Article

Enhancing Handover for 5G 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[2], Hui Li[3]

[1] Department of Electrical and Electronic Engineering, Auckland University of Technology, Auckland, New Zealand ; masoto.chiputa@aut.ac.nz, mizhang@aut.ac.nz, peter.chong@aut.ac.nz, hakilo.sabit@aut.ac.nz
[2] Department of Computer Science and Information Systems, Bradley University, Peoria, IL 61625, USA; nali@fsmail.bradley.edu
[3] Shenzhen Graduate School, Peking University, Shenzhen, China; lih64@pku.edu.cn
* Correspondence: peter.chong@aut.ac.nz

Abstract: The fifth Generation (5G) mobile networks use millimeter Waves (mmWaves) to offer giga bit 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 cause too early, too late, or wrong handoffs (HOs). To mitigate HO challenges, sustain connectivity and avert unnecessary HO, we propose a HO scheme based on 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 behaviour by predicting likely deterioration patterns of target links. The prediction is optimized by meta training techniques that also reduces training sample size. Thus, the JMLS-DRL platform formulates intelligent and versatile HO policies for 5G. Results show our proposed prediction scheme about target link behavior post HO to be highly reliable. The scheme also averts unnecessary HOs thus ably supports longer dew time.

Keywords: millimeter bands; fifth Generation; Handover; Deep Reinforcement Learning; and Jump Markov Linear System.

1. Introduction

The fifth generation (5G) mobile users will need uninterrupted connectivity while consuming large amounts of data and media content when commuting [1]. The millimeter wave (mmWave) bands (i.e., between 30−300 GHz of the radio spectrum) hold great potential to enabling 5G mobile users experience in Gigabit rates and networks meet traffic demands. However, a caveat to this is that mmWave communication is very susceptible to topographic and user dynamics. Common materials like concrete, water, and even human bodies/movements among others [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 solutions have been awash 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 [5, 7] select target cells based on initial maximum network performance values. However, the challenge is the optimum initial value do not always guarantee reliability of the connection after HO. For instance, the selection of mmWave target links based on highest SINR values [4-7], does not always reveal the reliability of the link after a HO event. In most cases, HOs end up getting executed too early, too late, wrongly, or wastefully. To that effect, 5G mobile network performance is punctuated with gradual and abrupt changes. To reduce inconsistences in network performance, selection of best target links requires understanding not just the immediate behavior after HO but also the long-term behavior i.e., post HO.

To that effect, we propose a 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) 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 are known to account for abrupt changes[7] 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 set to make new decisions. This reduces request for new training data sets when making new decisions in novel location. At HO, the predicted deterioration pattern of a target link is then compared against an optimal global desired deterioration patterns to understand reliability of a target link and select the most stable one.

1.1. Contributions

- We propose to use JMLS to model deterioration behavior/patten of mmWave target links and 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, thus have ample time to track pattern changes of rapid-varying channel in real-time, we propose using Meta-learning techniques. Meta-learning is a technique that automatically adapts reuse of training data from related past, tasks or neighbours to make new decision. This reduces the need for new CSI/training data set to make new decisions
- We use Kaiser-Meyer-Olkin (KMO) test to measure the expected divergence of target links from optimum deterioration pattern post HO to know their reliability.

1.2. Related Works

The surging role/potential of mmWave bands in mobile networks such as 5G/beyond cannot be ignored. However, so are its challenges, particularly in the mobility management support of 5G networks. The authors in [23] for instance claim higher propagation losses inherent in mmWaves must be addressed to sustain connectivity especially at ranges beyond 100 meters and in non-line-of-sight (NLOS) settings. The authors in [24] takes four directions 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 include, a physical layer distance-aware design, ultra-massive MIMO communication, reflect arrays, and intelligent surfaces. These use Machine and Artificial intelligence (AI) learning for 5G. The author in [24] suggest a move from centralized (used in most 4G systems) to decentralized mobility management algorithms using DRL. DRL in 5G ably learn and build knowledge about different dynamics of mmWave channels . For instance, by interacting with environment data, authors in [11] 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.

Exploiting actor-critic DRL, authors in [8] proposed to jointly solve offload and resource allocation problems in fog networks. Authors in [12] used deep Q-learning based task offloading scheme to select optimal BSs for users and maximize task offloading utility. In [13], Q-learning integrates Mobility Robustness Optimization (MRO) scheme with

doi:10.20944/preprints202111.0514.v1

Mobility-Load-Balancing (MLB) scheme to tackle traffic Load and speed effects in 5G. However, in all these schemes high mobile and dynamic users are hardly considered. Additionally, DRL requires thousands of samples to gradually learn useful policies [15]. Besides, DRL acts 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, [20], uses JMLS with Expected Maximization (EM) to predict abrupt deterioration behaviour. It then enhances predictions using Viterbi algorithms. The Viterbi algorithm however requires accurate Channel State Information (CSI) to converge. In such cases, paper [26] argues that inaccurate training gradually cripples the accuracy of predictions, particularly in low signal-to-noise ratios (SNRs). To that effect, the author combines it with the meta data training, making the Viterbi proposed approach more reliable and less dependable on changing and accuracy of the data. In [18], to tackle distributed making decision scenario, the author extends the JMLS formulation into game theoretic technique. Similarly [17], incorporates particle-filter-based RL in JMLS to predict a finite number of disturbances within a randomly chosen sample of trajectories. This allows the scheme to track/adjust to time-varying conditions in real-time

1.3. Organization

The remainder of this paper is organized as follows. Section II describes the proposed framework and its operation. Section III describes resource allocation and optimization problems. Sections IV and V present adoptions of JMLS-DRL solution. Sections VI and VII analyse simulation results and the conclusion, respectively.

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 received power deterioration pattern for four different types of users with respect to mmWave BSs. The four types of users include cars, pedestrians, cyclers and ebikers. For each user type, prior to HO selection, the scheme learns the likely mmWave user received power deterioration pattern given 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 modelled using JMLS. It models how likely received power will deteriorate for a user given NLOS and distance effects on mmWave channel. Thus, in the first instance, the model learns and determines desired optimal received power deterioration patterns for different user types using Expected Maximization (EM) [6]. 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 by 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 large data variance analyzed, the global pattern is regarded to be more accurate.

