Reinforcement Learning based Per-antenna Discrete Power Control for Massive MIMO Systems

Reinforcement Learning based Per-antenna Discrete Power Control for Massive MIMO Systems

Navneet Garg, Mathini Sellathurai†, Tharmalingam Ratnarajah
The University of Edinburgh, UK, †Heriot-Watt university, Edinburgh, UK.

***Abstract*—Power consumption is one of the major issues in massive MIMO (multiple input multiple output) systems, causing increased long-term operational cost and overheating issues. In this paper, we consider per-antenna power allocation with a given finite set of power levels towards maximizing the long-term energy efficiency of the multi-user systems, while satisfying the QoS (quality of service) constraints at the end users in terms of required SINRs (signal-to-interference-plus-noise ratio), which depends on channel information. Assuming channel states to vary as a Markov process, the constraint problem is modeled as an unconstraint problem, followed by the power allocation based on Q-learning algorithm. Simulation results are presented to demonstrate the successful minimization of power consumption while achieving the SINR threshold at users.**

I. Introduction

Massive MIMO systems are the central part of 5G and next generation wireless networks. Due to large number of antennas in the array, the increased power consumption i.e. reduced energy efficiency (EE), causes increased operational cost and overheating problems which leads to reduced lifespan of the array. The power allocation problem has been widely investigated in literature via different schemes such as antenna selection schemes [1]–[11], machine/deep learning (ML/DL) schemes [12]–[14], convex approximation based [15], [16], etc. In massive MIMO systems, transmit correlation with mutual coupling is studied in [17], while with hybrid precoding, power consumption cost is minimized in [18]. The antenna selection methods require NP-hard non-convex problem to be solved, and power allocation step is still needed, which reduces its preference of usage in practice. The drawback of ML/DL approaches is that they require the huge data for training and the optimal solution is not guaranteed. Convex-approximation based approaches approximate the non-convex EE expressions into convex ones and obtain sub-optimal power allocation. Therefore, a unified power allocation and antenna selection approach is essential in improving the energy efficiency.

In this paper, we present the discrete power allocation scheme using reinforcement Q-learning for downlink multi-user massive MIMO system towards that maximization of the long-term energy efficiency subject to the total power constraint, per-antenna power constraint, and the quality of service (QoS) constraints at the end users in terms of SINR. Discrete power allocation can also be considered as a generalization of antenna selection schemes, which has only two power levels. Assuming the channel changes as a Markov process in the time-slotted model with unknown transition probabilities, the long term energy efficiency maximization problem is presented subjected to total power constraint and per-antenna power constraint. This constraint problem is formulated as an unconstraint problem and Q-learning is used to obtain the solution. Simulation results demonstrate that Q-learning algorithm converges and minimizes the power consumption, while satisfying the QoS constraint at users.

The work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grants EP/P009549/1 and EP/P009670/1.

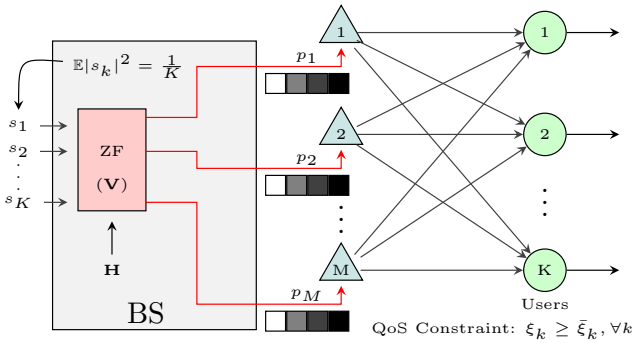

Fig. 1. BS with discrete power control $\mathbf{P} = \mathcal{D}\{p_1, \ldots, p_M\}$ and $\mathbb{E}\{\mathbf{s}\mathbf{s}^\dagger\} = \frac{1}{K}\mathbf{I}_K$ and $\mathbb{E}\{\mathbf{s}\} = \mathbf{0}$.

