MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

Spectrum Sharing-inspired Safe Motion Planning

Kim, Kyeong Jin; Vinod, Abraham P.; Guo, Jianlin; Deshpande, Vedang M.; Parsons, Kieran

TR2023-049 May 31, 2023

Abstract

In this paper, the problem of safe motion planning of the mobile agent in the presence of multiple static obstacles is investigated. In addition to the collision avoidance, an additional objective of joint minimizing the energy consumption for con- trolling its dynamic movement and maximizing the instantaneous post-processing signal-to-noise ratio (ISNR) that determines the accessing capability of spectrum allocated to the licensed users is taken into account. Due to a non-existing system setup for single carrier transmissions in the spatial-temporal correlated frequency selective fading channel and non-existing feasible mathematical analysis to maximize the distribution of the ISNR over the energy conscious motion planning, we propose a model- free and off-policy soft actor critic (SAC) to learn and make an action by the mobile agent to achieve the following three objectives simultaneously from interactions with the environment: i) achieve the safe motion planning that avoids collision with the static obstacles; ii) minimize the control energy consumption; and iii) maximize the ISNR. Simulation results verify that these three objectives can be achieved efficiently by the proposed SAC-based safe motion planning.

IEEE International Conference on Communications Workshops (ICC) 2023

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

Spectrum Sharing-inspired Safe Motion Planning

Kyeong Jin Kim, Abraham P. Vinod, Jianlin Guo, Vedang Deshpande, and Kieran Parsons

***Abstract*—In this paper, the problem of safe motion planning of the mobile agent in the presence of multiple static obstacles is investigated. In addition to the collision avoidance, an additional objective of joint minimizing the energy consumption for controlling its dynamic movement and maximizing the instantaneous post-processing signal-to-noise ratio (ISNR) that determines the accessing capability of spectrum allocated to the licensed users is taken into account. Due to a non-existing system setup for single carrier transmissions in the spatial-temporal correlated frequency selective fading channel and non-existing feasible mathematical analysis to maximize the distribution of the ISNR over the energy conscious motion planning, we propose a model-free and off-policy soft actor critic (SAC) to learn and make an action by the mobile agent to achieve the following three objectives simultaneously from interactions with the environment: i) achieve the safe motion planning that avoids collision with the static obstacles; ii) minimize the control energy consumption; and iii) maximize the ISNR. Simulation results verify that these three objectives can be achieved efficiently by the proposed SAC-based safe motion planning.**

***Index Terms*—Motion planning, cyclic prefixed single carrier transmissions, spectrum sharing, reinforcement learning, soft actor critic, optimal policy.**

I. Introduction

Unmanned Autonomous ground vehicles (AGVs) and mobile robots become indispensable tools for Industry 4.0, smart manufacturing, and the future revolutionary industrial, manufacturing, and smart factory innovations [1], [2]. They can be deployed for transportation, environmental monitoring, and accomplishing tasks impossible and dangerous for human workers.

Motion planning is mainly related with the movement of AGV or robots between multiple points under uncertain environments. Via safe motion planning, they can find a route under an evaluation criterion, while avoiding static and dynamic obstacles [3]–[5]. AGVs and transportation robots are seen as unsustainable equipment that demand a high level of energy consumption for the movement, communications, sensing, and computation since batteries are usually used to provide energy for them. In particular, how to optimize energy consumption for dynamic movement is an important issue for economic and environmental reasons since this accounts for most of it. Thus, energy-efficient trajectory planning of an industrial robot have been proposed by many existing works such as [3], [6], [7]. In the cloud networked multi-robot system [3], the authors considered cloud computing to reduce execution time and energy consumption, which is possible by offloading its computation into cloud. In [6], the authors proposed an energy-efficient motion planner for multi-robot coordination. Each robot is forced to reach an exploration target with the lowest energy consumption. For a material handling robot, the authors in [7] investigated an energy conscious scheduling under the bound of movement. In recent years, model free or data-driven machine learning (ML) techniques that do not require to know the parametric model have led to exceptional improvements in a wide range of applications. The ability of ML to learn complex hidden models from data has proven quite successful, quickly surpassing most state-of-the-art human-designed algorithms. In particular, the authors in [4], [8]–[10], leveraged reinforcement learning (RL) in developing the safe motion planning. In their approaches, a mobile robot is recognized as a mobile agent in the environment to learn its motion policy by maximizing a specific criterion while avoiding collision with static and mobile obstacles. By employing the deep Q-learning (DQL), the mobile robot is trained to make its action with no explicit information about the environment [9]. Jointly considering competition and cooperation among multi-unmanned aerial vehicles (UAVs), a multi-agent deep deterministic policy gradient (MADDPG) algorithm was used in [10] for the safe motion planning.