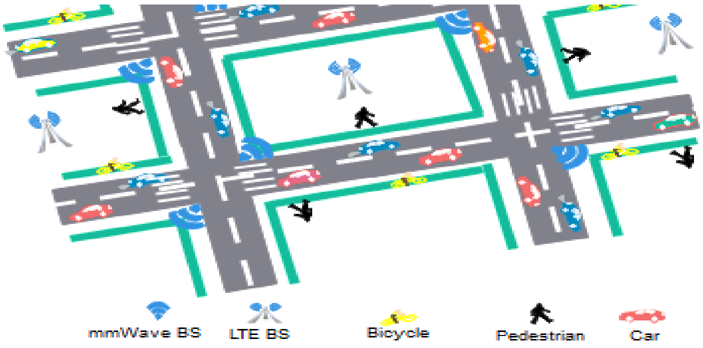

Figure 1. Multiuser type Mobility Model.

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 the past HOs. 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 to 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 Fig.1) in an urban scenario. The road network area is 500m x 100m. We have four types of users: pedestrian, cycler, and cars. A quarter are pedestrian with speeds of 1.4m/s. Another quarter are cyclers with speeds of 3 - 7 m/s, the other quarter are ebikers between 8 - 9 m/s. The other quarter are car users with velocities of between 10-14m/s. Cars in the space of 3m or more to each other and adjust velocities by 1-3m/s to avert crashes. Car speeds are updated every 3s to decrease/increase. Each street consists of the right and left lanes for each user-type.

Given user directions i.e., $\mathfrak{y}$ ={moving towards/away from a mmWave BS}, users traverse different streets. The probability of recovering channel link just after being blocked $\mathbb{P}_{\mathfrak{y}r}$ and remaining blocked $\mathbb{P}_{\mathfrak{y}b}$ is [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 non-blocking and blocking windows within a transmission range d. The rate of channel links switching from blocked to recovered/vice versa within d is $1/T^r$ and $1/T^b$ respectively. Otherwise $\mathfrak{y}$ is binary and assumed 1 when the users are moving towards a target BS. $\mathfrak{y}$, is assumed to be 0 as recovery of reconnection over the serving cell is minimal if user is moving away. The argument is link recovery chances are high if a user is moving toward the direction of mmWave BS.

2.1. Outage Probability

Assuming, 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 [2] and [11]:

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

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 all users have a receiver sensitivity of x_{kmin}.Thus, each serving mmWave BS (with either LOS or NLOS link) given, P, must satisfy 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 (2). α is the path loss exponent in LOS and NLOS path-loss 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 target BS are unable to communicate owing to insufficient received power x_t. The data rate is defined as:

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

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

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. $|h^H\boldsymbol{p}|^2$ is channel gain. $F_x(|\theta_k^l|)$ denotes the Fejér kernel value. As user speed approaches zero and $F_x(|\theta_k^l|) \rightarrow 1$, SINR approaches to the maximum. F_x approaches 0 as v increases [4].

[redacted]*Resource Allocation Problem*

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

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

where $P_\pi^{m|x_t}$ and $P_\pi^{m|u_t}$ are 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 (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 current received power in LOS. $\hat{\gamma}$ and γ^{min} are targeted and measured SINR needed to satisfy $\mathbb{R}^m$. It must be noted that if there exists an infeasible SINR target in certain user state, the resulting power demand, x_{t+1}, by users may diverge to infinity. This is due to each user link attempt 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 consumptions for a given x_{t+1} is (3c) [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 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, $e_0 * \zeta c(t-w)$

denotes energy lost due to lost packets (expected number less the actual number of received packets) at t during window w. Given receivable,x_t, and transmittable power, P, constraints (see section 2A), for optimum packet delivery latency, the maximum link utility problem is formulated as[19]:

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

where δ, expected latency scaling factor given, x_{t+1} within w. ζE_c is latency discrepancy following x_t to x_{t+1} change as the user moves away from the serving BS. We learn to predict the long-term deterioration pattern $\{x_t, \dots, x_T\}$ of the target links to ascertain its reliability to meeting 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 resource allocation problem in (3a to 3d) 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 describe (3a)-(3d) as in (4a):

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

where $x_t \in X$ is the current received power in LOS given state, s_t, $u_t \in \mathcal{U}$ 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 (3b); $A(s_t)$, and $B(s_t)$ are SINR/power coefficient matrices in (3b). $v_t \sim \mathcal{N}(0, Q(s_t))$ and $w_t \sim \mathcal{N}(0, R(s_t))$ are data rate and received power measurement noise, respectively. Measurement noises are influenced by competing effect of change in gain, angular and linear transmission distance, user speed etc. for the same SINR requirements (see eqn. (3a)-(3c)). s_t denotes a state governing for parameter set $\Theta = \{A, B, R, r^{min}, Q, P(S)\}$. s_t belongs to a set of Markov stochastic decisions $\mathcal{M} = \{m_1, m_2, \dots, m_M\}$ and m_M determine which state is active at time t.

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

where:

$v = [v_1, \dots, v_T]$:	is a vector of User velocity
$r_t = [r_1, \dots, r_t]$:	is a vector of possible user data rate
$T_t = [t_1^m, \dots, t_N^m]$	is a vector of average service time
$d = [d_t^m, \dots, d_T^m]$	is a vector of transmission distances with same SINR.
$\eta = [\eta_1, \dots, \eta_N]$	is a vector of user direction in 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:

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

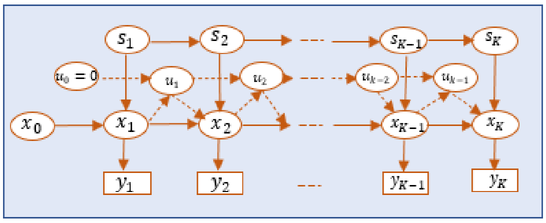

Figure 2. Dynamic Bayesian representation of JMLS composed by 3 variables and relates deterioration variables to each other over adjacent time steps K.