II. System Model

Consider a downlink multi-user system, where a base station (BS) is equipped with a large number of antennas (M). The BS serves simultaneously a set of K users indexed by $\mathcal{K} = \{1, \ldots, K\}$. The transmitted signal from the BS can be expressed as

$$\mathbf{x} = \mathbf{P}^{1/2} \sum_{k \in \mathcal{K}} \mathbf{v}_k s_k = \mathbf{P}^{1/2}\mathbf{V}\mathbf{s}, \tag{1}$$

where $\mathbf{s} = [s_1, \ldots, s_K]^T$ is $K \times 1$ symbol vector to be transmitted such that for each k^{th} user, $\mathbb{E}\{s_k\} = 0$, $\mathbb{E}\{s_k s_j^*\} = \frac{1}{K}\delta_{kj}$ and $\mathbb{E}\{\mathbf{s}\mathbf{s}^\dagger\} = \frac{1}{K}\mathbf{I}$ with δ_{kj} being the Kronecker delta having value 1 when $k = j$ and 0 otherwise; the matrix $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_K]$ is an $M \times K$ orthonormal precoder such that $\mathbf{V}^\dagger\mathbf{V} = \mathbf{I}_K$; the quantity $\mathbf{P} = \mathcal{D}(p_1, \ldots, p_M)$ is an $M \times M$ diagonal power allocation matrix with non-negative entries. Using the above, the per-antenna and the total power

constraints at the BS can be obtained as

$$T_{per}(p_m) = \left[\mathbb{E}\left\{\mathbf{x}\mathbf{x}^H\right\}\right]_{m,m} \quad (2)$$

$$= \frac{p_m}{K}\left[\mathbf{V}\mathbf{V}^H\right]_{m,m} \leq \bar{P}_m, \forall m \quad (3)$$

$$T_{tot}(\mathbf{P}) = \mathbb{E}\left\|\mathbf{x}\right\|^2 = \frac{1}{K}\text{tr}(\mathbf{P}\mathbf{V}\mathbf{V}^H) \leq \bar{P}_T, \quad (4)$$

where $\bar{P}_m$ and $\bar{P}_T$ are the m^{th} antenna and the total power constraints. For simplicity, we assume equal power constraint per antenna i.e. $\bar{P}_m = \bar{P}_{per}, \forall m = 1, \ldots, M$. Towards discrete power control, let the set $\mathcal{P} = \left\{p^{(1)}, \ldots, p^{(|\mathcal{P}|)}\right\}$ denote all the power levels for each antenna i.e. $p_m \in \mathcal{P}, \forall m$ and $\mathbf{P} \in \mathcal{P}^M$ such that $0 = p^{(1)} \leq \cdots \leq p^{(|\mathcal{P}|)} = \bar{P}_{per}$, where $\bar{P}_{per}$ also denotes the maximum power transmitted by a single antenna. Let $\mathbf{h}_k$ denote the channel state information (CSI) from BS at origin to the k^{th} user. Through this channel, the received signal at the k^{th} user can be written as

$$y_k = \mathbf{h}_k^H\mathbf{x} + n_k, \quad (5)$$

$$= \mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{V}\mathbf{s} + n_k, \quad (6)$$

$$= \mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{v}_k s_k + \mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{V}_{-k}\mathbf{s}_{-k} + n_k, \quad (7)$$

where $n_k \sim \mathcal{CN}(0, \sigma^2)$ is the circularly symmetric complex Gaussian noise; $\mathbf{V}_{-k} = [\mathbf{v}_1, \ldots, \mathbf{v}_{k-1}, \mathbf{v}_{k+1}, \ldots, \mathbf{v}_K]$ and $\mathbf{s}_{-k} = [s_1, \ldots, s_{k-1}, s_{k+1}, \ldots, s_K]^T$. At the k^{th}user, the resultant SINR can be given as

$$\xi_k(\mathbf{P}|\mathbf{H}) = \frac{\left|\mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{v}_k\right|^2 \frac{1}{K}}{\text{tr}\left(\mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{V}_{-k}\mathbf{V}_{-k}^H\mathbf{P}^{1/2}\mathbf{h}_k\right)\frac{1}{K} + \sigma^2}, \quad (8)$$

which depends on CSI $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K]$. Thus, stacking all the received signals gives $\mathbf{y} = \mathbf{H}^H\mathbf{P}^{1/2}\mathbf{V}\mathbf{s} + \mathbf{n}$. If the CSI variations follow a Markov process, the resultant SINR process will also be Markov. In other words, the power in the elements of $\mathbf{P}$ needs to be adjusted according to CSI to satisfy QoS constraints at the k^{th} user. Further, the achievable sum-rate is given as

$$R(\mathbf{P}|\mathbf{H}) = \sum_{k\in\mathcal{K}} \log_2\left(1 + \xi_k(\mathbf{P}|\mathbf{H})\right). \quad (9)$$

