-1}(i)}\left|\eta^j\right|\varepsilon_\kappa, \tag{69}$$

where (a) follows by the triangle and Cauchy-Schwarz inequalities and (b) follows by Part (ii) and Assumption 6. □

Theorem 2 provides approximate policy gradient expressions that can be computed using only local neighborhood information and establishes corresponding error bounds depending on the system model, communication radius, choice of κ, policy design, and placement of radars within the network. The bounds arise due to signal decay inherent in the radar range equations (2)-(3), manifested in the reward (5). These bounds provide a guide to the selection of κ, design of g from Assumption 1, radar placement based on the effective communication range R, properties of the system model captured by M, properties of the parametric policy class through the Lipschitz constants L_i, and the number of radars n. Importantly, the result demonstrates that, for appropriate choice of design parameters given the underlying system, the expressions presented in Theorem 2 provide good approximations of the desired policy gradients while using only local neighborhood information. Armed with these policy gradient expressions, we next turn to development of decentralized MARL algorithms leveraging them to solve the problems proposed in Section II-B.

IV. Algorithms

In this section, we derive decentralized, policy gradient-based MARL methods for solving problems (P_κ^{max}) and (P_κ^{min}) proposed in Section II-B. In each case, we solve the problem using a decentralized, stochastic policy gradient descent-ascent procedure, or decentralized saddle point policy gradient (D-SP-PG), on the Lagrangian relaxation corresponding to the original problem. Performing gradient ascent-descent on the Lagrangian relaxation is a standard solution technique for constrained RL problems [22], while our extension to the MARL setting is inspired by decentralized approaches to optimization over networks [23], [24]. For each of problems (P_κ^{max}) and (P_κ^{min}), we first formulate the Lagrangian of the corresponding problem, then provide gradient expressions and local approximations for each agent, and finally state the corresponding decentralized algorithm.

Algorithm 1 SINR Maximization with Cost Constraints

```
1: Input: stepsizes $\alpha_t, \beta_t, \zeta_t$
2: Initialize: initialize $s_0, \theta_0, \nu_0$, set $t = \widehat{\mu_0^{c^i}} = \widehat{\mu_0^{r^i}} = 0$, and set all entries of $\widetilde{Q}^{r^i}, \widetilde{Q}^{c^i}$ to 0, for all $i \in \mathcal{N}$
3: for agent $i \in \mathcal{N}$ do
4:     share $s_t^i$ with $\mathcal{N}_\kappa(i)$, receive $s_t^{\mathcal{N}_\kappa(i)}$ from $\mathcal{N}_\kappa(i)$
5:     take action $a_t^i \sim \pi^i_{\theta^i_t}(\cdot|s_t^{\mathcal{N}_\kappa(i)})$
6:     observe $r_t^i = r^i(s_t, a_t)$, $c_t^i = c^i(s_t, a_t)$
7:     $\widehat{\mu_t^{c^i}} = (1-\zeta_t)\widehat{\mu_{t-1}^{c^i}} + \zeta_t c_t^i$
8:     $\widehat{\mu_t^{r^i}} = (1-\zeta_t)\widehat{\mu_{t-1}^{r^i}} + \zeta_t r_t^i$
9:     share $a_t^i$ with $\mathcal{N}_\kappa(i)$, receive $a_t^{\mathcal{N}_\kappa(i)}$ from $\mathcal{N}_\kappa(i)$
10:    $\widetilde{Q}_t^{c^i} = \texttt{UPDATE\_Q}(\widetilde{Q}_{t-1}^{c^i}, c_t^i, \widehat{\mu_t^{c^i}}, s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}, \zeta_t)$
11:    $\widetilde{Q}_t^{r^i} = \texttt{UPDATE\_Q}(\widetilde{Q}_{t-1}^{r^i}, r_t^i, \widehat{\mu_t^{r^i}}, s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}, \zeta_t)$
12:    share $\widetilde{Q}_t^{r^i}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}), \widehat{\mu_t^{c^i}}, \nu_t^i$ with $\mathcal{N}_\kappa(i)$, receive $\widetilde{Q}_t^{r^j}(s_t^{\mathcal{N}_\kappa(j)}, a_t^{\mathcal{N}_\kappa(j)}), \widehat{\mu_t^{c^j}}, \nu_t^j$ from $\mathcal{N}_\kappa(i)$
13:    form estimates:
           $\widehat{h^i_{r,t}} = \sum_{j \in \mathcal{N}_\kappa(i)} \widetilde{Q}_t^{r^j}(s_t^{\mathcal{N}_\kappa(j)}, a_t^{\mathcal{N}_\kappa(j)}) \nabla_{\theta^i} \log \pi^i_{\theta^i_t}(a_t^i|s_t^{\mathcal{N}_\kappa(i)})$
           $\widehat{h^i_{c^i,t}} = \widetilde{Q}_t^{c^i}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}) \nabla_{\theta^i} \log \pi^i_{\theta^i}(a_t^i|s_t^{\mathcal{N}_\kappa(i)})$
14:    update:
           $\theta^i_{t+1} = \theta^i_t + \alpha_t\big(\widehat{h^i_{r,t}} - \sum_{j \in \mathcal{N}_\kappa(i)} \nu^j \widehat{h^i_{c^i,t}}\big)$
           $\nu^i_{t+1} = \nu^i_t - \beta_t\big(u^i - \sum_{j \in \mathcal{N}_\kappa(i)} \widehat{\mu_t^{c^j}}\big)$
15:    $t \leftarrow t+1$
16: end for
```