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

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

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

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

Assuming 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 in (4f) :

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

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

3.2. Initial Deterioration Path Training

Given Y_T,$X_T$and S_T denotes a sequence of observed data rates $\{y_1, \dots, y_T\}$ over corresponding receivable power values $\{x_0, \dots x_T\}$, and $\{s_1, \dots, s_T\}$ states respectively till 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,.., $k_2 \in T$ at distance $k_1 \le k_2$. The initial deterioration pattern estimator upon which we design our framework for received power pattern X, is an EM algorithm in [12]. EM uses Bayesian inference to automatically infer optimal value set of Θ for X_T[12] at each step k as seen in Fig. 2 and the value function can be written as:

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

Such that

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

where Θ(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 (5d) to avoid abrupt changes or shocks in data rate:

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

where $\mu(k)^v$ is the averaged data-rate-discrepancy between state s_k and s_{k+1} for a user with velocity v. In (5d), the smaller the difference, the lower the 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. Besides EM, cannot handle such switching dynamics [12]. Thus, instead of recomputing steps in (5a)-(5c) to refine pattern X, as more CSI about Y_T is obtained, we use online DRL with EM estimations of X, as initial experience to determine user target data rates Y_T as will be seen in Fig. 3.

3.3. Deep Reinforcement Learning in EM-Estimates

As seen in (5b), EM estimates gives the maximum obtainable received power, x, i.e., the upper bound of desirable received power in each state needed to obtain high SINR, $a \in A_x$, 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 high transmit power (which is energy inefficient and is beyond the limit) to meet desirable receivable power and receive the same amount of user data within a given QoS/SINR requirement. However, if gain of the channel is high in the peak, even if received power is lower (e.g., in NLOS), this will permit using lower receivable power to receive the same/similar amount of data while maintaining the same given QoS/SINR. Thus, knowing not just the pattern of the maximum but minimum receivable power prior to HO decision is vital. Hence the need to use DRL is to determine minimum desirable power given the maximum by EM estimation. Here DRL uses EM data as 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 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 Fig. 2

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

Technically, for a given optimum pattern, $-X^*$, in (5d), the algorithm uses corresponding optimized parameter sets θ_π 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 initial experience and improves it by minimizing expected energy cost, $\mathcal{J}(x_t)$. The policy, $\pi(s_t|\theta\pi)$, is defined as:

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

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 (4f), satisfying the following Bellman equations.

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

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

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

3.4. Deep Deterministic Policy Gradient (DDPG)

We use Deep Deterministic Policy Gradient (DDPG) to do improve the accuracy of the pattern. DDPG combines with DQN on the premise of 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's the experience replays memory and the target network to solve the problem of non-convergence to approximate the EM function values in neural networks. It thus is 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 by 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, 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 in (5d). Experience tuples include the current and next state, the SINR ratio value following the transition between state, and reward for choosing the received power level in X_T. Replay memory updates are randomly sampled for training primary critic network and setting target in the target network for eqn. (5d) eventualities.

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 corresponding maximum change $\Delta r^{min}(-x_t, a_t, x_t, s_t)$ for each value ,x_t, transitioning from $-x_t$ and is defined as:

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

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 ∀s ∈ S, $|\mathcal{J}_k$(x) $-\mathcal{J}_k$(-k)| $\leq \varepsilon$ and termination is guaranteed for $\varepsilon > 0$. ε is similar to Greedy strategy with probability $1 - \varepsilon$[27]. Here ε decays as more iteration hence experience is gained. The primary critic network updates θa by minimizing loss function $Ls(\theta\pi)$ defined as:

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

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

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

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

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

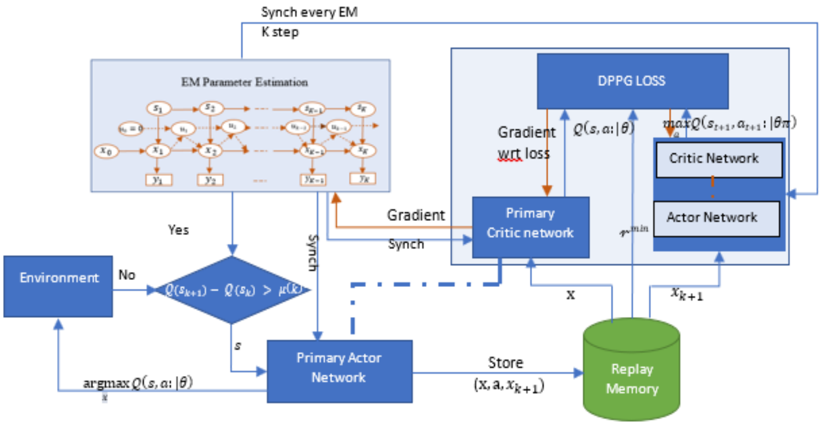

Figure 3: Deep Deterministic Policy Gradient (DDPG) algorithm structure

According to (7d), the parameter θ_Q of primary critic neural network can be updated. Specifically, at each training step, with a mini-batch experience $< s_t, a_t, R^{imm}, s_{t+1} >, t \in \{1, \dots, k\}$, 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 θ_π^T and θ_Q^T. At each iteration, θ_π^T and θ_Q^T in (7c) and (7d) are updated with a weighted combination of the previous state. The prediction of target path takes the form of a weighted combination of models :

$$\theta_\pi^T(\tilde{x}_k) = w\theta_\pi(-\tilde{x}_k) + (1-w)\theta_\pi^T(-\tilde{x}_k)$$
$$\theta_Q^T(\tilde{x}_k) = w\theta_Q(\tilde{x}_k) + (1-w)\theta_Q^T(\tilde{x}_k), \tag{7e}$$

Where $\omega \in [0,1]$ is weight computed using Gaussian kernel parameterized by the transmission distance metric $d_k \in \tilde{s}_k$:

$$w_k = \exp\left(-0.5(x-\mu_k)^T d_k (x-\mu_k)\right) \quad , \tag{7f}$$

Target neural networks generate target or ideal values for training and re-optimizing deterioration pattern, X_T from $-X_T$ based on EM and replay updates. Thus EM estimations in each iteration are used as meta data to 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 target Actor and Critic network are updated in a way of slow change (Soft Replace) by EM estimated values. Here, instead of directly and randomly training parameters of primary Actor and Critic network to further enhance the stability of the training process, we copy EM estimations as ideal initial values. Replay memory stores EM experience tuples formulating X_T, and each value update $x_t \in X_T$ include tuple $< -x_t, a_t, R^{imm}, x_t >$ update.