Dedicated private spectrum is allocated by the German, Japan, and USA governments for industrial use to support innovation and enable new use cases in the private network such as control, resilient and reliable low latency wireless automation, extensive Internet of Things (IoT) and industrial IoT (IIoT), and secured network within premise [11]. Unlicensed spectrum can be used for their operation. However, its usage may raise some concerns due to external interference and malicious jamming that can result in service disruptions. The role of spectrum sharing has become key in spectrum management, allowing access to new bands and carrier aggregation to boost network performance, while protecting their operations and the access rights of incumbents or those of multiple different use cases, with distinct priorities. Many mathematical analyses have been developed for the frequency selective fading channel. However, to the best our knowledge, performance analysis of spectrum sharing in the frequency selective fading channel with spatially and temporally correlated shadow fading has never been investigated.

In contrast to the mentioned works, the following problems are still open for further investigation with the safe motion planning.

- How to conduct a safe motion planning of the mobile agent while maximizing the instantaneous post-processing signal-to-noise ratio (ISNR) and minimizing

Authors are with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA.

control energy consumption being used for its dynamic movement. Thus, this will be a novel problem compared with those of [4], [8]–[10], and [12]
- How to incorporate a more realistic frequency selective fading channel jointly taking into account of spatial-temporal correlation in shadow fading over the safe motion control. Thus, in contrast to the work [5], [12], it is desirable to obtain the motion planning policy of the mobile agent without exact knowledge of the channel model.
- How much spectrum sharing and energy saving can benefit from the joint optimization under the framework of the safe motion planning is also an important problem.

II. System and Channel Models

Fig. 1 illustrates the considered network operating in the environment with multiple obstacles that work as the primary-user (PU)-receivers (PU-RXs), $\{\text{obs}_j, \forall j\}$, one secondary-user (SU)-TX, and a mobile agent that works as the SU-receiver (SU-RX). We assume that every node is equipped with a single antenna to transmit and receive the signals. All PU-RXs are coexistent in the same licensed frequency band. However, SUs are assumed to be operating in a unlicensed different frequency band. However, when the mobile agent wants to accomplish a reliable task with a higher priority, it is necessary to access the licensed frequency band.

Definition 1: (Collective safety [4]) The agent is assumed to be collectively safe with its motion planning at a particular time epoch when i) the agent with its size r_a avoids collision with obstacles with size r_o, located at $\boldsymbol{p}_{o_j} \in \mathbb{R}^{2\times 1}$, $\forall j$; and ii) the agent is moving within the working area. In particular, since the SU-TX is mounted at a higher height over that of the agent, it is not recognized as the static obstacle.

A. Dynamic model of the mobile agent

We setup a discrete-time dynamic model of the mobile agent, similar to [4]. We describe the dynamics as follows:

$$\begin{aligned} \boldsymbol{x}(k+1) &= \boldsymbol{A}\boldsymbol{x}(k) + \boldsymbol{B}\mathbf{u}(k), \\ \boldsymbol{p}_r(k) &= \boldsymbol{H}\boldsymbol{x}(k) \end{aligned} \tag{1}$$

where discrete-time state $\boldsymbol{x}(k) \triangleq [x(k), v_x(k), y(k), v_y(k)]^T$, input $\mathbf{u}(k) \triangleq [u_x(k), u_y(k)]^T$, matrices $\boldsymbol{A} \triangleq \begin{bmatrix} \boldsymbol{\Phi}_T & \mathbf{0}_{2\times 2} \\ \mathbf{0}_{2\times 2} & \boldsymbol{\Phi}_T \end{bmatrix}$ and $\boldsymbol{B} \triangleq \begin{bmatrix} \boldsymbol{\phi}_T & \mathbf{0}_{2\times 1} \\ \mathbf{0}_{2\times 1} & \boldsymbol{\phi}_T \end{bmatrix}$ with $\boldsymbol{\Phi}_T \triangleq \begin{bmatrix} 1 & \Delta T \\ 0 & 1 \end{bmatrix}$ and $\boldsymbol{\phi}_T \triangleq \begin{bmatrix} (\Delta T)^2/2 \\ \Delta T \end{bmatrix}$ for some sampling time $\Delta T > 0$, and $\boldsymbol{H} \triangleq \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}$. We define the observation vector, $\boldsymbol{p}_r(k) \triangleq [x(k), y(k)]^T$ as the position coordinates of the agent at $k\Delta T$ in the 2D-space. Similarly, we define an agents's velocity vector $\boldsymbol{v}(k) = [v_x(k), v_y(k)]^T$.

For the ease of training an RL policy to control (1), we use following low-level, tracking controller,

$$\boldsymbol{u}^{\text{RL}}(k) = -\boldsymbol{K}\boldsymbol{x}(k) + \boldsymbol{F}\boldsymbol{r}(k) \tag{2}$$

where $\boldsymbol{r}(k)$ is the command set by the RL-based motion planner [4], and the matrices $\boldsymbol{K}, \boldsymbol{F}$ are determined via standard LQR theory [14]. In particular, the controller gain $\boldsymbol{K}$ is obtained by $\boldsymbol{K} = (\boldsymbol{B}^T\boldsymbol{S}\boldsymbol{B} + \boldsymbol{R})^{-1}\boldsymbol{B}^T\boldsymbol{S}\boldsymbol{A}$, where $\boldsymbol{S}$ is the solution to the algebraic Riccati equation given by

$$\boldsymbol{S} = \boldsymbol{A}^T\left(\boldsymbol{S} - \boldsymbol{S}\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{S}\boldsymbol{B} + \boldsymbol{R})^{-1}\boldsymbol{B}^T\boldsymbol{S}\right)\boldsymbol{A} + \boldsymbol{Q},$$

and pre-specified matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ are respectively positive semidefinite and positive definite matrices associated with the quadratic costs defined for state and control vectors. We then solve the following equation to compute $\boldsymbol{F}$,

$$\boldsymbol{H}(\boldsymbol{I}_{4\times 4} - (\boldsymbol{A} - \boldsymbol{B}\mathbf{K}))^{-1}\mathbf{B}\boldsymbol{F} = \boldsymbol{I}_{2\times 2} \tag{3}$$