A. Sum-of-SINRs Maximization

The Lagrangian of problem (P_κ^{max}) is given by

$$\mathscr{L}_Q(\theta, \nu) = J_r(\theta) + \sum_{j \in \mathcal{N}} \nu^j \big(u^j - \sum_{k \in \mathcal{N}_\kappa(j)} J_{c^k}(\theta)\big), \tag{70}$$

where we recall that $J_r(\theta) = \sum_{j \in \mathcal{N}} J_{r^j}(\theta)$. In order to solve ($P_\kappa^{max}$), our goal is to instead solve the problem $\max_\theta \min_\nu \mathscr{L}_Q(\theta, \nu)$ by alternating between gradient ascent in θ and gradient descent in ν using stochastic approximates of the gradient expressions

$$\nabla_{\theta^i} \mathscr{L}_Q(\theta, \nu) = \nabla_{\theta^i} J_r(\theta) - \sum_{j \in \mathcal{N}} \nu^j \sum_{k \in \mathcal{N}_\kappa(j)} \nabla_{\theta^i} J_{c^k}(\theta) \tag{71}$$

$$= \nabla_{\theta^i} J_r(\theta) - \sum_{j \in \mathcal{N}_\kappa(i)} \nu^j \nabla_{\theta^i} J_{c^i}(\theta) \tag{72}$$

$$\nabla_{\nu^i} \mathscr{L}_Q(\theta, \nu) = u^i - \sum_{j \in \mathcal{N}_\kappa(i)} J_{c^j}(\theta). \tag{73}$$

Here equation (72) follows from Assumption 2 combined with our assumption that the κ-hop neighbor relation is symmetric, i.e., that $i \in \mathcal{N}_\kappa(j)$ if and only if $j \in \mathcal{N}_\kappa(i)$.

We can now present a decentralized learning algorithm for approximately solving this problem, where each agent only needs information available within its local κ-hop neighborhood. Using the approximations provided by Theorem 2, for suitable choice of κ and R we have

$$\nabla_{\theta^i} \mathscr{L}_Q(\theta, \nu) \approx \widehat{h^i_r}(\theta) - \sum_{j \in \mathcal{N}_\kappa(i)} \nu^j \widehat{h^i_{c^i}}(\theta), \tag{74}$$

$$\nabla_{\nu^i} \mathscr{L}_Q(\theta, \nu) \approx u^i - \sum_{j \in \mathcal{N}_\kappa(i)} \widehat{\mu^{c^j}}(\theta), \tag{75}$$