Fig. 3 shows the structure of the proposed JMLS-DDPG algorithm. The DDPG algorithm takes EM parameter data set and maximum receivable power values ,-X, as initial input to determine minimum receivable power values of a pattern. Given power effects on SINR can be reduced in high channel gain locations, afterwards, the DDPG agents outputs an minimum receivable power values X, needed to maintain the same SINR ratio previous 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 and, on the contrary, it gives a negative reward if condition (5d) 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, neural network trains experience and continuously adjusts SINR strategy by randomly extracting sample data from EM pool and uses the gradient descent approach to update and iterate network parameters, so as to further enhance the stability of pattern

Proposed Algorithm: 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 $Y = [y_1, y_2, y_3, \dots, y_N] \in \mathbb{R}$,
Output: mmWave Deterioration path $X = [x_1, x_2, x_3, \dots, x_N]$ for target link
1.  Initialize the deterioration path estimations
2.  for $t = 1$ do
3.    Draw $y_t$ for JMLS parameter estimation $\Theta$, where $(X_T, S_T, Y_T|\Theta)$
4.  Estimate Maximization (EM):
5.        $\mathcal{Q}(\Theta|\Theta^k) = \mathbb{E}[log\, P(X_T, S_T, Y_T|\Theta)|\, Y_T, \Theta^k]$,
6.    $\Theta^k = \arg\max_{x \in X} \mathcal{Q}(\Theta|\Theta^k)$
7.  Define Pattern : $X = [x_1, x_2, x_3, \dots, x_N]$
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 $\mathscr{r}^{min}(a_t, s_t)$ and $\mathcal{Q}(s, a|\theta_\pi)$
23.        Update EM tuple < $s_t$,$a_t$, $\mathscr{r}^{min}$, $s_{t+1}$ > 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 $\theta_\pi$ and recompute step 6-12 for all $X$
27.            end for
28.        end for
29. end for
```

X and accuracy of the algorithm. Using EM experiences as initial training data input to DDPG restricts search range for optimal minimum receivable power values. Thus, any observed mmWave BS data rate not meeting corresponding receivable power is immediately discarded for training or consideration. This in itself technically reduces training sample for DRL hence convergence time . Ultimately, the improved DRL HO is obtained by combining DDPG with EM predictions acting as meta training sample. Finally, the pattern model is integrated into HO platform for HOs.

3.5. Online Update of Target Deterioration Path

DDPG subdivides the training network structure into online network and target network (See Fig.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 online (primary) Actor network and online Critic network, respectively. The target network includes target Actor network and target Critic network, which get updated by EM values. The target Actor network system does not however 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 large than 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, will be regarded to provide reliable training sample for the target network in DDPG. EM data is thus re-encoded to generate new training samples for the DRL and set new targets over $\tilde{S}_t$ henceforth as meta-training. If indeed the pattern of link deterioration is successfully followed by the target mmWave network, then $\tilde{X}_t$ represents the true channel link deterioration behaviour from which Y_t is obtained. Consequently, the corresponding pair $\tilde{S}_t$and Y_t parameter set $\theta_\pi(s_t)$ can continue being used to re-train DDPG instead of requesting new CSI from the environment in Fig. 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 $-\tilde{X}$ of $\tilde{X}$ over $\tilde{S}_t$ do not accurately represent the channel behaviour 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 SINRs areas where link deterioration pattern errors occur frequently. Nonetheless, when pattern errors are less in EM, the effects of decision estimate errors of ε, namely, the number of errors in a pattern, can be used to decide when to generate meta-training. For instance, we re-train 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 is used, and the effect of decision errors is controlled. When using new training samples, we cleverly focus attention on states with un converged pattern values i.e., where (5d) is not fulfilled. Our online training mechanism is summarized in the proposed Algorithm 1.

3.6. Global Path and Local Path Optimization Formulation

