sensors

MDPI

Article

Enhancing Handover for 5G mmWave Mobile Networks Using Jump Markov Linear System and Deep Reinforcement Learning

Masoto Chiputa [1], Minglong Zhang [1], G. G. Md. Nawaz Ali [2], Peter Han Joo Chong [1,*], Hakilo Sabit [1], Arun Kumar [3] and Hui Li [4]

[1] Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland 1010, New Zealand; masoto.chiputa@aut.ac.nz (M.C.); mizhang@aut.ac.nz (M.Z.); hakilo.sabit@aut.ac.nz (H.S.)
[2] Department of Computer Science and Information Systems, Bradley University, Peoria, IL 61625, USA; nali@fsmail.bradley.edu
[3] Department of Computer Science & Engineering, National Institute of Technology, Rourkela 769008, Odisha, India; kumararun@nitrkl.ac.in
[4] Shenzhen Graduate School, Peking University, Shenzhen 518055, China; lih64@pku.edu.cn
* Correspondence: peter.chong@aut.ac.nz

Abstract: The Fifth Generation (5G) mobile networks use millimeter waves (mmWaves) to offer gigabit data rates. However, unlike microwaves, mmWave links are prone to user and topographic dynamics. They easily get blocked and end up forming irregular cell patterns for 5G. This in turn causes too early, too late, or wrong handoffs (HOs). To mitigate HO challenges, sustain connectivity, and avert unnecessary HO, we propose an HO scheme based on a jump Markov linear system (JMLS) and deep reinforcement learning (DRL). JMLS is widely known to account for abrupt changes in system dynamics. DRL likewise emerges as an artificial intelligence technique for learning highly dimensional and time-varying behaviors. We combine the two techniques to account for time-varying, abrupt, and irregular changes in mmWave link behavior by predicting likely deterioration patterns of target links. The prediction is optimized by meta training techniques that also reduce training sample size. Thus, the JMLS–DRL platform formulates intelligent and versatile HO policies for 5G. When compared to a signal and interference noise ratio (SINR) and DRL-based HO scheme, our HO scheme becomes more reliable in selecting reliable target links. In particular, our proposed scheme is able to reduce wasteful HO to less than 5% within 200 training episodes compared to the DRL-based HO scheme that needs more than 200 training episodes to get to less than 5%. It supports longer dew time between HOs and high sum rates by ably averting unnecessary HOs with almost half the HOs compared to a DRL-based HO scheme.

Keywords: millimeter bands; Fifth Generation; handover; deep reinforcement learning; jump Markov linear system

Citation: Chiputa, M.; Zhang, M.; Ali, G.G.M.N.; Chong, P.H.J.; Sabit, H.; Kumar, A.; Li, H. Enhancing Handover for 5G mmWave Mobile Networks Using Jump Markov Linear System and Deep Reinforcement Learning. *Sensors* **2022**, *22*, 746. https://doi.org/10.3390/s22030746

Academic Editor: Adrian Bekasiewicz

Received: 26 November 2021
Accepted: 13 January 2022
Published: 19 January 2022

1. Introduction

Fifth Generation (5G) mobile users need uninterrupted connectivity while consuming large amounts of data and media content when commuting [1]. Millimeter wave (mmWave) bands (i.e., 30−300 GHz on the radio spectrum) hold great potential, enabling 5G mobile users to experience gigabit rates and networks to meet traffic demands. However, a caveat to this is that mmWave communication is very susceptible to topographic and user dynamics. Common materials such as concrete, water, and even human bodies/movements [2] severely alter its cell patterns and ultimately its performance. This level of vulnerability in mmWave bands severely impacts mobility management in 5G mobile networks. To reduce that impact, research on efficient mobility management in 5G mmWave communication continues to gain momentum. In the recent past, 5G mobility management has been explored with machine and artificial intelligence (AI) learning solutions. Some of these include deep

and reinforcement learning (RL) handoffs (HOs). The challenge is that most of the previous HO works [3–6] selected target cells on the basis of initial maximum network performance values. However, the challenge is that the optimum initial value does not always guarantee reliability of the connection after HO. For instance, the selection of mmWave target links based on the highest SINR values [4–7] does not always reveal the reliability of the link after a HO or caching [7] event. In most cases, HOs end up getting executed too early, too late, wrongly, or wastefully. Poor HOs negatively affect the selection of caching points [7] in edge computing too. To that effect, 5G mobile network performance is punctuated with gradual and abrupt changes. To reduce inconsistences in network performance, selection of the best target links requires understanding not just the immediate behavior after HO but also the long-term behavior.

To that effect, we propose an HO scheme that learns not just the immediate behavior of target links but also the likely behavior/pattern post HO. In this regard, we learn to predict the deterioration patterns of potential target links post HO. We use the jump Markov linear system (JMLS) [8,9] and deep reinforcement learning (DRL) to learn the feasible optimal deterioration pattern that chosen target links must adhere to for them to avoid wasteful HO. JMLS is known to account for abrupt changes [8] in system dynamics. We exploit this capability to predict the likely receivable power deterioration pattern of target links at the user. We strategically update the initial JMLS deterioration pattern with online DRL and meta training techniques. Meta training is a technique that reuses similar past training data to make new decisions. This reduces the request for new training datasets when making new decisions in a novel location. At HO, the predicted deterioration pattern of a target link is then compared against an optimal global desired deterioration pattern to understand the reliability of a target link and select the most stable one.

1.1. Related Works

The surging role/potential of mmWave bands in mobile networks such as 5G/beyond cannot be ignored. However, this also applies to its challenges, particularly in the mobility management support of 5G networks. The authors in for [10] instance claimed that higher propagation losses inherent in mmWaves must be addressed to sustain connectivity especially at ranges beyond 100 m and in non-line-of-sight (NLOS) settings. The authors in took four approaches to tackle the crucial problem of distance limitation owing to high spreading loss and molecular absorption that often limit the mmWave transmission distance and coverage range. These were a physical layer distance-aware design, ultra-massive MIMO communication, reflect arrays, and intelligent surfaces. These methods use machine and artificial intelligence (AI) learning for 5G. Various author between [11–31] suggested a move from centralized (used in most 4G systems) to decentralized mobility management algorithms using DRL. DRL in 5G ably learns and builds knowledge about different dynamics of mmWave channels [11–18]. For instance, by interacting with environment data, the authors utilized DRL to observe the available resource at network edges and provide a resource allocation scheme. This enhances user mobility management at the edge given user mobility context, transitions, and signaling exchange [11–26].

Exploiting various actor-critic different DRLs in [11–27], the authors e.g., in [18] proposed to jointly solve offload and resource allocation problems in 5G networks. The authors in [12] used a deep Q-learning-based task offloading scheme to select optimal BSs for users and maximize task offloading utility. In [13], Q-learning was integrated with the mobility robustness optimization (MRO) scheme and mobility-load-balancing (MLB) scheme to tackle traffic load and speed effects in 5G. In [29], a paradigm shift for leveraging time-consecutive camera images in handover decision problems was presented. DRL was used for deciding the handover timings. In [30], a DRL-based approach to solving the problem of joint server selection, task offloading, and handover in a multi-access edge computing (MEC) wireless network was proposed. On the other hand, in [31], HO and the power allocation problem in a two-tier HetNet, consisting of a macro base station and mmWave small base stations, were explored. The author developed a multi-agent reinforcement

learning (MARL) algorithm based on the proximal policy optimization (PPO) method, by introducing centralized training with a decentralized execution framework. However, in all these schemes, highly mobile and dynamic users were hardly considered. Additionally, DRL requires thousands of samples to gradually learn useful policies [15]. Furthermore, DRL becomes terribly unstable/stochastic when learning systems with large local variances [16]. Thus, to guarantee continuous connectivity for 5G mobility, i.e., by not just satisfying channel input/state bounds but also considering abrupt and continuous disturbances, control approaches using Markov systems have been proposed in the literature. For instance, the authors in [20] used JMLS with expected maximization (EM) to predict abrupt deterioration behavior. Predictions were then enhanced using Viterbi algorithms. The Viterbi algorithm, however, requires accurate channel state information (CSI) to converge. In such cases, the authors in [26] argued that inaccurate training gradually cripples the accuracy of predictions, particularly at low signal-to-noise ratios (SNRs). To that effect, it was combined with meta data training, making the Viterbi proposed approach more reliable and less dependable on the changing and accuracy of the data. In [18], to tackle a distributed decision-making scenario, the author extended the JMLS formulation into game theory. Similarly, the authors in [17] incorporated particle-filter-based RL in JMLS to predict a finite number of disturbances within a randomly chosen sample of trajectories. This allowed the scheme to track/adjust to time-varying conditions in real-time. It is worth reiterating that none of the mentioned works analyzed the deterioration pattern of mmWaves to make an HO decision or utilized multiple users with very different levels of impact on mmWave propagation characteristics; they were all designed to operate in a single frequency band or with one user type.

1.2. Contributions

- We propose to use the JMLS to model the deterioration behavior/patten of mmWave target links and the formulation of HO policies for 5G mmWave networks. Given JMLS's ability to account for abrupt changes [7], we analyze the pattern and learn to predict the extent of abrupt performance changes in the chosen target mmWave links before HO.
- We use DRL to update and optimize JMLS deterioration pattern predictions and learning. To help reduce training samples and, thus, have ample time to track pattern changes of rapid-varying channels in real time, we propose using meta learning techniques. Meta learning is a technique that automatically reuses training data from related past tasks or neighbors to make a new decision. This reduces the need for a new CSI/training dataset to make new decisions.
- We use the Kaiser–Meyer–Olkin (KMO) [25] test to measure the expected divergence of target links from the optimum deterioration pattern post HO to know their reliability in advance.

1.3. Organization

The remainder of this paper is organized as follows: Section 2 describes the proposed framework and its operation. The section further describes the resource allocation and optimization problems; Section 3 present adoptions of the JMLS–DRL solution; Section 4 analyzes the simulation results; Section 5 provides the conclusion.