By construction, the use of $u^{\text{RL}}(k)$ in the dynamics (1) guarantees $\lim_{k\to\infty} \boldsymbol{p}_r(k) = \boldsymbol{r}$ for a constant $\boldsymbol{r}(k) = \boldsymbol{r}$. In other words, the controller (2) ensures that the system (1) drives the system such that when the position $r(k)$ specified by the RL-based planner is held (nearly) constant to r, the position of the agent converges to r.

Definition 2: (Energy) We define the energy needed to control the mobile agent,

$$E_{\text{control}} \triangleq \sum\nolimits_{k=1}^{K} \|\boldsymbol{u}^{\text{RL}}(k)\|^2 \tag{4}$$

where K denotes a finite horizon to arrive at the target position, $\boldsymbol{t} \in \mathbb{R}^{2\times 1}$, within the displacement threshold, denoted by r_d.

B. Frequency selective fading channel model

A communication channel between two nodes can be modeled by using small-scale fading for multipath, large-scale fading for shadowing and path loss. For two nodes i and j, respectively located at $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$, the logarithm of the squared channel magnitude of $f(\boldsymbol{p}_i, \boldsymbol{p}_j, \boldsymbol{h}_{i,j}) \in \mathbb{C}^{N_h\times 1}$ with N_h multipath components is expressed as follows [15]:

$$\begin{aligned} F_{\text{dB}}(\boldsymbol{p}_i, \boldsymbol{p}_j, \boldsymbol{h}_{i,j}) &\triangleq 10\log_{10}(\|f(\boldsymbol{p}_i, \boldsymbol{p}_j, \boldsymbol{h}_{i,j})\|^2) \\ &= -\eta 10\log_{10}(\|\boldsymbol{p}_i - \boldsymbol{p}_j\|) + F_{\text{SH}}(\boldsymbol{p}_i, \boldsymbol{p}_j) + \\ &\quad F_{\text{SF}}(\boldsymbol{p}_i, \boldsymbol{p}_j, \boldsymbol{h}_{i,j}) \end{aligned} \tag{5}$$

where η is the path loss exponent. A fading channel between these two node is assumed to be a frequency selective fading channel denoted by $\boldsymbol{h}_{i,j}$. In addition, $F_{\text{SH}}(\boldsymbol{p}_i, \boldsymbol{p}_j)$ and $F_{\text{SF}}(\boldsymbol{p}_i, \boldsymbol{p}_j, \boldsymbol{h}_{i,j})$ respectively represent the effects of shadow fading and frequency selective fading channel, $\boldsymbol{h}_{i,j}$, in dB. A zero-mean Gaussian distribution with an exponential spatial correlation is used to model $F_{\text{SH}}(\boldsymbol{p}_i, \boldsymbol{p}_j)$. How large-scale shadow fading components are changing spatially, the covariance between $\boldsymbol{p}_i$ and $\boldsymbol{p}_j$ is given by $\Omega_{i,j} = (\epsilon_{dB})^2 e^{-\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|}{\eta_S}}$, where $(\epsilon_{dB})^2$ and η_S respectively denote the variance of the shadow fading component in dB and decorrelation distance that controls the spatial correlation [15].