Algorithm 2 Cost Minimization with SINR Constraints

```
1: Input: stepsizes $\alpha_t, \beta_t, \zeta_t$
2: Initialize: initialize $s_0, \theta_0, \eta_0, \nu_0$, set $t = \widehat{\mu_0^{c^i}} = \widehat{\mu_0^{r^i}} = 0$, and set all entries of $\widetilde{Q}^{r^i}, \widetilde{Q}^{c^i}$ to 0, for all $i \in \mathcal{N}$
3: for agent $i \in \mathcal{N}$ do
4:     share $s_t^i$ with $\mathcal{N}_\kappa(i)$, receive $s_t^{\mathcal{N}_\kappa(i)}$ from $\mathcal{N}_\kappa(i)$
5:     take action $a_t^i \sim \pi^i_{\theta^i_t}(\cdot|s_t^{\mathcal{N}_\kappa(i)})$
6:     observe $r_t^i = r^i(s_t, a_t)$, $c_t^i = c^i(s_t, a_t)$
7:     $\widehat{\mu_t^{c^i}} = (1-\zeta_t)\widehat{\mu_{t-1}^{c^i}} + \zeta_t c_t^i$
8:     $\widehat{\mu_t^{r^i}} = (1-\zeta_t)\widehat{\mu_{t-1}^{r^i}} + \zeta_t r_t^i$
9:     share $a_t^i$ with $\mathcal{N}_\kappa(i)$, receive $a_t^{\mathcal{N}_\kappa(i)}$ from $\mathcal{N}_\kappa(i)$
10:    $\widetilde{Q}_t^{c^i} = \texttt{UPDATE\_Q}(\widetilde{Q}_{t-1}^{c^i}, c_t^i, \widehat{\mu_t^{c^i}}, s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}, \zeta_t)$
11:    $\widetilde{Q}_t^{r^i} = \texttt{UPDATE\_Q}(\widetilde{Q}_{t-1}^{r^i}, r_t^i, \widehat{\mu_t^{r^i}}, s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}, \zeta_t)$
12:    share $\widetilde{Q}_t^{r^i}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}), \widehat{\mu_t^{c^i}}, \nu_t^i$ with $\mathcal{N}_\kappa(i)$, receive $\widetilde{Q}_t^{r^j}(s_t^{\mathcal{N}_\kappa(j)}, a_t^{\mathcal{N}_\kappa(j)}), \widehat{\mu_t^{c^j}}, \nu_t^j$ from $\mathcal{N}_\kappa(i)$
13:    form estimates:
           $\widehat{h^i_{c^i,t}} = \widetilde{Q}_t^{c^i}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}) \nabla_{\theta^i} \log \pi^i_{\theta^i}(a_t^i|s_t^{\mathcal{N}_\kappa(i)})$
           $\widehat{h^i_{r^j,t}} = \widetilde{Q}_t^{r^j}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(j)}) \nabla_{\theta^i} \log \pi^i_{\theta^i}(a_t^i|s_t^{\mathcal{N}_\kappa(i)})$,
               for all $j \in \mathcal{N}_\kappa(i)$
14:    update:
           $\theta^i_{t+1} = \theta^i_t - \alpha_t\Big(\big(1 + \sum_{j \in \mathcal{N}_\kappa(i)} \nu^j\big)\widehat{h^i_{c^i,t}} - \sum_{j \in \mathcal{N}_\kappa(i)} \eta^j \widehat{h^i_{r^j,t}}\Big)$
           $\eta^i_{t+1} = \eta^i_t + \beta_t\Big(\gamma_{min} - \widehat{\mu_t^{r^i}}\Big)$
           $\nu^i_{t+1} = \nu^i_t + \delta_t\Big(u^i - \sum_{j \in \mathcal{N}_\kappa(i)} \widehat{\mu_t^{c^j}}\Big)$
15:    $t \leftarrow t+1$
16: end for
```

Algorithm 3 `UPDATE_Q`

```
1: Input: $\widetilde{Q}_{t-1}^{f^i}, f_t^i, \widehat{\mu_t^{f^i}}, s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}, \zeta_t$
2: perform updates
       $\widetilde{Q}_t^{f^i}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)}) = (1-\zeta_t)\widetilde{Q}_{t-1}^{f^i}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)})$
               $+ \zeta_t \Big(f_t^i - \widehat{\mu_t^{f^i}} + \widetilde{Q}_{t-1}^{f^i}(s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)})\Big)$
       $\widetilde{Q}_t^{f^i}(s^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa(i)}) = \widetilde{Q}_{t-1}^{f^i}(s^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa(i)})$, for all
               $(s^{\mathcal{N}_\kappa(i)}, a^{\mathcal{N}_\kappa(i)}) \neq (s_t^{\mathcal{N}_\kappa(i)}, a_t^{\mathcal{N}_\kappa(i)})$
3: return $\widetilde{Q}_t^{f^i}$
```

where $\widehat{\mu^{c^j}}(\theta) \approx J_{c^j}(\theta)$ is a suitable approximation, such as a cumulative or exponential moving average. Using these expressions, we provide the D-SP-PG scheme for solving (P_κ^{max}) in Algorithm 1.