2. Proposed Framework

We propose to use the likely received power pattern supplemented with SINR values to determine the best mmWave target cell/link. We first learn to predict and then analyze the received power deterioration pattern for four different types of users with respect to mmWave BSs. The four types of users are cars, pedestrians, cyclers, and e-bikers. For each user type, prior to HO selection, the scheme learns the likely mmWave user received power deterioration pattern given the effects of speed, topography, and channel state. The best target link is one whose likely deterioration pattern with distance is gradual

and follows the global deterioration pattern generated from aggregative data samples from multiple mmWave BSs. The received power deterioration pattern is modeled using JMLS. It models how likely received power will deteriorate for a user given the NLOS and distance effects on the mmWave channel. Thus, in the first instance, the model learns and determines the desired optimal received power deterioration patterns for different user types using expected maximization (EM) [9]. EM automatically infers missing values of the link deterioration pattern over some states. Even though EM is robust, dynamic channel changes are not anticipated [10]. The EM estimations are, thus, optimized using DRL and meta training techniques.

Meta learning is loosely defined as an automatic learning and adaptation mechanism that improves accuracy by typically acquiring training from related tasks/users. The scheme only requires new training samples when the prediction error is bigger than the assumed predicted threshold. At HO, we have two deterioration patterns to consider: a global deterioration pattern formulated with aggregative data from all mmWave BSs, and a current local deterioration pattern formulated using local/individual BS channel data. Owing to the large data variance analyzed, the global pattern is regarded to be more accurate.

Thus, at HO, KMO test index values are used to determine the similarity levels between the global and local deterioration pattern for target links whose SINR is above the threshold. The level of divergence between the target link's deterioration behavior and global pattern determines how reliable the target link is post HO. This is vital because mmWave links have a tendency of deteriorating from excellent to very poor performance immediately after HO. Thus, understanding the long-term connectivity endurance post HO is paramount for a reliable connection.

2.1. Manhattan Grid Mobility Model

A Manhattan grid model is used to model the road network with streets and intersections (as shown in Figure 1) in an urban scenario. The road network area is 500 m $\times$ 100 m. We have four types of users, distributed evenly: pedestrians with speeds of 1.4 m/s, cyclers with speeds of 3–7 m/s, e-bikers with speeds of 8–9 m/s, and cars with speeds of 10–14 m/s. Cars within 3 m of each other adjust velocities every 3 s by 1–3 m/s to avert crashes. Each street consists of right and left lanes for each user type.

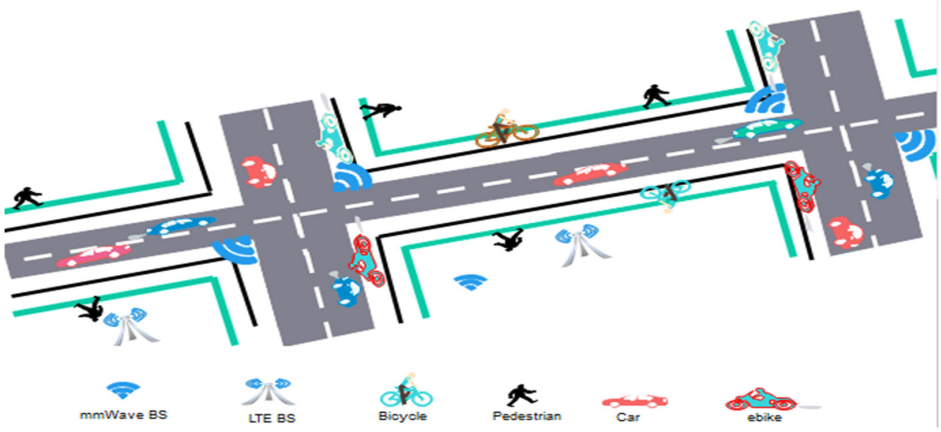

Figure 1. Multiuser type mobility model.

Given user directions, i.e., $\mathfrak{y}$ ={moving toward/away from a mmWave BS}, users traverse different streets. The probability of recovering the channel link just after being blocked $\mathbb{P}_{\mathfrak{y}r}$ and of remaining blocked $\mathbb{P}_{\mathfrak{y}b}$ is expressed as follows [12]:

$$\mathbb{P}_{\mathfrak{y}r} = \frac{\mathfrak{y}}{K}\sum_{i=1}^{k}\frac{T^r}{T^r + T^b}, \tag{1a}$$

$$\mathbb{P}_{\mathfrak{y}b} = \frac{\mathfrak{y}}{K}\sum_{i=1}^{k}\frac{T^b}{T^r + T^b}, \tag{1b}$$

where K is the total number of samples, whist $k \in K$ is the number of possible blockings; T^b and T^r are the mean nonblocking and blocking windows within a transmission range d. The rate of channel links switching from blocked to recovered and vice versa within d is $1/T^r$ and $1/T^b$, respectively. Accordingly, $\mathfrak{y}$ is binary and assumed 1 when the users are moving toward a target BS. Otherwise, $\mathfrak{y}$ is assumed to be 0, as the recovery of reconnection over the serving cell is minimal if user is moving away. The argument is that link recovery chances are high if a user is moving toward the direction of mmWave BS.

2.2. Outage Probability

Assuming that Θ is a set of optimization parameters for a given access policy π, the outage probability P_π for the observable set of signals Y_k can be defined as follows [2,11]:

$$P_\pi(Y_k|\Theta) \triangleq P\Big(\sum_l \sum_{s_k} b_l \log_2(1+\gamma_t(x)) \geq r^{m\zeta}(\hat{\gamma}_t)\Big), \tag{2a}$$

where γ_t and $\hat{\gamma}_t$ are the measured and target SINR, respectively, and $r^{m\zeta}$ is the targeted data rate given channel state $s_t \in S$. b_l is the bandwidth for the given channel link l. We assume that all mmWave BSs directionally transmit equal maximum power P, and that all users have a receiver sensitivity of x_{kmin}. Thus, each serving mmWave BS (with either LOS or NLOS link) given, P, must satisfy the average received power of at least x_{kmin}. Moreover, given a threshold x_{k0}, where $x_{k0} > x_{kmin}$, any user–mmWave BS link that requires transmit power that exceeds P or does not meet x_{k0} will not be established or lose connection, i.e., such a connection experiences a truncation outage at a given distance $d = \left(\frac{P}{x_{k0}}\right)^{\frac{1}{\propto^{k_L}}}$ despite satisfying Equation (2). α is the path loss exponent in LOS and NLOS pathloss exponents [25]. Equally, given the cutoff threshold x_{k0}, LOS and NLOS users located at distances beyond $\left(\frac{P}{x_{k0}}\right)^{\frac{1}{\propto^{k_L}}}$ and $\left(\frac{P}{x_{k0}}\right)^{\frac{1}{\propto^{k_{NL}}}}$, respectively, from the target BS are unable to communicate owing to insufficient received power x_t. The data rate is defined as

$$r^m = b\ \log_2\ \left(1 + \frac{P\left|h^H p\right|^2}{(1+d^\propto)}\ F_x\left(\left|\theta_k^l\right|\right)\right)\ , \tag{2b}$$

$$\varphi_k^l(\cdot) = \frac{1.4\times 10^4}{f_c(GHz)\cdot v(km/h)}, \tag{2c}$$

where $\theta_k^l = \frac{2d\sin\varphi_k^l}{\lambda}$ is the normalized central angle of arrival for beam p, v is user velocity under 50 km/h, f_c is the carrier frequency. $\left|h^H p\right|^2$ is channel gain. and $F_x\left(\left|\theta_k^l\right|\right)$ denotes the Fejér kernel value. As user speed approaches zero and $F_x\left(\left|\theta_k^l\right|\right) \to 1$, SINR approaches the maximum. F_x approaches 0 as v increases [4].

2.3. Resource Allocation Problem

The minimum rate $\mathbb{R}^m$ requirement problem given outage and power constraints at d from a BS is defined as

$$\max_{\Theta}\sum_t \sum_{S_t,\, l}\Big(1 - \mathbb{P}_{\mathfrak{y}b}\Big(P_\pi^{m|x_t} + P_\pi^{m|u_t}\Big) r_l^m(y)\Big) \geq \mathbb{R}^m, \tag{3a}$$

where $P_\pi^{m|x_t}$ and $P_\pi^{m|u_t}$ are the LOS and NLOS conditional outage probability for a user in the m-th state, respectively. r_l^m is the maximum attainable data rate at user–BS distance d.

The target receivable power x_{t+1} at d needed to meet $\mathbb{R}^m$ in condition (2a) given outage constraints (1a)–(2b) is proposed in Equation (3b).

$$x_{t+1} = \max \sum_{x_t, u_t} \left\{ \frac{\hat{\gamma}}{\gamma^{min}} x_t - \frac{\alpha {x_t}^2}{\beta \hat{\gamma}^2} \right\}, \tag{3b}$$

where $\{.\}^+ = \{max, 0\}$. x_t is the current received power in LOS. $\hat{\gamma}$ and γ^{min} are the targeted and measured SINR needed to satisfy $\mathbb{R}^m$. It must be noted that, if there exists an infeasible SINR target in a certain user state, the resulting power demand, x_{t+1}, by users may diverge to infinity. This is due to each user link attempting to meet its own required SINR no matter how high the power consumption can be. Thus, α and β are power and SINR scaling factors, respectively, to substantially enhance reasonable deviations of x_{t+1} in NLOS. The corresponding energy consumption for a given x_{t+1} is as follows [24]:

$$E_c = \beta \left\{ x_t \delta \frac{c(t-w)}{\mathbb{R}^m} + e_0 * \zeta c(t-w) \right\}, \tag{3c}$$

where β denotes the price per unit energy consumption, $c(t-w)$ denotes the actual number of packets received by the user at t during window w, $c(t-w)/\mathbb{R}^m$ is the latency, x_t is the current received power at time t, e_0 is the unit energy per packet, and $e_0 * \zeta c(t-w)$ denotes the energy lost due to lost packets (expected number minus the actual number of received packets) at t during window w. Given receivable x_t and transmittable power P constraints (see Section 2.1), for optimum packet delivery latency, the maximum link utility problem is formulated as follows [19]:

$$\max \sum_{P} \sum_{x_t,} \{ x_{t+1} \delta c(t-w) - \zeta P E_c \}, \tag{3d}$$

where δ is the expected latency scaling factor given x_{t+1} within w. ζE_c is the latency discrepancy following a change from x_t to x_{t+1} as the user moves away from the serving BS. We learn to predict the long-term deterioration pattern $\{x_t, \ldots, x_T\}$ of the target links to ascertain its reliability in meeting the desired data rate prior to the next HO. We utilize JMLS properties to predict the likely gradual/abrupt deterioration behavior of target links [7].