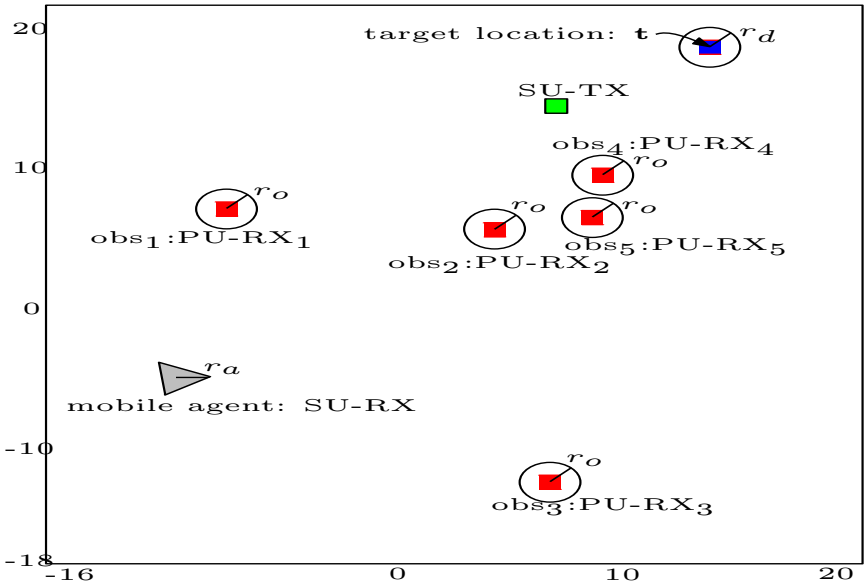

Fig. 1. Illustration of one use case of the cyclic prefixed single-carrier (CP-SC) transmissions in the wireless network in which a mobile agent that employs CP-SC transmissions [13] makes a motion planning to arrive at the target location, $\boldsymbol{t}$, in the environment coexisting with multiple static obstacles. Centers of node location are denoted by $\boldsymbol{p}_{\text{SU-TX}}$, $\boldsymbol{p}_{o_j}$, and $\boldsymbol{p}_r$, for a secondary-user (SU)-transmitter (SU-TX), jth obstacles, obs_j, and mobile agent.

For a frequency selective fading channel, $\boldsymbol{h}_{i,j}$, $F_{\text{SF}}(\boldsymbol{p}_i, \boldsymbol{p}_j, \boldsymbol{h}_{i,j})$ is modeled by the following probability density function (PDF):

$$f_{\Upsilon_{\text{SF}}(\boldsymbol{p}_i,\boldsymbol{p}_j)}(w) = \frac{\log(10)}{10}\frac{10^{\frac{w}{10}}}{\Gamma(N_h)}e^{-10^{\frac{w}{10}}}(10^{\frac{w}{10}})^{N_h-1} \quad (6)$$

where $\Gamma(\cdot)$ denotes the Gamma function.

C. Spectrum sharing

Many measurement campaigns have demonstrated that a large amount of licensed radio spectrum is inefficiently utilized due to the existence spatial and temporal holes. To counter these inefficient use of scarce radio spectrum, the cognitive radio (CR) network [16] was proposed. How the SU is allowed to access spectrum, three different types of spectrum sharing, namely, overlay, underlay, and interweave have been proposed. As an underlay scheme, spectrum sharing is effective in implementing dynamic spectrum management. Spectrum sharing allows the SU to co-occupy target spectrum as long as its interference to the PU is under a threshold, i.e., the PU can tolerate to interference produced on its assigned spectrum [12]. Thus, it can reduce the need of strict interference management [17], [18] for spectrum sharing.

Referring to Fig. 1, let us model that to meet the reliable transportation at a particular time the mobile agent that operates in the unlicensed spectrum wants to occupy the target spectrum assigned to the PUs. Let P_T denote the peak transmit power at the SU-TX, and I_P be the maximum allowable interference at all the PU-RXs. Under these two constraints, the transmit power allocation at the SU-TX is given by [12]:

$$P_s = \min\left(P_T, \frac{I_p}{\max_k \|f(\boldsymbol{p}_{\text{SU-TX}}, \boldsymbol{p}_{o_k}, \boldsymbol{h}_k)\|^2}\right) \quad (7)$$

where $\boldsymbol{h}_k$ is the frequency selective fading channel with N_g multipath elements from the SU-TX to the kth obstacle. Based on received signals transmitted from the SU-TX, the ISNR is given by

$$\gamma_{\text{spectrum}} = \min\left(P_T, \frac{I_p}{\max_k \|f(\boldsymbol{p}_{\text{SU-TX}}, \boldsymbol{p}_{o_k}, \boldsymbol{h}_k)\|^2}\right) \\ (\|f(\boldsymbol{p}_{\text{SU-TX}}, \boldsymbol{p}_r, \boldsymbol{g})\|^2) \quad (8)$$

where $\boldsymbol{g}$ is the frequency selective fading channel with N_g multipath elements from the SU-TX to the receive antenna at the mobile agent. As was verified by [12], the spectrum access capability of the mobile user is determined by the magnitude of γ_{spectrum}, which is a random variable. To simplify the model, we assume that N_g and N_h are independent of indices of k, that is, the number of multipath elements is independent of the connection to obstacles.

III. Soft Actor-Critic Network for Motion Planning

We first define the deterministic Markov decision process (MDP) for the mobile agent with the defined linear dynamics, expressed by (1), with a prescribed target position $\boldsymbol{t}$ in the 2D space. Unlike the description in [4], the following MDP explicitly considers the energy costs of motion planning.

A. Deterministic MDP

1) Observation space: The observation space is composed by the state vector and the displacement of the agent's current position to the target and obstacles' positions as follows:

$$\boldsymbol{o}(k) = [\boldsymbol{x}(k)^T, (\boldsymbol{p}_r(k) - \boldsymbol{t})^T, \sum_{n=1}^{N_O}(\boldsymbol{p}(k) - \boldsymbol{p}_{o_n})^T]^T \quad (9)$$

where N_O denotes the number of obstacles.