The local pattern is formulated based on 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 $\tilde{S}_t$. The long-term function for local deterioration pattern is:

$$Q_{LP}(a_t, x_t|\theta\pi) \triangleq \mathbb{E}\left[\sum_{t=0}^{T} \delta^t\{r^{min}(a_t, x_t) + \varepsilon Q(x_{t+1},\pi(x_{t+1}|\theta\pi))\}\right], \quad (9a)$$

where δ ∈ (0,1) is the discount factor and approaches 1 with more training sample. The global deterioration pattern is formulated based on a collective SINR ratio a_t and received power x_t , values from different mmWave BSs over $\tilde{S}_t$.The value function Q_{GP} is:

$$Q_{GP}(a_t, x_t|\theta\pi) \triangleq \sum_{a\in A_{x_k}} P_\pi(a_t|x_t) * \frac{\alpha}{K}\{Q(x_{t+1}a_t, x_t|\theta\pi) r^{min}(x_{t+1}, a_t, x_t)\}, \quad (9b)$$

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

3.7. Hand-Off 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 a report table (RT) user data in [4]. Given all the Target BSs with at least 3dB SINR above threshold, the KMO indexing test is used to find the level of correlation between an optimized global deterioration pattern and that of target link at the time of the HO request . KMO overall index value correlation defined as (14):

$$KMO_{\hat{x}} = \frac{\sum_{x\neq\hat{x}} R_{x\hat{x}}^2}{\sum_{x\neq\hat{x}} R_{x\hat{x}}^2 + \sum_{x\neq\hat{x}} a_{x\hat{x}}^2}, \quad (10a)$$

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

$$a_{x\neq\hat{x}.m} = \frac{r_{x\hat{x}} - r_{x.m}r_{\hat{x}.m}}{(1-r_{xm}^2)(1-r_{\hat{x}m}^2)}, \tag{10b}$$

and

$$r_{x\hat{x}} = \frac{\sum_{t=0}^{T}(x_t - \hat{x}_t)(d_t - \hat{d}_t)}{\sqrt{\sum_{t=0}^{T}(x_t - \hat{x}_t)^2 \sum_{t=0}^{T}(d_t - \hat{d}_t)^2}}, \tag{10c}$$

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 $\hat{x}_t$ and $\hat{d}_t$ are values for the global deterioration path. KMO test takes values between 0 and 1 and Table I summarizes index values. The general rule for interpreting measurements are in Table I. In this study, we select the target cells with KMO index of 0.751. If the KMO index value is less than 0.7, most likely the target link is not suitable for HO consideration though it might have the highest initial SINR. Additionally, during HO phase, If the serving BS still has a SINR value of 3dB, 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	Interpretaion
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 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 as well analyzed the HO overhead for different schemes. The principle is the higher the overhead the more wasteful the HO scheme is with the bandwidth. Lastly, we analyzed the performance of our proposed scheme against another scheme dubbed DDPG only scheme. DDPG-only scheme does not use meta training technique and does not consider condition (5d). It particularly uses random training samples than EM refined samples. We also analyzed performance against an existing soft-HO DC model HO scheme in [3]. The scheme only selects the best target cell by averaging their SINR/data rate.

4. Simulation Results

We propose to use 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 an end-to-end mmWave cellular network performance. It uses ns-3 simulator and features 3GPP channel model for frequencies above 6 GHz and a 3GPP-like cellular protocol stack [1]. The JMLS-DRL algorithm is developed using OpenAI Gym [24] toolkit . Open AI Gym is a RL development too and is integrable with ns-3 simulator: 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 1000s simulation time with a resolution of one Transmission Time Interval (TTI) (1 ms)

Table 2. The Simulation Parameter Table

Parameter	**Value**
mmWave	28GHz
mmWave bandwidth	1GHz
3GPP Channel Scenario	Urban Micro, Urban Macro
MMWave max outage	-5dB
mmWave transmission Power	46dBm
mmWave max PHY Rate	3.2Gbps
X2 link latency	1ms
S1 link latency	10ms
RLC buffer Size	5MB
S1 MME link latency	10ms
User speed	[1,50] m/s
UDP Source rate	200Mbits/sec

was used for analysis. The main parameters used are summarized in Table II. For a more detailed review of simulators, refer to [15]. Figs. 3 and 4 compare the number of wasteful HOs against the number of training episodes in JMLS Viterbi 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 later uses EM estimated data as training sample so that long condition (5d) is satisfied. It only requests new training samples when EM data estimates fail to meet (5d) conditions. Results show that our proposed scheme quickly reduces the number of wasted