3. JMLS System Definition

We first reformulate the resource allocation problem in Equation (3a–d) into a JMLS learning form with system state, action, and reward defining the deterioration pattern.

3.1. The JMLS Representation

We propose the deterioration pattern learning algorithm and JMLS by describing Equation (3a–d) as follows:

$$\begin{cases} x_{t+1} = A(s_t) x_t + B(s_t) u_t + w_t \\ y_t = r^{min}(s_t) x_{t+1} + v_t, \\ \mathcal{M} = (\Theta,\ P(S), \pi, P_\pi) \end{cases} \tag{4a}$$

where $x_t \in X$ is the current received power in the LOS given state s_t, and $u_t \in \mathcal{U}$ is the estimated received power discrepancy due to blockage/NLOS effects. It is related to x_t by $u_t = -Kx_t$ where K is the control factor of the power and SINR scaling factor in Equation (3b); $A(s_t)$ and $B(s_t)$ are the SINR/power coefficient matrices in Equation (3b). $v_t \sim \mathcal{N}(0, \mathcal{Q}(s_t))$ and $w_t \sim \mathcal{N}(0, R(s_t))$ are the data rate and received power measurement noise, respectively. Measurement noises are influenced by the competing effect of change in gain, angular and linear transmission distance, user speed, etc. for the same SINR requirements (see Equation (3a–c)). s_t denotes a state governing for parameter

set $\Theta = \{A,\ B,\ R, r^{min}, \mathcal{Q}, P(S)\}$. s_t belongs to a set of Markov stochastic decisions $\mathcal{M} = \{m_1, m_2, \ldots, m_M\}$, and m_M determines which state is active at time t.

$$s_t = \{v, r_t, T_t, d_t, \eta_t\}, \tag{4b}$$

where $v = [v_1, \ldots, v_T]$ is a vector of user velocity, $r_t = [r_1, \ldots, r_t]$ is a vector of possible user data rate, $T_t = [t_1^m, \ldots, t_N^m]$ is a vector of average service time, $d = [d_t^m, \ldots, d_T^m]$ is a vector of transmission distances with the same SINR, and $\eta = [\ \eta_1, \ldots, \eta_N]$ is a vector of user direction in the n-th sample.

Following a transition to x_{t+1}, the immediate reward $r^{min}(s_t)$ for the observed signal $y_t \in Y$ is defined as a function of energy efficiency.

$$r^{min}(s_t) = \frac{r^m(s_t, a_t)}{P_t}, \tag{4c}$$

where $r^m(s_t, a_t)$ is a data rate greater than $\mathbb{R}^m$ in Equation (3a). The likely rate discrepancy between a user and mmWave BS is expressed as

$$\mathcal{Q}(s_t) = \delta_k \frac{(P_t - x_t) r^m(s_t, a_t)}{P_t}, \tag{4d}$$

where δ_k is the scaling factor of the rate discrepancy for each state, s at time t given maximum rate $r^m(s_t, a_t)$. The transition probability between states with x_t and x_{t+1} is

$$P(S) \triangleq P\big(s_{t+1} = m_j \big| s_t = m_i\big). \tag{4e}$$

Assuming that N samples from different mmWave BSs at time t are collected within each window w and arranged in ascending order of the users' distance from serving BSs, the transmission energy cost function is defined as follows:

$$\mathcal{J}(x_t) = E\Big\{\sum_{j=1}^{N} \|x_j\|^2_{\mathcal{Q}(s_t)} + \sum_{j=0}^{N-1} \|u_t\|^2_{R(s_t)}\Big\}, \tag{4f}$$

where first and second factors in Equation (4f) represent the sum-weighted norm energy cost for received packets and lost packets over $X_N = \{x_0, \ldots\ x_N\}$, respectively. $\mathcal{J}(x_t) \triangleq \sum_{j=1}^{N} E_c$.

3.2. Initial Deterioration Path Training

Y_T, X_T, and S_T denote a sequence of observed data rates $\{y_1, \ldots, y_T\}$ over corresponding receivable power values $\{x_0, \ldots\ x_T\}$ and $\{s_1, \ldots, s_T\}$ states until time T. The JMLS learning problem in each user type is to define the likely sequence X_T and parameter Θ that maximize the likelihood function $P(X_T|\Theta, Y_T)$ given a finite observation in Y_T over S_T for all $k_1, \ldots, k_2 \in\ T$ at distance $k_1 \leq k_2$. The initial deterioration pattern estimator upon which we design our framework for received power pattern X is the EM algorithm in [12]. EM uses Bayesian inference to automatically infer the optimal value set of Θ for X_T [12] at each step k, as seen in Figure 2; the value function can be written as

$$\mathbb{Q}\Big(\Theta|\Theta^k\Big) = \mathbb{E}\Big[\log P(\,X_T, S_T, Y_T|\Theta)\Big|Y_T,\ \Theta^k\Big], \tag{5a}$$

such that

$$\Theta^k = \arg\max_{x\ \in\ X} \mathbb{Q}\Big(\Theta\Big|\Theta^k\Big), \tag{5b}$$

where $\Theta^{(k)}$ is the current parameter estimate at iteration k. The change, $\Delta Q(s_t)$, between s_k and s_{k+1} states must satisfy condition (5c) to avoid abrupt changes or shocks in data rate.

$$|\mathrm{Q}(s_{k+1}) - \ \mathrm{Q}(s_k)| < \mu(k)^v\ , \tag{5c}$$

where $\mu(k)^v$ is the averaged data rate discrepancy between states s_k and s_{k+1} for a user with velocity v. In Equation (5c), a smaller difference denotes a lower change between x_k and x_{k+1}, defining deterioration pattern X_T. $\mathcal{Q}(s_{k+1})$ and $\mathcal{Q}(s_k)$ can be chosen independently. Obtaining full or accurate CSI to determine the pattern may be difficult owing to rapid changes in mmWave channels. Furthermore, EM cannot handle such switching dynamics [12]. Thus, instead of recomputing the steps in Equation (5a–c) to refine pattern X, as more CSI about Y_T is obtained, we use online DRL with EM estimations of X as the initial experience to determine user target data rates Y_T, as seen in Figure 3.

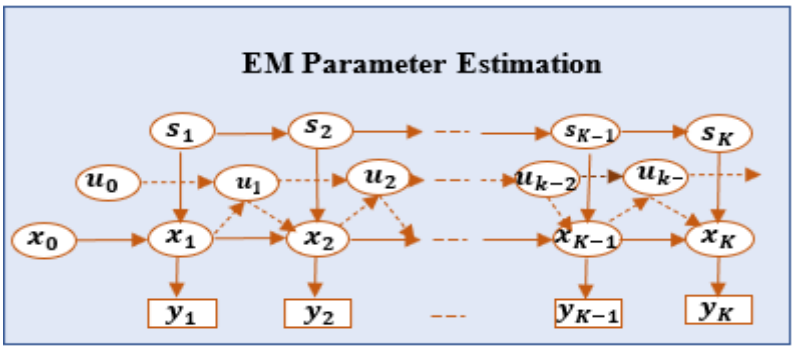

Figure 2. Dynamic Bayesian representation of JMLS composed of three variables and related deterioration variables over adjacent timesteps *K*.

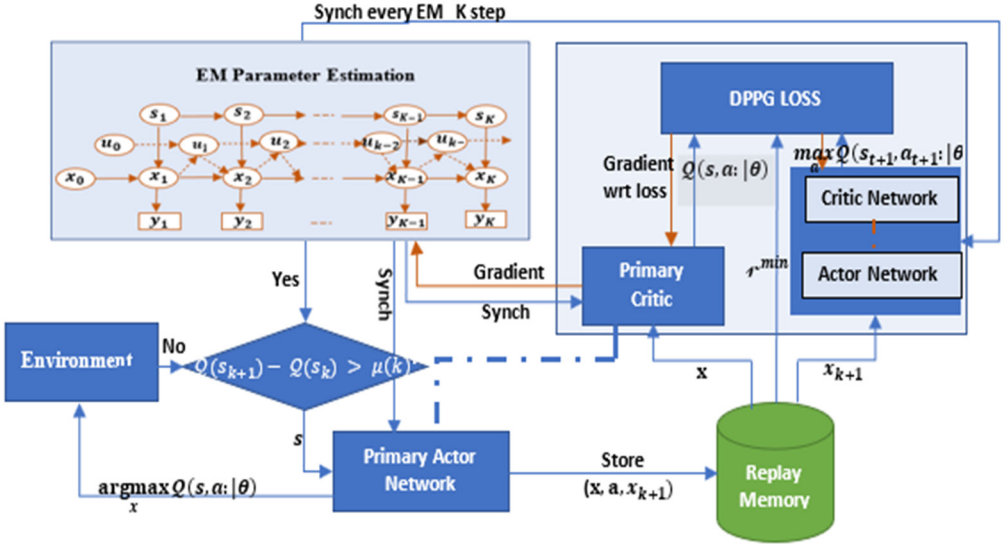

Figure 3. Deep deterministic policy gradient (DDPG) algorithm structure.