The resultant energy efficiency can be defined as the ratio of the sum rate over the total power incurred in the transmission as

$$\eta(\mathbf{P}|\mathbf{H}) = \frac{R(\mathbf{P}|\mathbf{H})}{T_{tot}(\mathbf{P})}, \quad (10)$$

where the circuit power is ignored as it is a constant. In the following, we simplify the sum rate for two popular precoding schemes based on ZF (zero-forcing) and MRT (maximal ratio transmission).

A. Zero-forcing

For zero forcing transmission, to find the precoder satisfying $\mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{v}_j = 0, \forall k \neq j$, we normalize the columns of $\mathbf{V}' = \mathbf{P}^{1/2}\mathbf{H}\left(\mathbf{H}^H\mathbf{P}\mathbf{H}\right)^{-1}$ to be unit norm columns. The above precoder results in the received signal $y_k = \mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{v}_k s_k + n_k$, resulting into the sum rate

$$R_{ZF}(\mathbf{P}|\mathbf{H}) = \sum_{k\in\mathcal{K}} \log_2\left(1 + \frac{\left|\mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{v}_k\right|^2}{\sigma^2 K}\right). \quad (11)$$

B. Maximal ratio transmission

For MRT based precoding, the precoder is set as $\mathbf{v}_k = \frac{\mathbf{P}^{1/2}\mathbf{h}_k}{\sqrt{\mathbf{h}_k^H\mathbf{P}\mathbf{h}_k}}$. Note that MRT precoding is used for low complexity operations, thus, the precoding vectors are not orthonormalized. The sum rate can be simplified as

$$R_{MRT}(\mathbf{P}|\mathbf{H}) = \sum_{k\in\mathcal{K}} \log_2\left(1 + \frac{\mathbf{h}_k^H\mathbf{P}\mathbf{h}_k}{\sum_{j\neq k}\frac{\mathbf{h}_k^H\mathbf{P}\mathbf{h}_j\mathbf{h}_j^H\mathbf{P}\mathbf{h}_k}{\mathbf{h}_j^H\mathbf{P}\mathbf{h}_j} + K\sigma^2}\right).$$

C. Problem formulation

Our goal is to maximize the energy efficiency of transmissions via discrete power allocations. However, note that in each time slot, finding discrete power levels for each antenna in the massive MIMO system is an NP-hard search problem and a non-convex problem. Moreover, the estimation of CSI in massive MIMO consumes resources. Therefore, for faster operations, utilizing the CSI correlation via Markov process, reinforcement learning is utilized to obtain these power levels. Thus, assuming the channel information varies as a finite state Markov chain, our objective to find the discrete power allocation to maximize the long term efficiency subject to the QoS constraints satisfied for each user, can be expressed as

$$\max_{\mathbf{P}(t)} \sum_{\tau=t}^{\infty} \gamma^{\tau-t}\eta(\mathbf{P}(t)|\mathbf{H}(t)) \quad (12)$$

$$\text{subject to } T_{tot}(\mathbf{P}(t), \mathbf{H}(t)) \leq \bar{P}_T, \mathbf{P}(t) \in \mathcal{P}^M,$$

$$\xi_k(\mathbf{P}(t)|\mathbf{H}(t)) \geq \bar{\xi}_k, \forall k \in \mathcal{K}, \quad (13)$$

where $\bar{\xi}_k$ represents the SINR requirements for QoS at k^{th} user, and (t) denotes their time dependent behavior. Note that the total power is also considered here a function of $\mathbf{H}$. It is due to the fact that the precoders are computed using channel information $\mathbf{H}$. This makes the problem non-convex and difficult to solve.