2) Action space: The continuous action space, $\boldsymbol{a}(k)$, is defined to refine the position command for control-loop control input defined by (2) as follows :

$$\boldsymbol{r}(k) = \alpha_{\text{control}}\boldsymbol{a}(k) + \boldsymbol{t} \tag{10}$$

where α_{control} is a constant determined by the size of the working area. To reduce the overshooting by the action, we limit the range of the action as: $|\boldsymbol{a}(k)| < \mathbf{1}^{2\times 1}$, where $\mathbf{1}$ is a column vector of ones.

3) Step function: With the position command $\boldsymbol{r}(k)$, this function generates the next state $\boldsymbol{x}(k+1)$.

4) Expected reward function: The expected reward function that minimizes the energy consumption for controlling and maximizes the ISNR is given by

$$\begin{aligned} R(\boldsymbol{p}(k)) = \alpha_{\text{obs}} \sum_{n=1}^{N_O} \frac{1}{\|\boldsymbol{p}(k) - \boldsymbol{p}_{o_n}\|^2} + \alpha_{\text{tgt}}\|\boldsymbol{p}(k) - \boldsymbol{t}\|^2 \\ + \mathbb{I}_E \cdot \alpha_{\text{E}} \cdot E + \mathbb{I}_S \cdot \alpha_{\text{CR}} \cdot \gamma_{\text{spectrum}} + p_{\text{collision}} \\ + p_{\text{out\_of\_range}} + r_{\text{reach\_target}} \end{aligned} \tag{11}$$

where the expected reward is inversely proportional to the displacement to obstacles whereas it is proportional to the displacement to the target position. α_{obs} and α_{tgt} are the corresponding penalty parameters for them. In addition, α_{E} is another penalty parameter related with the control energy consumption. They are all real negative. $p_{\text{collision}} = -c$ and $p_{\text{out\_of\_range}} = -c$ denote actual penalties when the agent collides with any of the obstacles and moves out from the working area. $r_{\text{reach\_target}} = c$ denotes the reward when $\|\boldsymbol{p}(k) - \boldsymbol{t}\|^2 \leq r_d$. Furthermore, a positive α_{CR} denotes a parameter related with the ISNR. Two indicator functions $\mathbb{I}_E \in \{0, 1\}$ and $\mathbb{I}_{\text{CR}} \in \{0, 1\}$ are also specified to compare with the case with non-optimized control energy and ISNR.

5) Optimal policy π^*: The objective of the motion planning is to determine its subsequent safe movement meanwhile minimizing the energy consumption used by the agent to control its movement and maximizing the ISNR to access the shared spectrum.

B. SAC

The SAC [19], [20] has been developed under the RL framework with an objective of maximizing the entropy, in which the actor that tries to learn a stochastic policy, described as the distribution over continuous action space, attempts to maximize the expected rewards and entropy. The optimal policy, is modeled as Gaussian over the action space, $\boldsymbol{a}$, with mean and covariance estimated by the neural network. We can summarize SAC as follows:

- The SAC model defines soft functions instead of general functions defined for actor critic. It enables exploration by adding an entropy term to the general function.
- Maximum entropy processing conducts a search to avoid suboptimal local minima. The entropy coefficient can be fixed as a constant, but can be updated through training as well.
- Policy network, $\pi_\phi(\boldsymbol{a}|\boldsymbol{x})$, two value networks, $\{V_\psi(\boldsymbol{x}), V_{\bar{\psi}}(\boldsymbol{x})\}$, Q-network, $Q_\theta(\boldsymbol{x}, \boldsymbol{a})$, are required. Neural networks parameterize $V_\psi(\boldsymbol{x})$ and $Q_\theta(\boldsymbol{x}, \boldsymbol{a})$.
- The value can be obtained from the Q value, but in SAC, a separate value network, $V_{\bar{\psi}}(\boldsymbol{x})$, is provided for the stability of the model. Note that $V_{\bar{\psi}}(\boldsymbol{x})$ is not a train target, but is updated as an exponential moving average from $V_\psi(\boldsymbol{x})$.
- SAC is the off-policy model that uses relay buffer $\mathbb{D}$, expressed as: $\mathbb{D} = \big\{\big(\boldsymbol{p}(k); \boldsymbol{a}(k); \boldsymbol{p}(k+1); R(\boldsymbol{p}(k))\big)\big\}$.
- Value function loss being used to train $V_\psi(\boldsymbol{x})$ is given by (12), where $\bar{\boldsymbol{a}}(k)$ is generated from the policy network.
- Q-function loss being used to train $Q_\theta(\boldsymbol{x}, \boldsymbol{a})$ is given by (13). A batch of experiences is uniformly sampled from $\mathbb{D}$ and used for the expectation. Note that $V_{\bar{\psi}}(\boldsymbol{x}(k+1))$ is used as for the target value instead of $V_\psi(\boldsymbol{x}(k+1))$.
- Policy loss function is given by (14), expressed by means of the expected KL-divergence. In SAC, $\exp(Q_\theta(\boldsymbol{x}(k), \cdot))$ is normalized by $Z_\theta(\boldsymbol{x}(k))$ and regarded as a probability distribution. Thus, the new policy and the KL-divergence of $\exp(Q_\theta(\boldsymbol{x}(k), \cdot))$ is forced to be minimized. As for the value network training, the expectation is computed by the samples from $\mathbb{D}$.
- Target value network is updated periodically.
- Without using the value network, SAC can be implemented via two Q-networks and policy network [21] to reduce the possible overestimation.