3.3. Deep Reinforcement Learning in EM-Estimates

As seen in Equation (5b), EM estimates the maximum obtainable received power x, i.e., the upper bound of desirable received power in each state needed to obtain a high SINR, $a \in A_x$, and, hence, data rate, y, efficiently. The role of DRL, given optimum maximum receivable power $-x^*$ per state, is to determine the minimum/lower bound of receivable power x^* needed to obtain the same JMLS value, $a \in A_x$, efficiently about the same state. It must be both noted and emphasized that the power at the receiver can randomly vary with time, space, and frequency. This may trigger erroneous reception at the receiver. Rectifying or averting the errors may need a high transmit power (which is energy-inefficient and is beyond the limit) to meet the desirable receivable power and receive the same amount of user data within a given QoS/SINR requirement. However, if the gain of the channel is high in the peak, even if the received power is lower (e.g., in NLOS), this permit using lower receivable power to receive the same/similar amount of data while maintaining

the same given QoS/SINR. Thus, knowing the pattern of noy only the maximum but also the minimum receivable power prior to the HO decision is vital. Hence, DRL is used to determine minimum desirable power given the maximum by EM estimation. Here, DRL uses EM data as the initial experience (meta data) to determine the least expected receivable power needed to give $a \in A_x$. In that case, the DRL agent has to consider only the SINR value, $a \in A_x$, possible for $-$x in EM and find the power x that gives the highest directly obtainable reward plus expected accumulated future reward of the resulting states s. The EM's Q value for the $(-x,\ a)$ pair is used as meta data by the agent to find the SINR that gives the smallest DRL Q value with a function value V. The optimal value function V^* is obtained by solving x_{ko} for each given $-x_{ko}$ in Figure 2.

$$V^*(x_{ko}) = \max_{\pi} \mathbb{E}\left\{ \nabla^{min}(-x^*{}_t, a(s_t), x^*{}_t)\Big|_{s_t, \pi(x_t|\theta\pi)} \right\}. \tag{6a}$$

Technically, for a given optimum pattern $-X^*$ in Equation (5c), the algorithm uses corresponding optimized parameter sets $\theta\pi$ and policy $\pi(s_t|\theta\pi)$ as input to DRL. The DRL scheme then determines the minimum desirable value x_t needed to achieve $a(s_t)$. It uses corresponding maximum value $-x_t$ determined by EM in each state s_t as the initial experience and improves it by minimizing the expected energy cost, $\mathcal{J}(x_t)$. The policy $(s_t|\theta\pi)$ is defined as

$$\pi = \underset{\mathcal{J}}{argmin}\left\{ \mathcal{Q}(a_t, x_t|\theta\pi) + \varepsilon \sum_{s_t \in S} P_\pi(x_t|-x_t, a_t)\mathcal{J}^*(x_t) \right\}, \tag{6b}$$

where $P_\pi(x_t|-x_t, a_t) \rightarrow\ [0,\ 1]$ denotes the probability of transition from $-x_t$ to x_t without change, $a \in A$, with least possible energy cost $\mathcal{J}^*(x_t)$, in s_t. The optimal policy π derives the smallest possible value of $\mathcal{Q}(-x_k, a_k, x_k|\theta\pi)$; hence, $\mathcal{J}^*(x_t)$ in Equation (4f) satisfies the following Bellman equations:

$$\mathcal{J}^*(x_t) = \text{if } s^* \in S \vee x_t, \text{ else} \tag{6c}$$

$$\mathcal{J}^*(x_t) \triangleq \min_{x} \mathbb{E}\left[r^{min}(a(s_t), x_t^*) + \sum_{x_t \in X} P_\pi(x_t|-x_t, a_t)\mathcal{J}(x_t^*) \right], \tag{6d}$$

where s^* are goal states where condition (5c) is satisfied.

3.4. Deep Deterministic Policy Gradient (DDPG)

We use the deep deterministic policy gradient (DDPG) to improve the accuracy of the pattern. DDPG is combined with DQN on the premise of the EM algorithm in order to further enhance the stability and effectiveness of network training. This makes it more conducive to solving issues of continuous state and action space. Technically, DDPG uses DQN as the experience replay memory and the target network to solve the problem of nonconvergence to approximate the EM function values in neural networks. It is, thus, an actor-critic and model-free algorithm. It learns policies using highly dimensional observation and action spaces. In this respect, agents use three modules: primary network, target network, and replay memory.

Primary networks match actions (SINR ratios in JMLS parameter sets) with expected received power using a policy gradient method. It consists of two deep neural networks, namely, primary actor and primary critic neural networks. On the other hand, the target network sets target values y_t for the optimal receivable power x_t with pattern X given by EM estimations. The replay memory stores the tuple experience from EM Bayesian estimators and environment via the actor network given condition (5c). Experience tuples include the current and next state, the SINR ratio value following the transition between states, and the reward for choosing the received power level in X_T. Replay memory updates

are randomly sampled for training the primary critic network and setting the target in the target network for the eventualities in Equation (5c).

Given EM parameter set θ and policy $\pi(s_t|\theta\pi)$, the cost policy gradient $\nabla_{\theta\pi}\mathcal{J}$ gives the values of $x_t \in X_T \forall\ y_t$ with a minimum change in $\nabla_{\theta a}\mathcal{Q}(a_t, x_t|\theta\pi)$ between $-x_t$ and x_t, and the corresponding maximum change $\Delta\nabla^{min}(-x_t, a_t, x_t, s_t)$ for each value x_t transitioning from $-x_t$ is defined as

$$\nabla_{\theta\pi}\mathcal{J} \approx \max_{\pi}\mathbb{E}\Big[\Delta\nabla^{min}(a_t, s_t)\Big|_{s_t,\pi(x_t|\theta\pi)}\nabla_{\theta a}\mathcal{Q}(a_t, x_t|\theta\pi)\Big]. \tag{7a}$$

The optimal value $\mathcal{J}^*(x_t)$ gives the highest possible expected future reward and lowest discrepancy from target values for each state. The policy gradient is explored by the primary actor neural network, and the value function $\mathcal{Q}$ for the (x, a) pair is used by the agent to find the SINR ratio a and received power x that gives the lowest $\mathcal{Q}$ value and highest reward. Value iteration in DDGP terminates when $\forall$s $\in$ S, $|\mathcal{J}_k(\mathrm{x}) - \mathcal{J}_k(-\mathrm{k})| \leq \varepsilon$, and termination is guaranteed for $\varepsilon > 0$. ε is similar to a greedy strategy with probability $1 - \varepsilon$ [27]. Here, ε decays as more iterations (and, hence, more experience) are gained. The primary critic network updates θa by minimizing loss function $Ls(\theta\pi)$, which is defined as

$$Ls(\theta Q) = \mathbb{E}(\hat{y}_t\ -\ \mathcal{Q}(a_k, -x_k|\theta\pi)), \tag{7b}$$

where $\hat{y}_t$ is the target network value and can be obtained by

$$\hat{y}_t = \nabla^{min}(a, x_t) + \varepsilon\mathcal{Q}^k(x_k, \pi^k\Big(s_{k+1}\Big|\theta_\pi^T\Big)\Big|\theta_a^T\Big). \tag{7c}$$

Here, $\varepsilon\mathcal{Q}^k(x_k, \pi^k\big(s_{t+1}\big|\theta_\pi^T\big)\Big|\theta_a^T\Big)$ is obtained through the target network, i.e., the network with parameters $\theta\pi$, from EM with $-$X values and θa from X generated over time for minimum desirable receivable power. The new values of Equation (5c), i.e., patterns, are updated by minimizing loss in Equation (7b). The gradient of $Ls(\theta Q)$ over X_T is calculated by its first derivative, which can be denoted as in [14].

$$\nabla_{\theta\pi}Ls(\theta Q) = \mathbb{E}(2(y_t - \mathcal{Q}(a, x_t|\theta\pi)\nabla_{\theta a}\mathcal{Q}(a, s_t|\theta\pi))). \tag{7d}$$

According to Equation (7d), the parameter $\theta_\mathcal{Q}$ of the primary critic neural network can be updated. Specifically, at each training step, a mini-batch experience $\langle\ s_t,\ a_t, R^{imm},\ s_{t+1}\rangle$, $t \in \{1,\ \ldots,\ k\}$ is randomly sampled from replay memory. For each point in X_K, the target network value is regarded as the previous and current version of EM parameters $\theta_\pi^T$and θ_Q^T. At each iteration, θ_π^T and θ_Q^T in Equation (7c,d) are updated with a weighted combination of the previous state. The prediction of target path takes the form of a weighted combination of the following models:

$$\begin{aligned}\theta_\pi^T(\widetilde{x}_k) &= \mathcal{w}\theta_\pi(-\widetilde{x}_k) + \Big(1 - \mathcal{w}\Big)\theta_\pi^T(-\widetilde{x}_k),\\ \theta_\mathrm{Q}^T(\widetilde{x}_k) &= \mathcal{w}\theta_\mathrm{Q}(\widetilde{x}_k) + \Big(1 - \mathcal{w}\Big)\theta_\mathrm{Q}^T(\widetilde{x}_k),\end{aligned} \tag{7e}$$

where $\omega\ \in\ [0,\ 1]$ is the weight computed using a Gaussian kernel parameterized by the transmission distance metric $d_k \in\ \widetilde{s}_k$.

$$\mathcal{w}_k = \exp\Big(-0.5(x - \mu_k)^T d_k(x - \mu_k)\Big). \tag{7f}$$

Target neural networks generate target or ideal values for training and reoptimizing the deterioration pattern X_T from $-X_T$ on the basis of EM and replay updates. Thus, EM estimations in each iteration are used as meta data for DDPG. The target neural network has a similar network structure to the primary network, i.e., similar neural network structure and initialization parameters. In the training process, the parameters of the target actor and critic networks are updated slowly (soft replace) by EM estimated values. Here, instead

of directly and randomly training parameters of the primary actor and critic networks to further enhance the stability of the training process, we copy EM estimations as ideal initial values. Replay memory stores EM experience tuples, thus formulating X_T, and each value update $x_t \in X_T$ includes a tuple $\langle -x_t,\ a_t, R^{imm},\ x_t \rangle$ update.