B. Power Minimization with SINR Threshold

The Lagrangian of problem (P_κ^{min}) is given by

$$\mathscr{L}_R(\theta, \eta, \nu) = J_c(\theta) + \sum_{j \in \mathcal{N}} \eta^j \left(\gamma_{min} - J_{r^j}(\theta)\right) \tag{76}$$

$$- \sum_{j \in \mathcal{N}} \nu^j \big(u^j - \sum_{k \in \mathcal{N}_\kappa(j)} J_{c^k}(\theta)\big), \tag{77}$$

where we recall that $J_c(\theta) = \sum_{j \in \mathcal{N}} J_{c^j}(\theta)$. In order to solve (P_κ^{min}), we instead solve the saddle point problem $\min_\theta \max_{\eta,\nu} \mathscr{L}_R(\theta, \eta, \nu)$ by alternating between stochastic gradient descent in θ and ascent in η and ν. Differentiating (76), (77) with respect to θ^i gives

$$\nabla_{\theta^i} \mathscr{L}_R(\theta, \eta, \nu) = \nabla_{\theta^i} J_c(\theta) - \sum_{j \in \mathcal{N}} \eta^j \nabla_{\theta^i} J_{r^j}(\theta) \tag{78}$$

$$+ \sum_{j \in \mathcal{N}} \nu^j \sum_{k \in \mathcal{N}_\kappa(j)} \nabla_{\theta^i} J_{c^k}(\theta) \tag{79}$$

$$= \nabla_{\theta^i} J_{c^i}(\theta) - \sum_{j \in \mathcal{N}} \eta^j \nabla_{\theta^i} J_{r^j}(\theta) + \sum_{k \in \mathcal{N}_\kappa(i)} \nu^k \nabla_{\theta^i} J_{c^i}(\theta) \tag{80}$$

$$= \Big(1 + \sum_{k \in \mathcal{N}_\kappa(i)} \nu^j\Big) \nabla_{\theta^i} J_{c^i}(\theta) - \sum_{j \in \mathcal{N}} \eta^j \nabla_{\theta^i} J_{r^j}(\theta), \tag{81}$$

where the first equality holds by Assumption 2 combined with our assumption that the κ-hop neighbor relation is symmetric, i.e., that $i \in \mathcal{N}_\kappa(j)$ if and only if $j \in \mathcal{N}_\kappa(i)$. Finally, differentiating with respect to η^i and ν^i yields

$$\nabla_{\eta^i} \mathscr{L}_R(\theta, \eta, \nu) = \gamma_{min} - J_{r^i}(\theta), \tag{82}$$

$$\nabla_{\nu^i} \mathscr{L}_R(\theta, \eta, \nu) = u^i - \sum_{j \in \mathcal{N}_\kappa(i)} J_{c^j}(\theta). \tag{83}$$

Our goal is again to perform the necessary updates in a decentralized manner, with each agent only using locally available information. Using the approximations provided by Theorem 2, for suitable choice of κ and R we have

$$\nabla_{\theta^i} \mathscr{L}_R(\theta, \eta, \nu) \approx \Big(1 + \sum_{k \in \mathcal{N}_\kappa(i)} \nu^k\Big) \widehat{h^i_{c^i}}(\theta) - \sum_{j \in \mathcal{N}_\kappa(i)} \eta^j \widehat{h^i_{r^j}}(\theta) \tag{84}$$

$$\nabla_{\eta^i} \mathscr{L}_R(\theta, \eta, \nu) \approx \gamma_{min} - \widehat{\mu^{r^i}}(\theta) \tag{85}$$

$$\nabla_{\nu^i} \mathscr{L}_R(\theta, \eta, \nu) \approx u^i - \sum_{j \in \mathcal{N}_\kappa(i)} \widehat{\mu^{c^j}}(\theta), \tag{86}$$

where $\widehat{\mu^{r^i}}(\theta) \approx J_{r^l}(\theta)$ and $\widehat{\mu^{c^j}}(\theta) \approx J_{c^j}(\theta)$ are suitable approximations. Using these expressions, we provide the D-SP-PG scheme for solving (P_κ^{min}) in Algorithm 2.

Remark 2. We note that, in order to lay the groundwork for convergence analysis of our algorithms in future work, we have presented the Q updates of Algorithm 3 in the easily analyzable tabular form, which is only applicable to finite state and action spaces. We emphasize that practical variants leveraging standard neural network architectures for Q function approximation can be substituted in Algorithm 3 to enable application to the continuous spaces of Section II.

V. Conclusion

In this work, we have considered the specific use-case of power allocation for target detection in a radar network to illustrate how signal strength decay properties inherent in wireless communications and radar networks can be used to develop decentralized, scalable MARL methods. Future directions include convergence analysis of Algorithms 1 and 2, experimental evaluation of neural network-based versions of our algorithms, extension of our approach to additional applications in communications and radar networks, and extension of our approach to joint motion planning and power allocation in target detection in mobile radar networks.