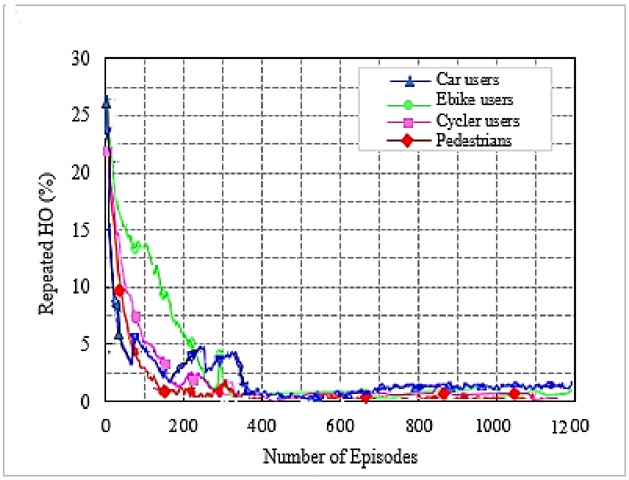

Figure 3. Number of wasteful HO vs Number of training Episodes for DDPG only HO scheme.

HOs than the DDPG only HO scheme. For instance, it requires 250 episodes to reduce repeated HOs to minimal levels of less than five whilst the DDPG only scheme requires close to 400 episodes. This also entails that it can strategically and ably predict deterioration patterns using less training samples. The fact that is more reliable accurate than one that keeps on getting new training samples is justified in [4]. The authors in [4] argue that the angles of arrival and received keeps on getting new training samples is justified in [4]. The authors in [4] argue that the angles of arrival and received power slowly varies with speeds because they are affected by the large-scale scattering environment and do not

doi:10.20944/preprints202111.0514.v1

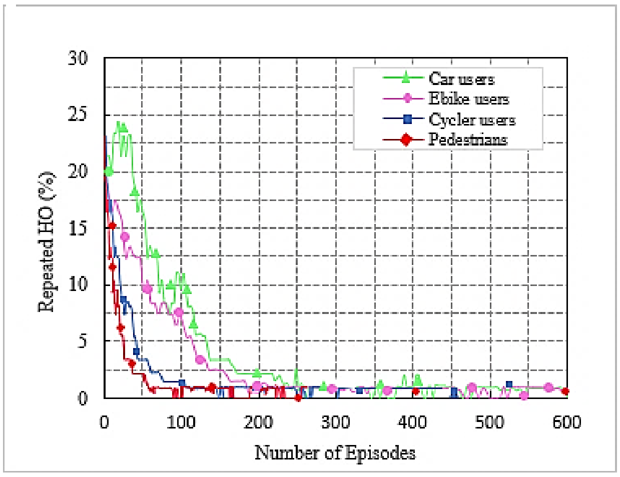

Figure 4. Number of wasteful HO vs Number of training Episodes for JMLS-DDPG only HO scheme.

change with small-scale mobility. Since the received power samples do not change significant from one sample to next, we can use the training samples of the received power in meta training. Figs. 5 and 6 compare cumulative average reward behavior against 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 JMLS-DDPG scheme. This explains why there is a high number of wasteful or repeated HOs in the early part of the training of JMLS-DRL as shown in Fig.4.

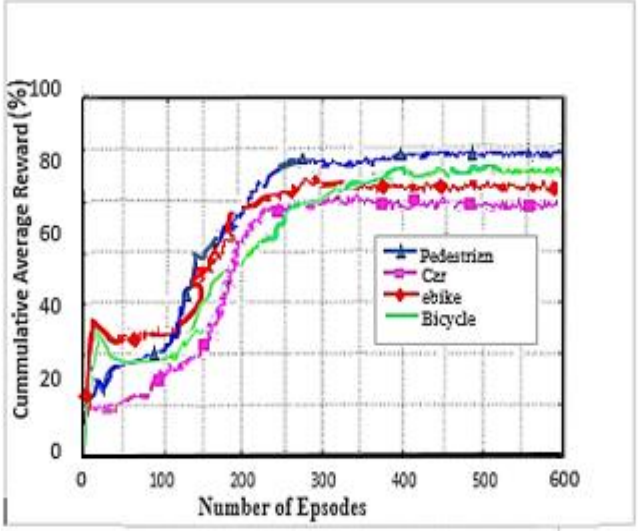

Figure 5. The cumulative reward vs Number of training Episodes for DDPG only HO scheme according to user type.

The blurriness is also seen when we compare deterioration pattern prediction after 200 episodes in Fig. 7 and that of Fig. 8 at 500 episodes. Figure 8 shows a more accurate prediction of likely received power for different user type than that of Fig. 7 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 Fig.5., the proposed JMLS-DRL-scheme converges with almost a common and higher reward for all user types. The implication is that after 200 training episodes, the JMLS

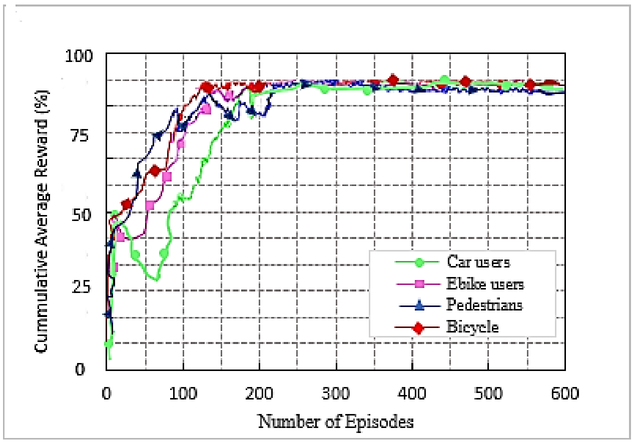

Figure 6. Compares the cumulative reward vs Number of training Episodes for our proposed JMLS-DDPG HO scheme according to user type.

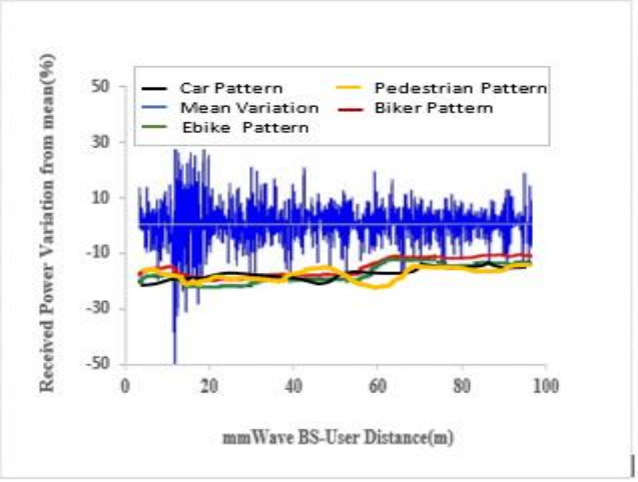

Figure 7. The actual average received power pattern variation at the UE in percentage about the mean value after 200 episodes in proposed JMLS-DDPG HO scheme.

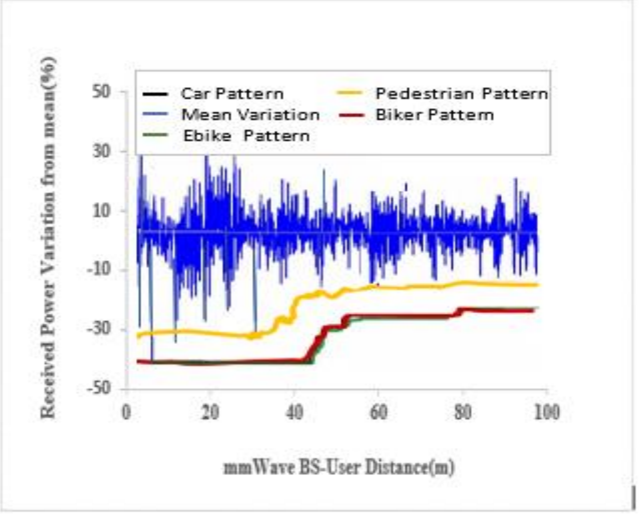

Figure 8. Shows the best expected received power pattern variation in percentage about the mean value after 500 episodes over 80m in proposed JMLS-DDPG HO scheme.

doi:10.20944/preprints202111.0514.v1

-DRL algorithm can have one common/global deterioration pattern to follow regardless of user type. On the other hand, for DDPG HO scheme each user type will need to follow a different type of deterioration pattern. This makes our proposed scheme easier to predict the expected target link behaviour than the later. In both, schemes, a 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 pattern in KMO are compared to 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 threshold. Therefore, HOs select a link based on the fact that it is expected to sustain connectivity at least for 80 m of assumed coverage of the mmWave BS. We also analysed a soft-HO DC-based scheme **[4]** using only SINR [2] and a DDPG [3] based scheme to make HO comparisons, and the former acted as a baseline for our case in Fig. 9 and Fig. 10.

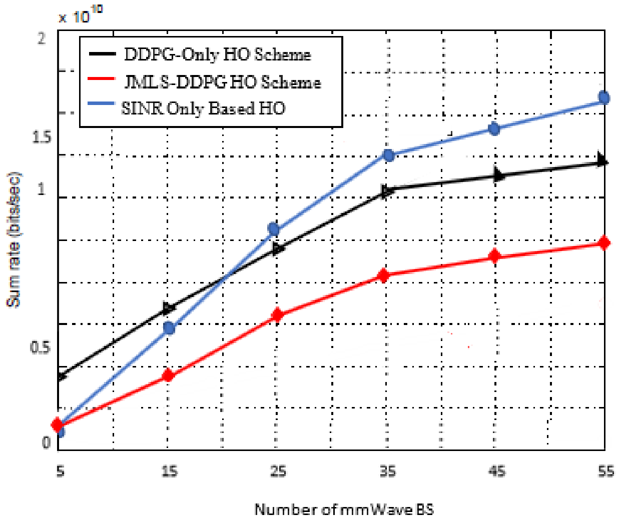

Figure 9. Compares sum rate for threes HO schemes vs Number of BSs.

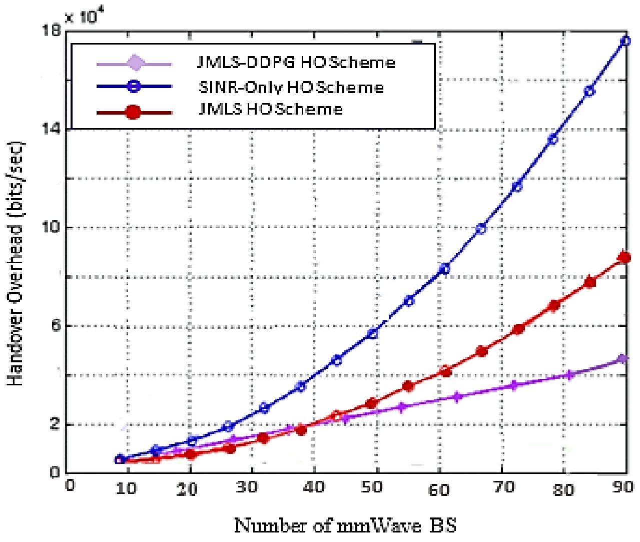

Fig. 10. Compares overhead vs Number of mmWave BSs.

In Fig. 9, we compare the sum rate against the number of BSs for three different HO schemes. SINR based scheme as explained in [4], just compares the SINR of the target and serving cell/link. The other scheme as earlier said gets new updates every episode whilst our proposed scheme uses both new and old CSI. We can see that the proposed scheme has a good efficiency on how it uses/selects BSs. The other two schemes seem to start saturating after 35-40 BSs. This can be attributed to low training sample requirement and thorough analysis of CSI in our scheme. Reuse of training sample gives our scheme ample time to analyse the behaviour of links. At the same time, having a small number of mmWave BS deprives the proposed scheme from learning more about target link deterioration pattern. This can be seen by the smaller sum date rate recorded at 5 to 15 mmWave BS. The more the mmWave BSs the diverse the amount of data looked at in each episode. On the other hand, despite a very small amount of BS, for DDPG only HO scheme, the acquisition of new training samples in each episode improves prediction of target link path but because it changes fast, the inaccuracy in the predictions quickly manifest.