Figure 3 shows the structure of the proposed JMLS–DDPG algorithm. The DDPG algorithm takes the EM parameter dataset and maximum receivable power values $-$X as initial input to determine the minimum receivable power values of a pattern. Given that the power effects on SINR can be reduced in high-channel-gain locations, afterward, the DDPG agents output the minimum receivable power values X needed to maintain the same SINR ratio previously predicted and set by EM estimations for S_K. The corresponding reward of x_k in EM is copied, and the SINR that is beneficial to the agent to achieve the goal gives a positive reward; on the contrary, it gives a negative reward if condition (5c) is not fulfilled. The current state information, the SINR ratio, the reward, and the state information of the next minimum desirable receivable power are stored in the replay pool. Meanwhile, the neural network trains the experience and continuously adjusts the SINR strategy by randomly extracting sample data from the EM pool, and it uses the gradient descent approach to update and iterate network parameters, so as to further enhance the stability of pattern X and the accuracy of the algorithm. Using EM experiences as initial training data input to DDPG restricts the search range for optimal minimum receivable power values. Thus, any observed mmWave BS data rate not meeting the corresponding receivable power is immediately discarded for training or consideration. This in itself technically reduces the training sample for DRL and, hence, convergence time. Ultimately, the improved DRL HO is obtained by combining DDPG with EM predictions acting as a meta training sample. Finally, the pattern model is integrated into the HO platform for HOs.

3.5. Online Update of Target Deterioration Path

DDPG subdivides the training network structure into an online network and target network (see Figure 3). The online network is used to output the minimum expected received power in real time, evaluate SINR ratio values, and update network parameters through online training, which includes the online (primary) actor network and online critic network. The target network includes the target actor network and target critic network, which get updated by EM values. The target actor network system, however, does not carry out online training. For each user type, the estimated path X_N is only re-estimated from new training samples when the pattern prediction error based on EM estimates is too much larger than the minimal desired received power pattern. It, therefore, follows that, when the error given the energy efficiency is small enough such that the channel gain compensates for the power loss to maintain the desired SINR, the corresponding EM information used to generate received power pattern X_t is regarded to provide reliable training sample for the target network in DDPG. EM data are, thus, re-encoded to generate new training samples for the DRL and to set new targets over $\widetilde{S}_t$, a process henceforth referred to as meta-training. If indeed the pattern of link deterioration is successfully followed by the target mmWave network, then $\widetilde{X}_t$ represents the true channel link deterioration behavior from which Y_t is obtained. Consequently, the corresponding pair $\widetilde{S}_t$ and Y_t parameter set $\theta_\pi(s_t)$ can continue being used to retrain DDPG instead of requesting new CSI from the environment in Figure 3. The model can be efficiently and quickly retrained with a relatively small number of new training samples. A natural drawback of decision-directed approaches such as the Bayesian in EM is their sensitivity to decision errors. For example, if the link fails to successfully sustain connectivity, then the meta training samples $-\widetilde{X}$ of $\widetilde{X}$ over $\widetilde{S}_t$ do not accurately represent the channel behavior results in Y_t. In such cases, the inaccurate training sequence may gradually deteriorate the accuracy of DDPG predictions, making the proposed approach unreliable, particularly in low-SINR areas where link deterioration pattern errors occur frequently. Nonetheless, when pattern errors are less frequent in EM, the effects of decision estimate errors of ε, i.e., the number of errors in a pattern, can be

used to decide when to generate meta training. For instance, we retrain with new training samples in DDPG only when the number of errors is larger than some threshold. Using this approach, only accurate meta training data are used, and the effect of decision errors is controlled. When using new training samples, we cleverly focus attention on states with non-converged pattern values, i.e., where Equation (5c) is not fulfilled. Our online training mechanism is summarized in Algorithm 1. The Workflow in Figure 4 summarizes the steps in Algorithm 1. In particular, EM estimates the initial receivable power pattern. If, however, the date rate discrepancy condition is not satisfied in Equation (5c), DDPG in Figure 4 is conditionally evoked to improve the prediction of the target link deterioration pattern when EM fails to meet the data rate condition in Equation (5c). DDPG, as earlier alluded to, cleverly uses the maximum SINR to find the minimum expected receivable power of each state defining the deterioration pattern.

Algorithm 1: JMLS–DRL-Based Pattern Algorithm.

Input: User mobility model parameters, $\mathbb{P}_{\mathfrak{y}}$. v
Parameters about DC communication: transmission power limits, bandwidth, channel gain, and NLOS and LOS path loss exponent.
Observed states S; Set of observed signals $Y = [y_1, y_2, y_3, \ldots, y_N] \in \mathbb{R}$,
Output: mmWave deterioration path $X = [x_1, x_2, x_3, \ldots, x_N]$ for target link

1. Initialize the deterioration path estimations
2. **for** $t = 1$ **do**
3. Draw y_t for JMLS parameter estimation Θ, where $(X_T, S_T, Y_T|\Theta)$
4. **Estimate Maximization (EM):**
5. $$\mathcal{Q}\left(\Theta\middle|\Theta^k\right) = \mathbb{E}\left[\log P(X_T, S_T, Y_T|\Theta)\middle|Y_T, \Theta^k\right],$$
6. $$\Theta^k = \arg\max_{x \in X} \mathcal{Q}\left(\Theta\middle|\Theta^k\right)$$
7. Define pattern: $X = [x_1, x_2, x_3, \ldots, x_N]$
8. **for** x_N **do**
9. **if** $\mathcal{Q}(s_{k+1}) - \mathcal{Q}(s_k) > \mu(k)^v$ **then**
10. **update** x_N **with DRL**
11. **else**
12. repeat step 6 for all X
13. **end if**
14. **end for**
15. **Update EM deterioration path estimations with DPPG**
16. Re-estimate $\mathcal{Q}(s_{k+1})$ using primary network $\mathcal{Q}(s, a|\theta_\pi)$
17. Initialize target network parameters with EM parameter set
18. Initialize replay memory using EM samples.
19. **for** each EM step **do**
20. Observe user state s_t and SINR ratio $a_t \in \theta_\pi$
21. Execute $a_t \in \theta_\pi$ and state x_t
22. Observe change in $r^{min}(a_t, s_t)$ and $\mathcal{Q}(s, a|\theta_\pi)$
23. Update EM tuple < s_t,a_t, r^{min}, s_{t+1} > in replay memory.
24. Compute target value $\hat{y}_t$, update $\mathcal{Q}(s, a|\theta_\pi)$ and minimizing loss
25. Update target neural networks
26. Update EM with θ_π and recompute steps 6–12 for all X
27. **end for**
28. **end for**
29. **end for**

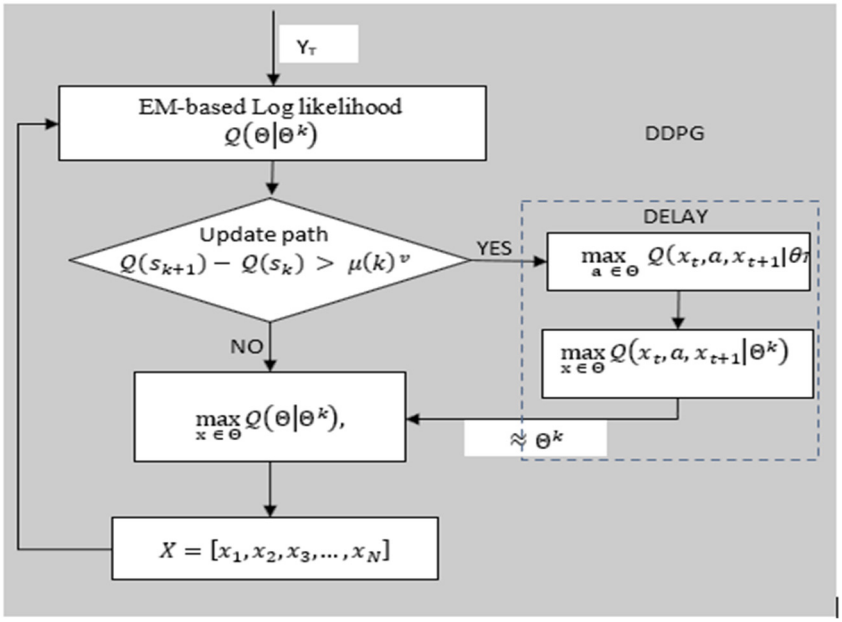

Figure 4. Workflow of JMLS–DDPG algorithm.

3.6. Global Path and Local Path Optimization Formulation

The local pattern is formulated on the basis of local CSI from one mmWave BS. The local agent, thus, considers only the SINR ratio $a \in A_x$ and corresponding received power x values possible in the local environment over given states $\widetilde{S}_t$. The long-term function for the local deterioration pattern is expressed as

$$\mathcal{Q}_{LP}(a_t, x_t|\theta\pi) \triangleq \mathbb{E}\left[\sum_{t=0}^{T} \delta^t \left\{ r^{min}(a_t, x_t) + \varepsilon \mathcal{Q}(x_{t+1}, \pi(x_{t+1}|\theta\pi)) \right\}\right], \tag{8a}$$

where $\delta \in (0,1)$ is the discount factor and approaches 1 with more training samples. The global deterioration pattern is formulated on the basis of collective SINR ratio a_t and received power x_t values from different mmWave BSs over $\widetilde{S}_t$. The value function $\mathcal{Q}_{GP}$ is

$$\mathcal{Q}_{GP}(a_t, x_t|\theta\pi) \triangleq \sum_{a \in A_{x_k}} P_\pi(a_t|x_t) * \frac{\alpha}{K} \left\{ \mathcal{Q}(x_{t+1}a_t, x_t|\theta\pi) r^{min}(x_{t+1}, a_t, x_t) \right\}, \tag{8b}$$

where $P_\pi(a_t|x_t)$ is the probability of receiving x given α in state s by EM; α is the learning rate over K samples in EM.