III. Reinforcement Learning

A. Dynamics of EEPA

We consider time varying channel across time slots. Within a time slot, the channel remains constant. The CSI in a cellular network varies if the user is walking, running or in a vehicle. In literature [19], [20], the time varying channel is modeled using a finite state Markov chain, where the ergodic channel in each time slot takes value in one of the Markov states. Let $\mathcal{H} = \left\{\mathbf{H}^{(1)}, \ldots, \mathbf{H}^{(|\mathcal{H}|)}\right\}$ denote the states in the Markov chain. The transition probability between channel states is fixed and unknown[1].

B. States, Actions and Rewards

For the above system dynamics, let $\underline{s}(t)$ be the state at time t, which is given as the CSI of the same slot as $\underline{s}(t) = \mathbf{H}(t) \in \mathcal{H}$. An action in the system corresponds to discrete power

[1]In some literature, first order auto-regressive process is used to model the channel variations due to the mobility, where the resulting channel model provides continuous state Markov process, rather than finite state chain.

control i.e. $\underline{a}(t) = \mathbf{P}(t) \in \mathcal{P}^M$. The action chosen is evaluated using the reward which is defined as the energy efficiency i.e.

$$\underline{r}(\underline{s}(t), \underline{a}(t)) = \frac{1}{\left(\sum_{k\in\mathcal{K}} \left|\xi_k(\mathbf{P}(t)|\mathbf{H}(t)) - \bar{\xi}_k\right|\right) T_{tot}(\mathbf{P}(t), \mathbf{H}(t))}, \tag{14}$$

where $|\cdot|$ ensures that the resulting SINR does not achieve values far from $\bar{\xi}_k$.

Here, the learner seeks the optimum action $\underline{a}(t)$ based on the previous observation $\mathbf{H}(t-1) = \underline{s}(t-1)$ by interactively making sequential decisions and observing the corresponding costs. In this way, the agent learns the best action policy against the random Markov chain transitions. Let the policy function be $\pi : \mathcal{H} \to \mathcal{P}^M$, which maps a state to an action. Under policy $\pi(\cdot)$, the power allocation is carried out via action $\underline{a}(t+1) = \pi(\underline{s}(t))$, dictating the allocation policy at time $t+1$. For the reward $\underline{r}_\pi(\underline{s}(t)) = \underline{r}(\underline{s}(t), \pi(\underline{s}(t)))$, power consumption performance is measured through the state value function as

$$V_\pi(\underline{s}(t)) = \sum_{\tau=t}^{\infty} \gamma^{\tau-T} \underline{r}_\pi(\underline{s}(t)), \tag{15}$$

which is the total average cost incurred over an infinite time horizon. The objective of this paper is to find the optimal policy π^* such that the average cost of any state is maximized

$$\pi^* = \arg\max_{\pi} V_\pi(\mathbf{S}).$$

C. Action set reduction

For the BS equipped with M antennas, there are huge number of $|\mathcal{P}|^M$ possible actions. However, not all actions are valid actions. Valid actions are those actions which satisfy the power constraint in (4). The total power $T_{tot}(\mathbf{P}, \mathbf{H})$ depends on the normalized precoder $\mathbf{V}$. To simplify the constraint in order to reduce the valid action set, we approximate the total power constraint as

$$\frac{1}{K}\text{tr}(\mathbf{P}\mathbf{V}\mathbf{V}^H) \approx \frac{1}{K}\text{tr}\left(\mathbf{P}\mathbb{E}\left\{\mathbf{V}_R\mathbf{V}_R^H\right\}\right) = \frac{1}{M}\text{tr}\left(\mathbf{P}\right) \leq \bar{P}_T, \tag{16}$$

where $\mathbf{V}_R$ is any random orthonormal precoder; the equality on the right follows from [21, Lem. 1]. Further, at least K actions should be non-zero i.e. $p_{i_k} > 0, \forall k \in \mathcal{K}$ that excludes $\sum_{k=1}^{K-1}\binom{M}{k}$ actions in $\mathcal{P}^M$. To get the minimum transmission power constraint to reduce huge number of possibilities, we approximate the QoS constraint as