Another criterion to evaluate the performance of the proposed HO methods, is the generated overhead. Fig. 10 shows the variation of the induced overhead related to 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 report must be exchanged to allocate new subcarriers resources. Nevertheless, using the DDPG-only based handover and tour proposed HO scheme, less overheads are experienced because the past link data needed to achieve reliability is reusable and exchanged in advance before the HO. For our proposed scheme, it is even much better because it can switch measurement data sources depending on conditions 6d (see Fig. 3). Hence the proposed scheme is better than both the DDPG only and SINR HO schemes.

5. Conclusion 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 multi-user set up was considered. We considered a downlink LTE-mmWave HetNet scenario. With a mmWave-link-behaviour pattern-analysis scheme applied to address the HO challenges. The resulting optimization solution considered modelling the link behaviour using JMLS, DRL and meta training techniques. Subsequently selection of optimal HO link used KMO test principles. Simulation results showed that our HO scheme outperformed DDPG only HO scheme and the SINR-only based HO scheme. This demonstrated the vital role deterioration pattern analysis can play in addressing mmWave link selection in 5G networks. Principally , we can conclude that our pattern analysis HO scheme envisages traits of long-term behaviour analysis for mmWave target links before HO execution . This is unlike unreliable techniques used in classic HO schemes where only the instantaneous behaviour of target links is analyzed prior to choosing the best target link . In future works, it would be interesting to consider the competing effects of pathloss, channel gain and transmission power when determining the receivable deterioration pattern of the target link. This is given the impact their variation have on the data rate. Further, while there is 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 behaviour pattern projections studies for target links. Finally, studying backhaul configurations that can efficiently support the proposed HO scheme would also be particularly interesting.

Author Contributions: This work was a collaborative effort of all authors. Specifically, conceptualization, Masoto Chiputa. and Peter Chong.; Methodology, Masoto Chiputa. and Peter Chong.; Software, Masoto Chiputa. and Peter Chong.; Validation, Minglong Zhang.; Hakilo Sabit and Peter Chong. Formal analysis, Masoto Chiputa.; G. Md. Nawaz Ali.; Hakilo Sabit. and Peter Chong.;

Investigation, Masoto Chiputa.; Minglong Zhang. and Peter Chong.; Writing original draft preparation , Masoto Chiputa.; Peter Chong and Minglong Zhang. ; Writing review and editing, Minglong Zhang. ; G. Md. Nawaz Ali.; Hakilo Sabit and Peter Chong.; Funding acquisition, Peter Chong. and Minglong Zhang. 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

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable

Data Availability Statement: Data sharing 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, 2018 IEEE Global Communications Conference (GLOBECOM), 2018, pp. 1-6, doi: 10.1109/GLOCOM.2018.8647528
2. Rangan,S.; Rappaport,T.S.; Erkip,E. Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges, in Proceedings of the IEEE, vol. 102, no. 3, pp. 366-385, March 2014, doi: 10.1109/JPROC.2014.2299397.
3. Dai,Y.; Xu,D.; Maharjan,S.; Zhang.Y. Joint load balancing and offloading in vehicular edge computing and networks, IEEE Internet Things J., vol. 6, no. 3, Jun. 2019, pp. 4377–4387.
4. Mwanje,S.; Zia,N.; Mitschele-Thiel,A. Self-Organized Handover Parameter Configuration for LTE , Proc. 9th International Symposium on Wireless Communication Systems (ISWCS'12), Paris, France, August 2012, pp. 26-30.