3.7. Handoff Considerations

We use the Kaiser–Meyer–Olkin (KMO) test [25] to test how much each individual/local mmWave target link's expected deterioration pattern, given the user speed, deviated from its optimized global deterioration pattern. The global deterioration pattern is formulated by collecting training sample from all mmWave BS with respect to user type/speed just like the complete report table (CRT) in [4]. The local deterioration pattern is based on data gathered from an individual BS's local environment with respect to a user's type. It is similar to the report table (RT) user data in [4]. Given all the target BSs with at least 3 dB SINR above the threshold, the KMO indexing test is used to find the level of correlation between an optimized global deterioration pattern and that of a target link at the time of the HO request. The KMO overall index value correlation is defined as follows:

$$\mathrm{KMO}_{\hat{x}} = \frac{\sum_{x \neq \hat{x}} R_{x\hat{x}}^2}{\sum_{x \neq \hat{x}} R_{x\hat{x}}^2 + \sum_{x \neq \hat{x}} a_{x\hat{x}}^2}, \tag{9a}$$

where $R = [r_{xd}]$ is the correlation matrix, and $A = [a_{xd}]$ is the partial covariance matrix, where a_{xd} is defined as

$$a_{x \neq \hat{x}.m} = \frac{r_{x\hat{x}} - r_{x.m} r_{\hat{x}.m}}{(1 - r_{\hat{x}m}^2)(1 - r_{\hat{x}m}^2)}, \tag{9b}$$

and

$$r_{x\hat{x}} = \frac{\sum_{t=0}^{T} (x_t - \hat{x}_t)\left(d_t - \hat{d}_t\right)}{\sqrt{\sum_{t=0}^{T} (x_t - \hat{x}_t)^2 \sum_{t=0}^{T} \left(d_t - \hat{d}_t\right)^2}}, \tag{9c}$$

where $x_t \in X_T$ is the optimum lower bound target link value of received power at state s_t. $d_t \in s_t$ is the minimum expected user–BS link distance, and $\hat{x}_t$ and $\hat{d}_t$ are values for the global deterioration path. The KMO test takes values between 0 and 1, as summarized in Table 1. The general rule for interpreting measurements is provided in Table 1. In this study, we selected the target cells with a KMO index of 0.751. If the KMO index value is less than 0.7, the target link is most likely not suitable for HO consideration although it might have the highest initial SINR. Additionally, during the HO phase, if the serving BS still has a SINR value of 3 dB, the user maintains the connection to the serving gNB. This avoids wasteful HOs. Otherwise, we execute the HO process and then go back to prediction phase.

Table 1. Interpretation Of KMO Measure.

KMO	Interpretation
0.9 and above	Marvelous
0.8–0.9	Meritorious
0.7–0.8	Middling
0.6–0.7	Mediocre
0.5–0.6	Miserable
Under 0.5	Unacceptable

3.8. Measurement Definition

We measured the number of repeated HOs to ascertain if the HO scheme can reduce the number of the wasteful HOs. Repeated HOs mean that the HO scheme is reselecting the same serving BS in which the user is already connected to for another HO. This is wasteful because there is no need to reselect the same BS for HO but rather maintain the link. We also analyzed the sum data rate of mmWave BSs using different HO schemes. Additionally, we analyzed the HO overhead for different schemes. The principle is that a higher overhead reflects a more wasteful HO scheme with the bandwidth. Lastly, we analyzed the performance of our proposed scheme compared to another scheme, dubbed the DDPG only scheme. The DDPG only scheme does not use the meta training technique and does not consider condition (5c). Specifically, it uses random training samples rather than EM refined samples. We also analyzed performance compared to the existing soft HO DC model HO scheme in [3]. This scheme only selects the best target cell by averaging the SINR/data rate.

4. Simulation Results

We used the DC LTE mmWave model introduced by the NYU and the University of Padova in our simulation [1]. The LTE BSs in the DC model manage mmWave BS. The model carefully considers the end-to-end mmWave cellular network performance. It uses an ns-3 simulator and features a 3GPP channel model for frequencies above 6 GHz, as well as a 3GPP-like cellular protocol stack [1]. The JMLS–DRL algorithm was developed using the OpenAI Gym [24] toolkit. Open AI Gym is an RL development that is integrable with the ns-3 simulator; it supports teaching agents for a variety of network applications including those in ns-3. We investigated the performance using system-level simulations. Data collected from over 1000 s of simulation time with a resolution of one transmission time interval (TTI) (1 ms) were used for analysis. The main parameters used are summarized in

Table 2. For a more detailed review of simulators, refer to [15]. Figures 5 and 6 compare the number of wasteful HOs as a function of the number of training episodes in the DRL HO scheme and JMLS–DDPG HO scheme, respectively. The former gets new training samples from the environment once the initial pattern has been defined by EM estimations for every other episode, while the latter uses EM estimated data as the training sample as long as condition (5c) is satisfied. It only requests new training samples when EM data estimates fail to meet condition (5c). Results show that our proposed scheme quickly reduces the number of wasted HOs compared to the DDPG only HO scheme. For instance, it required 250 episodes to reduce repeated HOs to minimal levels of less than five, whilst the DDPG only scheme required close to 400 episodes. This also suggests that it can strategically and ably predict deterioration patterns using fewer training samples. The fact that this is more reliable and accurate than a method that continuously receives new training samples was justified in [4]. The authors in [4] argued that the angles of arrival and received power slowly vary with speeds because they are affected by the large-scale scattering environment and do not change with small-scale mobility. Since the received power samples do not change significantly from one sample to the next, we can use the training samples of the received power in meta training. Figures 7 and 8 compare the cumulative average reward behavior as a function of training episodes under different user types. We can draw several observations. First, the early predictions or rewards of the deterioration pattern for different user types are very fuzzy in the JMLS–DDPG scheme. This explains why there are a high number of wasteful or repeated HOs in the early part of the training of JMLS–DRL, as shown in Figure 6.

Table 2. Simulation Parameter Table [1].

Parameter	Value
mmWave	28 GHz
mmWave bandwidth	1 GHz
3GPP Channel Scenario	Urban Micro, Urban Macro
MMWave max outage	−5 dB
mmWave transmission Power	46 dBm
mmWave max PHY Rate	3.2 Gbps
X2 link latency	1 ms
S1 link latency	10 ms
RLC buffer Size	5 MB
S1 MME link latency	10 ms
User speed	[1,50] m/s

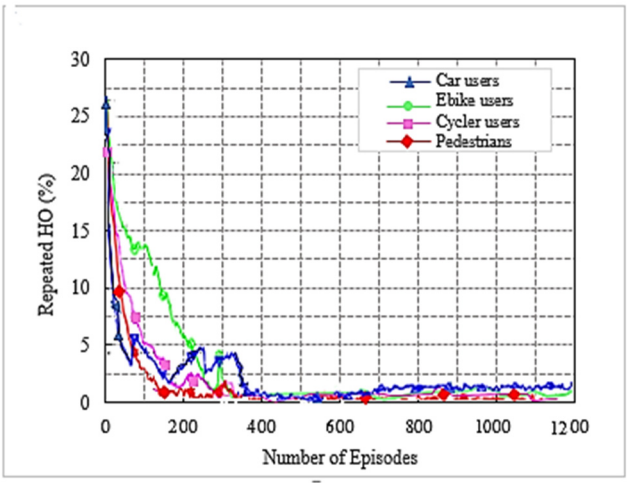

Figure 5. Number of wasteful HO as a function of the number of training episodes for DDPG only HO scheme.

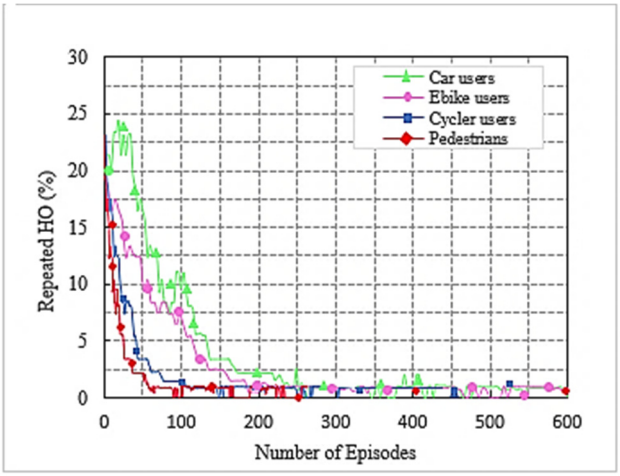

Figure 6. Number of wasteful HO as a function of the number of training episodes for JMLS–DDPG HO scheme.

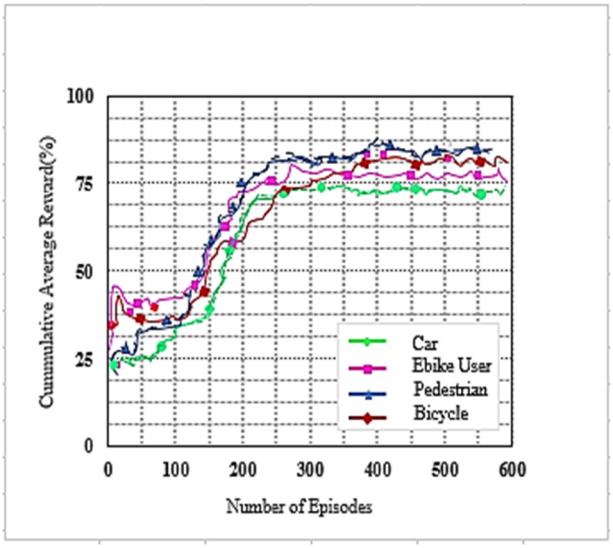

Figure 7. The cumulative reward as a function of the number of training episodes for DDPG only HO scheme according to user type.