IV. Simulation results

We use the following parameters for the evaluation of the proposed approach.

- Dimension of the working area: ± 20 [m]
- $\boldsymbol{Q} = 1\boldsymbol{I}_{4\times 4}$ and $\boldsymbol{R} = 0.1\boldsymbol{I}_{4\times 4}$ for LQR algorithm; and $T = 0.1$ [sec].
- $\alpha_{\text{control}} = 20$; $\alpha_{\text{obs}} = -0.2$; $\alpha_{\text{tgt}} = -2$; $\alpha_E = -0.001$; $\alpha_{\text{CR}} = 0.5$; and $c = 1000$.
- $r_o = 1$ [m]; $r_a = 0.5$ [m]; and $r_d = 1$ [m].
- $\gamma = 0.99$ for the discount factor; $lr = 0.0003$ for the learning rate; $\tau = 0.005$ for a target smoothing coefficient; $\alpha = 0.2$ for the reward scale of entropy regularization, that is, we do not train α; $B = 256$ for the batch size; and Gaussian-based policy for SAC.
- $N_o = 5$; $P_T = 10$ [dB]; $I_p = 1$ [dB]; $N_h = 4$; and $N_g = 3$.
- $\eta = 2.3$; $(\epsilon_{dB})^2 = 6$; and $\eta_S = 1.2$.

For the considered three cases, the proposed safe motion planning makes the mobile agent arrive at the target location successfully avoiding collision with obstacles, PU-RXs. The cases with $(\mathbb{I}_E = 1, \mathbb{I}_{\text{S}} = 1)$, where the control energy consumption and ISNR are computed by applying a joint optimization, have different trajectories over the case with $(\mathbb{I}_E = 0, \mathbb{I}_{\text{S}} = 0)$, where control energy consumption and ISNR are computed without a joint optimization. Furthermore,

$$J_V(\psi) = E_{\boldsymbol{x}(k)\sim\mathbb{D}}\left[0.5\left(V_\psi(\boldsymbol{x}(k)) - E_{\bar{\boldsymbol{a}}(k)\sim\pi_\phi}\left[Q_\theta(\boldsymbol{x}(k),\bar{\boldsymbol{a}}(k)) - \log\pi_\phi(\bar{\boldsymbol{a}}(k)|\boldsymbol{x}(k)\right]\right)^2\right], \quad (12)$$

$$J_Q(\theta) = E_{(\boldsymbol{x}(k),\boldsymbol{a}(k))\sim\mathbb{D}}\left[0.5\left(Q_\theta(\boldsymbol{x}(k),\boldsymbol{a}(k)) - R(\boldsymbol{p}(k)) - \gamma V_{\bar{\psi}}(\boldsymbol{x}(k+1))\right)^2\right], \quad (13)$$

$$J_\pi(\phi) = E_{\boldsymbol{x}(k)\sim\mathbb{D}}\left[D_{\mathrm{KL}}\left(\pi_\phi\left(\cdot|\boldsymbol{x}(k)\right)\middle\|\frac{\exp\left(Q_\theta\left(\boldsymbol{x}(k),\cdot\right)\right)}{Z_\theta\left(\boldsymbol{x}(k)\right)}\right)\right]. \quad (14)$$

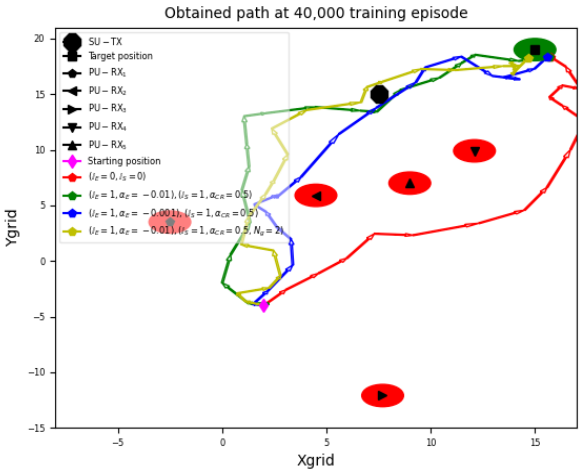

Fig. 2. One example of a determined path determined by a set of $\boldsymbol{p}(k)$s, i.e., $\{\boldsymbol{p}(k), 1 \le k \le K\}$.