$$\begin{aligned}\xi_k(\mathbf{P}|\mathbf{H}) &= \frac{\left|\mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{v}_k\right|^2}{\text{tr}\left(\mathbf{h}_k^H\mathbf{P}^{1/2}\mathbf{V}_{-k}\mathbf{V}_{-k}^H\mathbf{P}^{1/2}\mathbf{h}_k\right) + K\sigma^2}, \\ &\overset{(a)}{\approx} \frac{\text{tr}\left(\mathbf{P}\mathbf{v}_k\mathbf{v}_k^H\right)}{\text{tr}\left(\mathbf{P}\mathbf{V}_{-k}\mathbf{V}_{-k}^H\right) + K\sigma^2}, \\ &\overset{(b)}{\approx} \frac{\text{tr}\left(\mathbf{P}\right)\frac{1}{M}}{\text{tr}\left(\mathbf{P}\right)\frac{K-1}{M} + K\sigma^2} = \frac{1}{(K-1) + KM\frac{\sigma^2}{\text{tr}(\mathbf{P})}},\end{aligned}$$

where (a) follows from the massive MIMO channel hardening effect $\mathbf{h}_k\mathbf{h}_k^H \to \mathbf{I}_M$; and (b) follows similarly from (16). For ZF precoding, we have $\frac{1}{KM\frac{\sigma^2}{\text{tr}(\mathbf{P})}} \geq \bar{\xi}_k \implies \text{tr}\left(\mathbf{P}\right) \geq KM\sigma^2\bar{\xi}_k$. Let $\bar{P}_{\min}$ denote this lower bound on the transmission power. The new action space can now be expressed as

$$\bar{\mathcal{P}}_M = \left\{ \begin{pmatrix} p_1 \\ \vdots \\ p_M \end{pmatrix} : \begin{array}{c} \bar{P}_{\min} \leq \text{tr}(\mathbf{P}(t)) \leq M\bar{P}_T, \\ p_{i_k} > 0, \forall k \in \mathcal{K} \end{array} \right\}, \tag{17}$$

where $\bar{\mathcal{P}}_M \subset \mathcal{P}^M$. Note that the above approximations are to reduce the possible actions, and thus, it does not affect the optimal power allocations.

Algorithm 1 Q-learning algorithm.

Input: state $\underline{s}(0)$ randomly and $Q_0(\underline{s}, \underline{a}) = 0 \forall \underline{s}, \underline{a}$

1: **for** $t = 1, 2, \ldots,$ **do**

2: For given profile $\underline{s}(t-1)$, take action $\underline{a}(t)$ as

$$\underline{a}(t) = \begin{cases} \arg\max_{\underline{a}} Q_{t-1}(\underline{s}(t), \underline{a}) & \text{w.p. } 1-\epsilon \\ \text{random } \underline{a} \in \bar{\mathcal{P}} & \text{w.p. } \epsilon \end{cases}$$

3: Observe $\underline{s}(t)$ and compute $\underline{r}(\underline{s}(t), \underline{a}(t))$

4: Update

$$\begin{aligned} Q_t(\underline{s}(t), \underline{a}(t)) &= (1-\beta_t)Q_{t-1}(\underline{s}(t), \underline{a}(t)) \\ &+ \beta_t\left[\underline{r}(\underline{s}(t), \underline{a}(t)) + \gamma \max_{\underline{a}} Q_{t-1}(\underline{s}(t), \underline{a})\right]. \end{aligned} \tag{21}$$

5: **end for**

D. Bellman's Equations and Q-learning

Let $\Pr(\underline{s}, \underline{s}'|\underline{a})$ be the probability of transition from the current state $\underline{s}$ to the next state $\underline{s}'$ under action $\underline{a}$. Bellman equations express the state value functions in a recursive fashion as

$$V_\pi(\underline{s}) = \underline{r}_\pi(\underline{s}) + \gamma \sum_{\underline{s}'\in\Xi^K} \Pr(\underline{s}, \underline{s}'|\pi(\underline{s}))V_\pi(\underline{s}'), \forall \underline{s} \tag{18}$$

$$Q_\pi(\underline{s}, \underline{a}) = \underline{r}_\pi(\underline{s}) + \gamma \sum_{\underline{s}'\in\Xi^K} \Pr(\underline{s}, \underline{s}'|\underline{a})V_\pi(\underline{s}'), \forall \underline{s}, \underline{a}. \tag{19}$$