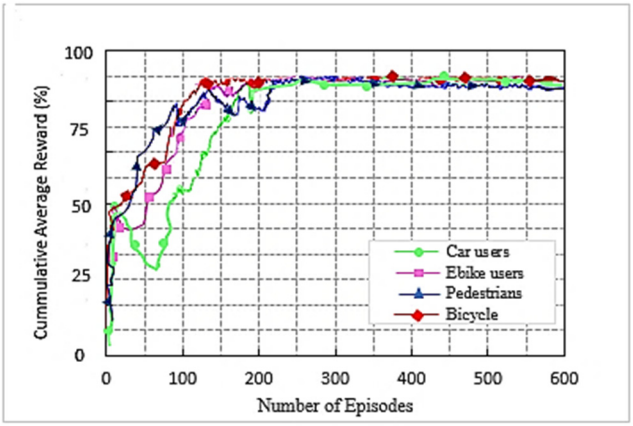

Figure 8. Cumulative reward as a function of the number of training episodes for our proposed JMLS–DDPG HO scheme according to user type.

The blurriness is also seen when we compare the deterioration pattern prediction after 200 episodes in Figure 9 and after 500 episodes in Figure 10. Figure 10 shows a more accurate prediction of likely received power for different user types than Figure 9 with 200 episodes or observations in our proposed JMLS–DRL-empowered HO algorithm. Secondly, while the DDPG scheme converges independently for each user type as seen in Figure 7, the proposed JMLS–DRL scheme converges with almost a common and higher reward for all user types (see Figure 8). The implication is that, after 200 training episodes, the JMLS–DRL algorithm can have one common/global deterioration pattern to follow regardless of user type. On the other hand, for the DDPG HO scheme, each user type will need to follow a different type of deterioration pattern. This facilitates our proposed scheme's prediction of the expected target link behavior. In both schemes, an HO is only issued when the received power at a particular given state/distance from the serving BS drops beyond the corresponding value of the expected local deterioration pattern. In this case, the global and local deterioration patterns in KMO are compared at least within a range of 80 m from a serving mmWave BS. While we can still try and predict beyond 80 m, the computation cost will be too high. Thus, a selected target link is deemed reliable if it is able to sustain connectivity within the 80 m transmission range. Beyond 80 m, HOs are evoked if the SINR drops to at least within 3 dB of the threshold. Therefore, HOs select a link on the basis of the fact that sustained connectivity is expected for at least for 80 m of assumed coverage of the mmWave BS. We also analyzed a soft-HO DC-based scheme [4] using only SINR [2] and a DDPG-based scheme [3] for comparison; the former acted as a baseline for our case in Figures 11 and 12.

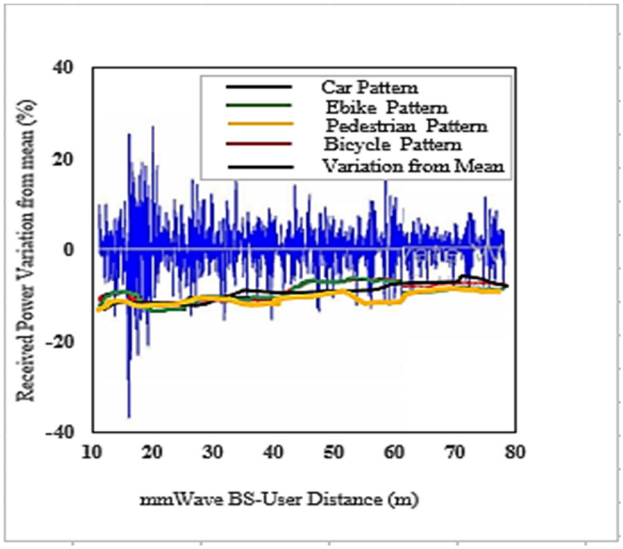

Figure 9. The actual average received power pattern variation at the UE as a percentage about the mean value after 200 episodes in the proposed JMLS–DDPG HO scheme.

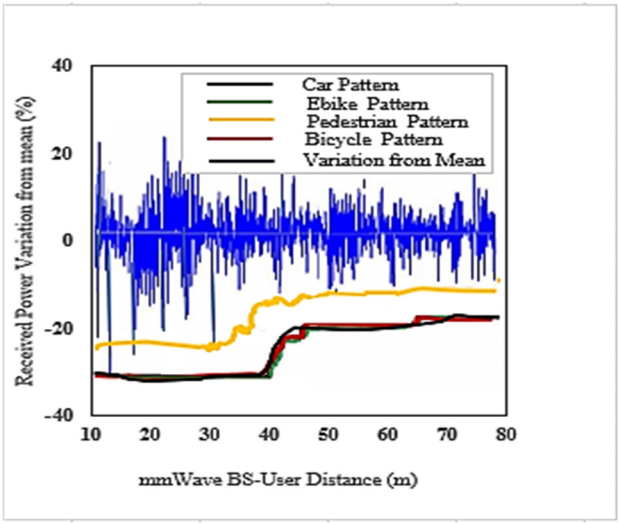

Figure 10. Best expected received power pattern variation as a percentage about the mean value after 500 episodes over 80 m in the proposed JMLS–DDPG HO scheme.

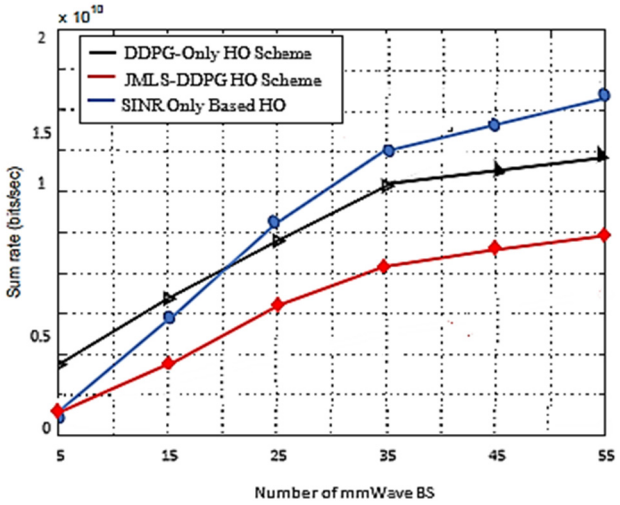

Figure 11. Sum rate for three HO schemes as a function of the number of BSs.

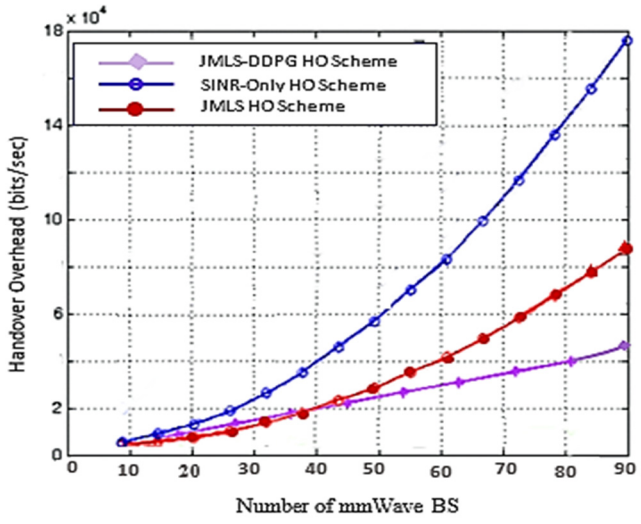

Figure 12. Overhead as a function of the number of mmWave BSs.

In Figure 11, we compare the sum rate as a function of the number of BSs for three different HO schemes. The SINR-based scheme, as explained in [4], only compares the SINR of the target and serving cell/link. The other scheme gets new updates every episode, whilst our proposed scheme uses both new and old CSI. We can see that the proposed scheme has good efficiency in terms of how it uses/selects BSs. The other two schemes seem to start saturating after 35–40 BSs. This can be attributed to the low training sample requirement and thorough analysis of CSI in our scheme. The reuse of training samples gives our scheme ample time to analyze the behavior of links. At the same time, having a small number of mmWave BSs prevents the proposed scheme from learning more about the target link deterioration pattern. This can be seen by the smaller sum date rate recorded at 5 to 15 mmWave BS. More mmWave BSs diversify the amount of data looked at in each episode. On the other hand, despite a very small number of BSs, for the DDPG only HO scheme, the acquisition of new training samples in each episode improved the prediction of the target link path; however, because it changed quickly, the inaccuracy in the predictions quickly manifested.

Another criterion to evaluate the performance of the proposed HO methods is the generated overhead. Figure 12 shows the variation of the induced overhead for the three proposed HO methods. It is obvious that the SINR-based HO induces more handover since, at each attachment to a new BS, a number of new measurement reports must be exchanged to allocate new subcarrier resources. On the other hand, using the DDPG only handover and our proposed HO scheme, fewer overheads are experienced because the past link data needed to achieve reliability are reusable and exchanged in advance before the HO. For our proposed scheme, this advantage is more evident because measurement data sources can be switched depending on condition (5c) (see Figure 3). Hence, the proposed scheme is better than both the DDPG only and the SINR HO schemes.