5. Shubyn,B.; Maksymyuk,T.Intelligent Handover Management in 5G Mobile Networks based on Recurrent Neural Networks, 2019 3rd International Conference on Advanced Information and Communications Technologies (AICT), 2019, pp. 348-351.
6. Shanmugam,K.; Golrezaei,N.; Dimakis,A.G.; Molisch,A.F.;Caire,G. Femtocaching: Wireless content delivery through distributed caching helpers, IEEE Trans. Inf. Theory, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.
7. Joud,M.; García-Lozano,M.; Ruiz,S. User specific cell clustering to improve mobility robustness in 5G ultra-dense cellular networks, 2018 14th Annual Conference on Wireless On-demand Network Systems and Services (WONS), 2018, pp. 45-50.
8. Blackmore,L.;Ono,M.;Bektassov,A.;Williams,B.C.A probabilistic particle-control approximation of chance-constrained stochastic predictive control.Tr.on Robotics,26(3),502–517,2010.
9. Chitraganti,S.;Aberkane,S.;Aubrun,C.;Valencia-Palomo,G.;and Dragan,V.On control of discrete-time state-dependent jump linear systems with probabilistic constraints :A receding horizon approach.Systems&ControlLetters,74,81–89,2014.
10. Zhou,Z.;Yu,H; Xu,C.;Zhang,Y.;Mumtaz,S.; and Rodriguez,J. Dependable content distribution in D2D-based cooperative vehicular networks: A big data-integrated coalition game approach, IEEE Trans. Intell. Transp. Syst., vol. 19, no. 3, pp. 953–964, Mar. 2018.
11. Zhou,Z.; Gao,C.; Xu.C.; Zhang,Y.; Mumtaz,S.; and Rodriguez,J. Social big-data-based content dissemination in Internet ofVehicles, IEEE Trans.Ind. Informat., vol. 14, no. 2, pp. 768–777, Feb. 2018.
12. 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., vol. 67, no. 9, pp. 1287–1300, Sep. 2018.
13. Maksymyuk,T.; Han,L.; Larionov,S.; Shubyn,B.; Luntovskyy,A. and Klymash,M. Intelligent Spectrum Management in 5G Mobile Networks based on Recurrent Neural Networks, 15th IEEE International Conference The Experience of Designing and Application of CADSystems (IEEE CADSM’2019), Polyana, Ukraine, February, 2019,
14. Rodrigues,T.G.; Suto,K.; Nishiyama,H.; and Kato,N. Hybrid method for minimizing service delay in edge cloud computing through vm migration and transmission power control, IEEE Trans. Comput., vol. 66, no. 5, pp. 810–819, May 2017.
15. Costa,O.L.V.; Fragoso,M.D.; andMarques,R.P. Discrete time Markov jump linear systems .Springer.,2005.
16. Edwards,C. Advanced Calculus of Several Vari-ables. Dover Publications.1973.
17. Kocvara,M.; and Stingl,M. Pennonacode for convex non linear and semi definite programming.Op-timization Methods and Software, 8(3),317–333,2003.
18. Dempster, A. P., N. M. Laird.; Rubin ,D.B. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society. Series B (methodological), 1–38,1977.
19. Bilmes, J. A. et al. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. International Computer Science Institute 4(510), 126,1998.
20. Costa, O. L. V.; Fragoso,M.D.; Marques,R.P . Discrete-time Markov jump linear systems. Springer Science & Business Media, 2006.
21. Zhou,Z.; Feng,J.; Chang,Z.; and Shen,X. Energy-efficient edge computing service provisioning for vehicular networks: A consensus admm approach, IEEE Trans. Veh. Technol., vol. 68, no. 5, pp. 5087–5099,May 2019.
22. Sorkhoh,I.; Ebrahimi,D.; Atallah,R.; and Assi,C. Workload scheduling in vehicular networks with edge cloud capabilities, IEEE Trans. Veh. Technol., vol. 68, no. 9, pp. 8472–8486, Sep. 2019.
23. Eason,G.; Noble,B.; Sneddon,I.N. On certain integrals of Lipschitz-Hankel type involving products of Bessel functions, Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.
24. Gawłowicz,P.; Zubow, ns-3 meets OpenAI Gym: The Playground for Machine Learning in Networking Research, ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, Nov 25–29, 2019, Miami Beach, USA. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/1122445.1122456.
25. Stephanie Glen. Kaiser-Meyer-Olkin (KMO) Test for Sampling Adequacy, From StatisticsHowTo.com: Elementary Statistics for the rest of us! https://www.statisticshowto.com/kaiser-meyer-olkin/

26. Shlezinger,N.; Farsad, N.;Eldar.Y.C.; Goldsmith,A.J. ViterbiNet: A Deep Learning Based Viterbi Algorithm for Symbol Detection, in IEEE Transactions on Wireless Communications, vol. 19, no. 5, pp. 3319-3331, May 2020,
27. Al-Nima, R.R.O.; Han, T.; Chen, T. Road tracking using deep reinforcement learning for self-driving car applications. Int. Conf. Comput. Recognit. Syst. 2019, doi:10.1016/j.future.2017.12.041.
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. https://doi.org/10.3390/s2101021.