The above equations can be used to obtain the optimal policy by minimizing Q-function as

$$\pi^* = \arg\max_{\underline{a}} Q_\pi(\underline{s}, \underline{a}), \forall \underline{s}, \tag{20}$$

where under π^*, $V_{\pi^*}(\underline{s}) = \max_{\underline{a}} Q_\pi(\underline{s}, \underline{a})$ and it gives the solution

$$\begin{aligned} Q_\pi(\underline{s}, \underline{a}) &= \underline{r}_\pi(\underline{s}) + \gamma \sum_{\underline{s}'\in\Xi^K} \Pr(\underline{s}, \underline{s}'|\underline{a}) \max_{\underline{a}} Q_\pi(\underline{s}, \underline{a}), \\ &= \sum_{\underline{s}'\in\Xi^K} \Pr(\underline{s}, \underline{s}'|\underline{a})\left[\underline{r}(\underline{s}, \underline{a}) + \gamma \max_{\underline{a}} Q_\pi(\underline{s}, \underline{a})\right]. \end{aligned}$$

The above solution demands an iterative solution for Q-function, which is given in Algorithm 1.

In a time slot t, after observing the state $\underline{s}(t)$, the ϵ-greedy action $\underline{a}(t)$ is taken and instantaneous cost $\underline{r}(\underline{s}(t), \underline{a}(t)) + \gamma \max_{\underline{a}} Q(\underline{s}(t), \underline{a})$ is incurred. Under mean squared error

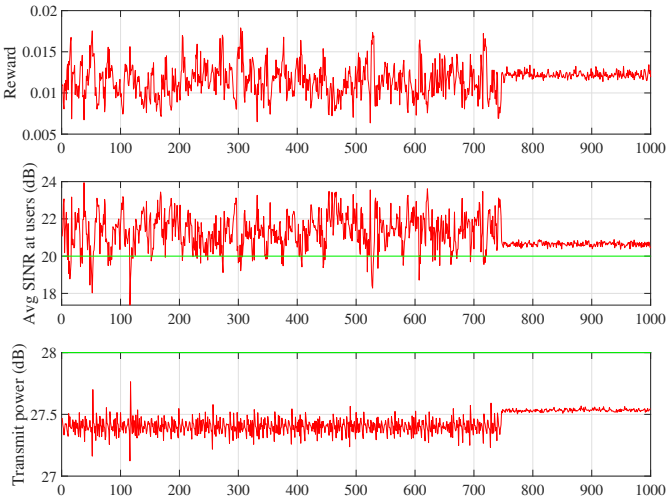

Fig. 2. Progress of average rewards, average SINR at users, and average transmit power at BS for different iterations for $M = 16$ and $|\mathcal{P}| = 3$ levels with per-antenna constraint, total transmit power constraint and user-SINR constraint of 30 dB, 28 dB and 20 dB (green lines) respectively.

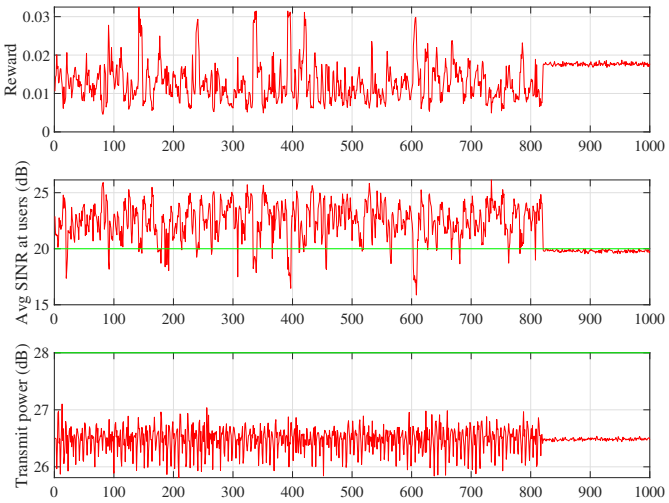

Fig. 3. Progress of average rewards, average SINR at users, and average transmit power at BS for different iterations for $M = 8$ and $|\mathcal{P}| = 5$ levels with per-antenna constraint, total transmit power constraint and user-SINR constraint of 30 dB, 28 dB and 20 dB respectively. .