5. Conclusions and Future Works

This paper proposed a new HO scheme given the distinct propagation characteristics of mmWaves in a HetNet structure. A resource allocation problem that considers the utilization of mmWave bands with LTE bands in a multiuser setup was considered. We considered a downlink LTE-mmWave HetNet scenario with an mmWave link behavior pattern analysis scheme applied to address the HO challenges. The resulting optimization solution consisted of modeling the link behavior using JMLS, DRL, and meta training techniques. Subsequently, the optimal HO link was selected using KMO test principles. Simulation results showed that our HO scheme outperformed the DDPG only HO scheme and the SINR-only based HO scheme in terms of the number of successful HOs. Additionally, the proposed scheme had fewer wasted (repeated) HOs and a quicker reduction in repeated HOs. In particular, as plotted in Figures 5 and 6, if we compare the number of repeated (wasted) HOs when using the existing DDPG (DRL) model and when using with our proposed JMLS–DRL scheme, results show that our scheme's performance was better. For instance, within 200 training episodes, our scheme was able to reduce the total percentage of wasted HOs to less than 5%. This is unlike the DDPG only HO scheme that exhibited over 5% after 200 training episodes. In addition, we compared our proposed scheme with the DRL and SINR HO schemes in terms of the sum rate and overhead performance (Figures 10 and 11, respectively). Our scheme also showed better performance in this regard. For instance, with a network of 55 mmWave BSs, the JMLS–DDPG HO scheme network had a sum rate of nearly 2×10^{10} bits/s and a corresponding overhead of less than 4×10^4 bits/s, as shown in Figures 10 and 11, respectively. This is unlike the DDPG only HO scheme which had a sum rate of less than 1.5×10^{10} bits/s and almost double the overhead (8×10^4 bits/s) for the same number of mmWave BSs. Thus, we can conclusively state that the proposed HO scheme offers longer dew times (time between HOs) than the SINR-based and DDPG only HO schemes. The results demonstrate the vital role that deterioration pattern analysis can play in addressing mmWave link selection in 5G networks. Principally, we can conclude that our pattern analysis HO scheme envis-

ages traits of long-term behavior analysis for mmWave target links before HO execution. This is unlike unreliable classic HO schemes (e.g., the SINR-based HO) where only the instantaneous behavior of target links is analyzed prior to choosing the best target link. In future work, it would be interesting to consider the competing effects of path loss, channel gain, and transmission power when determining the receivable deterioration pattern of the target link. This is given the impact that their variation has on the data rates. Furthermore, while there is a need for highly directional beam antennas at the PHY layer to have an acceptable link quality, how to effectively handle or dodge adverse effects of both mobile and static blockages when choosing mmWave links in HO schemes could be interesting to study in future behavior pattern projections studies for target links. Pattern analysis can also be extended to cell planning, coverage, or rate maximization. This is vital considering the vulnerability of mmWave to topographic and user dynamics. Lastly, studying backhaul configurations that can efficiently support the proposed HO scheme would also be interesting using the pattern-based HO scheme proposed.

Author Contributions: Conceptualization, M.C. and P.H.J.C.; methodology, M.C. and P.H.J.C.; software, M.C. and P.H.J.C.; validation, M.Z., H.S., H.L. and P.H.J.C.; formal analysis, M.C., G.G.M.N.A., H.S., A.K., H.L. and P.H.J.C.; investigation, M.C., M.Z., H.L. and P.H.J.C.; writing—original draft preparation, M.C., P.H.J.C. and M.Z.; writing—review and editing, M.Z., G.G.M.N.A., H.S., A.K. and P.H.J.C.; funding acquisition, P.H.J.C. and M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the School of Engineering, Computer and Mathematical Sciences, Faculty of Design and Creative Technologies, Auckland University of Technology, New Zealand. Funding Number: 16941452.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rebato, M.; Polese, M.; Zorzi, M. Multi-Sector and Multi-Panel Performance in 5G mmWave Cellular Networks. In Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM), Abu Dhabi, United Arab Emirates, 9–13 December 2018; pp. 1–6. [CrossRef]
2. Rangan, S.; Rappaport, T.S.; Erkip, E. Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges. *Proc. IEEE* **2014**, *102*, 366–385. [CrossRef]
3. Dai, Y.; Xu, D.; Maharjan, S.; Zhang, Y. Joint load balancing and offloading in vehicular edge computing and networks. *IEEE Internet Things J.* **2018**, *6*, 4377–4387. [CrossRef]
4. Mwanje, S.; Zia, N.; Mitschele-Thiel, A. Self-Organized Handover Parameter Configuration for LTE. In Proceedings of the 9th International Symposium on Wireless Communication Systems (ISWCS'12), Paris, France, 28–31 August 2012; pp. 26–30.
5. Shubyn, B.; Maksymyuk, T. Intelligent Handover Management in 5G Mobile Networks based on Recurrent Neural Networks. In Proceedings of the 2019 3rd International Conference on Advanced Information and Communications Technologies (AICT), Lviv, Ukraine, 2–6 July 2019; pp. 348–351.
6. Joud, M.; García-Lozano, M.; Ruiz, S. User specific cell clustering to improve mobility robustness in 5G ultra-dense cellular networks. In Proceedings of the 2018 14th Annual Conference on Wireless On-Demand Network Systems and Services (WONS), Isola, France, 6–8 February 2018; pp. 45–50.
7. Shanmugam, K.; Golrezaei, N.; Dimakis, A.G.; Molisch, A.F.; Caire, G. Femtocaching: Wireless content delivery through distributed caching helpers. *IEEE Trans. Inf. Theory* **2013**, *59*, 8402–8413. [CrossRef]
8. Blackmore, L.; Ono, M.; Bektassov, A.; Williams, B.C. A probabilistic particle-control approximation of chance-constrained stochastic predictive control. *IEEE Trans. Robot.* **2010**, *26*, 502–517. [CrossRef]
9. Chitraganti, S.; Aberkane, S.; Aubrun, C.; Valencia-Palomo, G.; Dragan, V. On control of discrete-time state-dependent jump linear systems with probabilistic constraints: A receding horizon approach. *Syst. Control Lett.* **2014**, *74*, 81–89. [CrossRef]
10. Du, J.; Valenzuela, R.A. How Much Spectrum is too Much in Millimeter Wave Wireless Access. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 1444–1458. [CrossRef]
11. Zhou, Z.; Yu, H.; Xu, C.; Zhang, Y.; Mumtaz, S.; Rodriguez, J. Dependable content distribution in D2D-based cooperative vehicular networks: A big data-integrated coalition game approach. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 953–964. [CrossRef]

12. Zhou, Z.; Gao, C.; Xu, C.; Zhang, Y.; Mumtaz, S.; Rodriguez, J. Social big-data-based content dissemination in Internet of Vehicles. *IEEE Trans. Ind. Inform.* **2018**, *14*, 768–777. [CrossRef]
13. Rodrigues, T.G.; Suto, K.; Nishiyama, H.; Kato, N.; Temma, K. Cloudlets activation scheme for scalable mobile edge computing with transmission power control and virtual machine migration. *IEEE Trans. Comput.* **2018**, *67*, 1287–1300. [CrossRef]
14. Maksymyuk, T.; Han, L.; Larionov, S.; Shubyn, B.; Luntovskyy, A.; Klymash, M. Intelligent Spectrum Management in 5G Mobile Networks based on Recurrent Neural Networks. In Proceedings of the 15th IEEE International Conference The Experience of Designing and Application of CADSystems (IEEE CADSM'2019), Polyana, Ukraine, 26 February–2 March 2019.
15. Rodrigues, T.G.; Suto, K.; Nishiyama, H.; Kato, N. Hybrid method for minimizing service delay in edge cloud computing through vm migration and transmission power control. *IEEE Trans. Comput.* **2017**, *66*, 810–819. [CrossRef]
16. Edwards, C. *Advanced Calculus of Several Variables*; Dover Publications: London, UK, 1973.
17. Kocvara, M.; Stingl, M. Pennonacode for convex nonlinear and semi definite programming. *Optim. Methods Softw.* **2003**, *8*, 317–333. [CrossRef]
18. Dempster, A.P.; Laird, N.M.; Rubin, D.B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B (Methodol.)* **1977**, *39*, 1–38.
19. Bilmes, J.A. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. *Int. Comput. Sci. Inst.* **1998**, *4*, 126.
20. Costa, O.L.V.; Fragoso, M.D.; Marques, R.P. *Discrete-Time Markov Jump Linear Systems*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 2006.
21. Zhou, Z.; Feng, J.; Chang, Z.; Shen, X. Energy-efficient edge computing service provisioning for vehicular networks: A consensus admm approach. *IEEE Trans. Veh. Technol.* **2019**, *68*, 5087–5099. [CrossRef]
22. Sorkhoh, I.; Ebrahimi, D.; Atallah, R.; Assi, C. Workload scheduling in vehicular networks with edge cloud capabilities. *IEEE Trans. Veh. Technol.* **2019**, *68*, 8472–8486. [CrossRef]
23. Eason, G.; Noble, B.; Sneddon, I.N. On certain integrals of Lipschitz-Hankel type involving products of Bessel functions. *Phil. Trans. R. Soc. Lond.* **1955**, *A247*, 529–551.
24. Gawłowicz, P.; Zubow, A. Ns-3 meets OpenAI Gym: The Playground for Machine Learning in Networking Research. In Proceedings of the ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Miami Beach, FL, USA, 25–29 November 2019. [CrossRef]
25. Glen, S. Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy, From StatisticsHowTo.com: Elementary Statistics for the Rest of Us! Available online: https://www.statisticshowto.com/kaiser-meyer-olkin/ (accessed on 20 November 2021).
26. Shlezinger, N.; Farsad, N.; Eldar, Y.C.; Goldsmith, A.J. ViterbiNet: A Deep Learning Based Viterbi Algorithm for Symbol Detection. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 3319–3331. [CrossRef]
27. Al-Nima, R.R.O.; Han, T.; Chen, T. Road tracking using deep reinforcement learning for self-driving car applications. *Futur. Gener. Comput. Syst.* **2018**, *108*, 1092–1111. [CrossRef]
28. Yang, P.; Chen, L.; Zhang, H.; Yang, J.; Wang, R.; Li, Z. Joint Optical and Wireless Resource Allocation for Cooperative Transmission in C-RAN. *Sensors* **2021**, *21*, 217. [CrossRef]
29. Koda, Y.; Nakashima, K.; Yamamoto, K.; Nishio, T.; Morikura, M. Handover Management for mmWave Networks with Proactive Performance Prediction Using Camera Images and Deep Reinforcement Learning. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 802–816. [CrossRef]
30. Ho, T.M.; Nguyen, K.-K. Deep Q-Learning for Joint Server Selection, Offloading, and Handover in Multi-access Edge Computing. In Proceedings of the ICC 2021—IEEE International Conference on Communications, Montreal, QC, Canada, 14–23 June 2021. [CrossRef]
31. Guo, D.; Tang, L.; Zhang, X.; Liang, Y.-C. Joint Optimization of Handover Control and Power Allocation Based on Multi-Agent Deep Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2020**, *69*, 13124–13138. [CrossRef]