References

[1] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International conference on machine learning*. PMLR, 2018, pp. 5872–5881.

[2] Wesley A Suttle, Zhuoran Yang, Kaiqing Zhang, Zhaoran Wang, Tamer Başar, and Ji Liu, "A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning," in *21st International Federation of Automatic Control (IFAC) World Congress*. IEEE, 2020, pp. 1549–1554.

[3] Krishna Chaitanya Kosaraju, Seetharaman Sivaranjani, Wesley A Suttle, Vijay Gupta, and Ji Liu, "Reinforcement learning based distributed control of dissipative networked systems," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 2, pp. 856–866, 2021.

[4] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of reinforcement learning and control*, pp. 321–384, 2021.

[5] Pablo Hernandez-Leal, Bilal Kartal, and Matthew E Taylor, "A survey and critique of multiagent deep reinforcement learning," *Autonomous Agents and Multi-Agent Systems*, vol. 33, no. 6, pp. 750–797, 2019.

[6] Sven Gronauer and Klaus Diepold, "Multi-agent deep reinforcement learning: a survey," *Artificial Intelligence Review*, vol. 55, no. 2, pp. 895–943, 2022.

[7] Changxi Zhu, Mehdi Dastani, and Shihan Wang, "A survey of multi-agent deep reinforcement learning with communication," *Autonomous Agents and Multi-Agent Systems*, vol. 38, no. 1, pp. 4, 2024.

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

[9] Guannan Qu, Adam Wierman, and Na Li, "Scalable reinforcement learning for multiagent networked systems," *Operations Research*, vol. 70, no. 6, pp. 3601–3628, 2022.

[10] Lijun Zhang, Lin Li, Wei Wei, Huizhong Song, Yaodong Yang, and Jiye Liang, "Scalable constrained policy optimization for safe multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 138698–138730, 2024.

[11] Mostafa M Shibl, Wesley A Suttle, and Vijay Gupta, "Scalable natural policy gradient for general-sum linear quadratic games with known parameters," in *7th Annual Learning for Dynamics & Control Conference*. PMLR, 2025, pp. 1–14.

[12] Andrea Goldsmith, *Wireless communications*, Cambridge university press, 2005.

[13] Mark A Richards, *Fundamentals of radar signal processing*, Mcgraw-hill New York, 2005.

[14] Simon Haykin, "Cognitive radar: a way of the future," *IEEE signal processing magazine*, vol. 23, no. 1, pp. 30–40, 2006.

[15] Anastasios Deligiannis and Sangarapillai Lambotharan, "A bayesian game theoretic framework for resource allocation in multistatic radar networks," in *2017 IEEE Radar Conference (RadarConf)*. IEEE, 2017, pp. 0546–0551.

[16] Chenguang Shi, Sana Salous, Fei Wang, and Jianjiang Zhou, "Power allocation for target detection in radar networks based on low probability of intercept: A cooperative game theoretical strategy," *Radio Science*, vol. 52, no. 8, pp. 1030–1045, 2017.

[17] Luke Snow, Vikram Krishnamurthy, and Brian M Sadler, "Identifying coordination in a cognitive radar network-a multi-objective inverse reinforcement learning approach," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[18] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour, "Policy gradient methods for reinforcement learning with function approximation," *Advances in neural information processing systems*, vol. 12, 1999.

[19] Yue Frank Wu, Weitong Zhang, Pan Xu, and Quanquan Gu, "A finite-time analysis of two time-scale actor-critic methods," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17617–17628, 2020.

[20] Xuyang Chen and Lin Zhao, "Finite-time analysis of single-timescale actor-critic," *Advances in Neural Information Processing Systems*, vol. 36, pp. 7017–7049, 2023.

[21] Wesley A Suttle, Amrit Bedi, Bhrij Patel, Brian M Sadler, Alec Koppel, and Dinesh Manocha, "Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level monte carlo actor-critic," in *International Conference on Machine Learning*. PMLR, 2023, pp. 33240–33267.

[22] Santiago Paternain, Luiz Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro, "Constrained reinforcement learning has zero duality gap," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[23] Alec Koppel, Felicia Y Jakubiec, and Alejandro Ribeiro, "A saddle point algorithm for networked online convex optimization," *IEEE Transactions on Signal Processing*, vol. 63, no. 19, pp. 5149–5164, 2015.

[24] Alec Koppel, Brian M Sadler, and Alejandro Ribeiro, "Proximity without consensus in online multiagent optimization," *IEEE Transactions on Signal Processing*, vol. 65, no. 12, pp. 3062–3077, 2017.