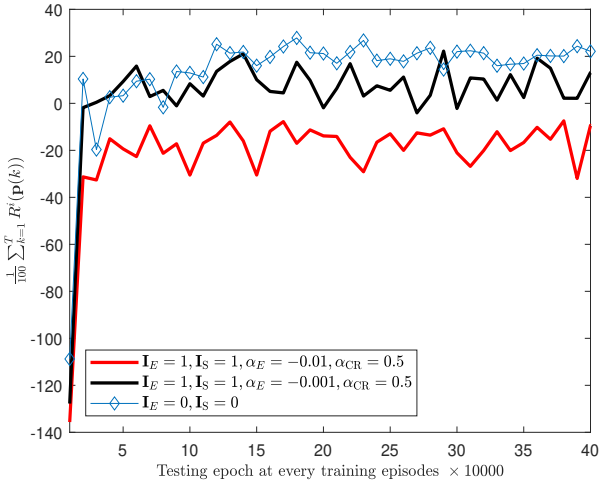

Fig. 3. Average rewards, $\frac{1}{100}\sum_{i=1}^{100}\sum_{k=1}^{K} R^i(\mathbf{p}(k))$, where $R^i(\mathbf{p}(k))$ denotes the received reward for the ith testing episode.

a different value of α_E results in somewhat different trajectory even for the joint optimization. For a different number of multipath components of the SU channel, i.e., $\{N_g = 2, N_g = 3\}$, this figure shows that the agent reaches the target position using the closed-loop RL based control policy.

At every 10,000 training episodes, we have tested the trained model by using 100 independent testing episodes. The average rewards for 100 testing episodes is plotted in Fig. 3 over the determined safe path with one example provided in Fig. 2. This figure shows that both cases with $(\mathbb{I}_E = 0, \mathbb{I}_S = 0)$ has greater rewards than the cases with $(\mathbb{I}_E = 1, \mathbb{I}_S = 1)$ since we give a more penalty for the energy consumption in the reward function, which is expressed by (11). In general, at least 10,000 training episodes is required to have a reliable safe motion planning. Furthermore, as $|\alpha_E|$ increases, the achieved rewards decrease due to a greater penalty for the energy consumption.

In the following two figures, Fig. 4 and 5, we have compared the prediction of the control energy consumption, E, and ISNR.

These two figures show that the SAC-based joint optimization can reduce the control energy consumption and increase ISNR. In particular, more than 10% energy can be saved. In addition, more than three times greater ISNR can be achieved. However, these two figures suggest that a more hyper-parameter tuning for α_E and α_{CR} is required to achieve separate goals of minimizing E and maximizing $\gamma_{\mathrm{spectrum}}$.

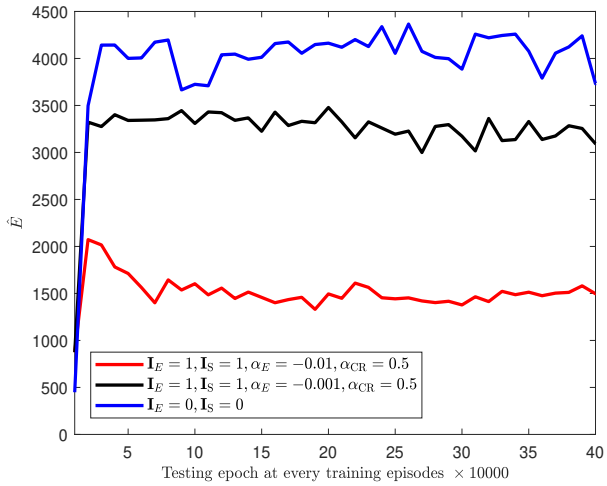

Fig. 4. Control energy consumption, E.

In Fig. 6, the impacts of multipath diversity gain on the ISNR is invested. As was verified by [12], multipath diversity gain exploited over the SU channel, expressed by N_g, influences the $\gamma_{\mathrm{spectrum}}$. In particular, this figure shows that with a less number of multipath components over the SU channel, the SAC-based joint optimization results in a greater $\gamma_{\mathrm{spectrum}}$ over the motion planning that does not employ the

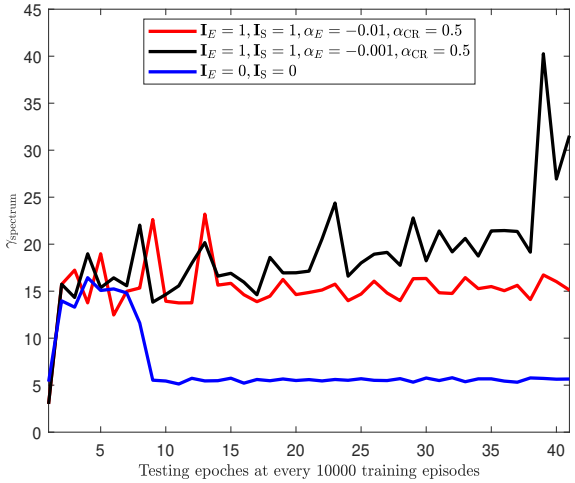

Fig. 5. ISNR, γ_{spectrum}.

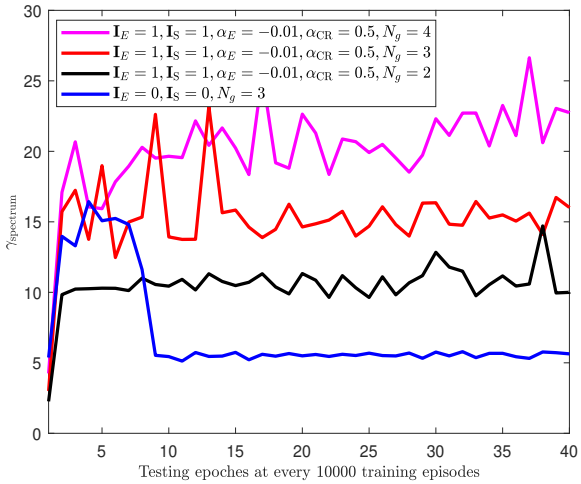

Fig. 6. ISNR, γ_{spectrum}, in terms of N_g, the number of multipath components of the SU channel.

joint optimization.

V. Conclusions and Future Works

In this paper, we have proposed SAC-based safe motion planning for a single mobile agent. To finish high priority tasks, a new spectrum-sharing scheme that support a high reliable movement is integrated into the energy conscious motion planning. Without explicit knowledge of the channel and dynamic environment, the simulation results have shown that the proposed SAC-based safe motion planning can achieve the desired three goals: i) avoiding collision with the static obstacles; ii) minimizing the control energy consumption; and iii) maximizing the ISNR. The future will consider multi-mobile agents coordination and competition to achieve a common goal, in which avoiding an intra-collision with other agents will be a challenging problem. How to minimize the overall control energy consumption and maximize the overall ISNR will be another open problem with an adaptive coordination and competition among the mobile agents.

References

[1] K. C. Chen *et al.*, "Wireless networked multirobot systems in smart factories," *Proc. IEEE*, vol. 109, no. 4, pp. 468–494, 2021.

[2] J. Wan *et al.*, "Artificial-intelligence-driven customized manufacturing factory: Key technologies, applications, and challenges," *Proc. IEEE*, vol. 109, no. 4, pp. 377–398, 2021.

[3] A. Rahman, *et al.*, "Energy-efficient optimal task offloading in cloud networked multi-robot systems," *Computer Networks*, vol. 160, pp. 11–32, 2019.

[4] A. P. Vinod *et al.*, "Safe multi-agent motion planning via filtered reinforcement learning," in *Proc. IEEE Int. Workshop on Robotics and Automation (ICRA)*, Philadelphia, PA, Jul. 2022, pp. 7270–7276.

[5] Y. Yan and Y. Mostofi, "Robotic router formation in realistic communication environments," *IEEE Trans. Robot.*, vol. 28, no. 4, pp. 810–827, Aug. 2012.

[6] A. Benkrid, A. Benallegue, and N. Achour, "Multi-robot coordination for energy-efficient exploration," *J. Control Autom. Electr. Syst.*, vol. 30, pp. 911–920, 2019.

[7] S. Gürel, H. Gultekin, and V. E. Akhlaghi, "Energy conscious scheduling of a material handling robot in a manufacturing cell," *Robotics and Computer-Integrated Manufacturing*, vol. 58, pp. 97–108, 2019.

[8] M. Everett, Y. F. Chen, and J. P. How, "Collision avoidance in pedestrian-rich environments with deep Reinforcement learning," *IEEE Access*, vol. 9, pp. 10 357–10 377, 2021.

[9] L. Lv, S. Zhang, D. Ding, and Y. Wang, "Path planning via an improved DQN-based learning policy," *IEEE Access*, vol. 7, pp. 67 319–67 330, 2018.

[10] H. Qie *et al.*, "Joint optimization of multi-UAV target assignment and path planning based on multi-agent reinforcement learning," *IEEE Access*, vol. 7, pp. 146 264–146 272, 2019.

[11] M. Wen *et al.*, "Private 5G networks: Concepts, architectures, and research landscape," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 1, pp. 7–25, Jan. 2022.

[12] K. J. Kim *et al.*, "Spectrum sharing single-carrier in the presence of multiple licensed receivers," *IEEE Trans. Wireless Commun.*, vol. 12, no. 10, pp. 5223–5235, Oct. 2013.

[13] ——, "QR decomposition-based cyclic prefixed single-carrier transmissions for cooperative communications: Concepts and research landscape," *IEEE Commun. Surveys Tuts.*, vol. 7, pp. 67 319–67 330, 2022.

[14] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal Control*, 3rd ed. John Wiley & Sons, Inc., Hoboken, New Jersey.

[15] Y. Yan and Y. Mostofi, "Robotic router formation in realistic communication environments," *IEEE Trans. Robot.*, vol. 28, no. 4, pp. 810–827, Aug. 2012.

[16] J. Mitola and G. Q. Maguire, "Cognitive radios: Making software radios more personal," *IEEE Personal Commun. Mag.*, vol. 6, pp. 13–18, Aug. 1999.

[17] K. V. Mishra *et al.*, "Toward millimeter-wave joint radar communications: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 36, no. 5, pp. 100–114, Sep. 2019.

[18] A. Aubry *et al.*, "A new radar waveform design algorithm with improved feasibility for spectral coexistence," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 51, no. 2, pp. 1029–1038, Apr. 2015.

[19] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proc. Int. Conf. on Machine Learning (ICML)*, Stockholm, Sweden, Jun. 2018, pp. 1861–1870.

[20] S. Evmorfos and A. P. Petropulu, "Deep actor-critic for continuous 3D motion control in mobile relay beamforming networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May. 2022, pp. 5353–5357.

[21] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018, pp. 1587–1596.