(MSE) criteria, the MSE expression for the estimated Q-function values can be written as

$$\epsilon(\underline{s}(t),\underline{a}(t)) = \Bigg[\underline{r}(\underline{s}(t),\underline{a}(t)) + \gamma \max_{\underline{a}} Q(\underline{s}(t),\underline{a}) - Q(\underline{s}(t),\underline{a}(t))\Bigg]^2.$$

Minimizing the above error expression for Q-values using gradient descent method yields the following

$$Q_t(\underline{s}(t),\underline{a}(t)) = (1-\beta_t)Q_{t-1}(\underline{s}(t),\underline{a}(t)) + \beta_t\left[\underline{r}(\underline{s}(t),\underline{a}(t)) + \gamma \max_{\underline{a}} Q_{t-1}(\underline{s}(t),\underline{a})\right], \quad (22)$$

where Q_t is estimated Q-values at time t. It can be noted that the convergence of the algorithm depends on the values of β_t. Choosing β_t such that $\sum_t \beta_t < \infty$ guarantees the convergence. These cases of convergence and several other related algorithms has been thoroughly studied in [22].

Note that the cardinality of action space is increased exponentially for increase in the number of antennas and the number of power levels. Therefore, to make it scalable, deep reinforcement learning based methods will be investigated as a part of future work.

IV. Simulation Results

The following values are assumed for Q-learning parameters: $M = 8, 16$ antennas; $K = 4$ downlink users $|\mathcal{P}| = 3, 5$; 1000 number of episodes for Q-learning with each episode having 2000 iterations; exploration decay factor per episode 0.1; transmit power constraint 28 dB; per-antenna maximum power constraint 30 dB; QoS constraint for SINR 20 dB; number of channel states $|\mathcal{H}| = 128$. Zero-forcing based precoding is assumed since M is not high enough and the present Q-learning algorithm is computationally time consuming.

Figure 2 shows the plots for the progresses of average reward over iterations, average SINR across users and iterations, and average transmit power across iterations, respectively for $M = 16$ antennas at BS and $|\mathcal{P}| = 3$ power levels for each antenna. The action set is reduced from 3^{16} to around 12000 entries. It can be seen that the Q-learning learns the optimum power allocation in terms of reward, and the learned actions provide SINR greater than the QoS constraint for each user, keeping the transmit power within the constraint. Due to larger size of Q-matrix, it takes around 750 iterations to learn the optimum converging action. Similar trends can be seen for the case, when five power levels are assumed as shown in Figure 3. It shows the successful application of Q-learning in quickly finding the optimum power allocation among such a large set of possibilities $3^{16} \approx 4 \times 10^7$.

V. Conclusion

In this paper, we have presented reinforcement learning solution for discrete power allocation, which is a combinatorial optimization problem and is NP-hard. By leveraging the correlation between channels for slowing moving scenarios in wireless cellular networks, we model the channel variations as a finite state Markov chain and presented the RL formulation where the constraints are transmit power constraint and the Quality of service guarantee in terms of received SINR at each user with an objective of maximizing the energy efficiency at the transmitter. Typically, to handle the constraints in Q-learning, primal-dual approaches are used. However, we model the reward function to incorporate these constraints, without needing any additional dual variables in design. Simulations shows the successful application of the power allocation while satisfying these constraints.

The future work is to make the algorithm scalable for larger number of power levels and larger number of antennas.

REFERENCES

[1] X. Gao, O. Edfors, F. Tufvesson, and E. G. Larsson, "Massive MIMO in real propagation environments: Do all antennas contribute equally?" *IEEE Transactions on Communications*, vol. 63, no. 11, pp. 3917–3928, 2015.

[2] Z. Liu, W. Du, and D. Sun, "Energy and spectral efficiency tradeoff for massive MIMO systems with transmit antenna selection," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 5, pp. 4453–4457, May 2017.

[3] A. Garcia-Rodriguez, C. Masouros, and P. Rulikowski, "Reduced switching connectivity for large scale antenna selection," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2250–2263, May 2017.

[4] H. Tang and Z. Nie, "RMV antenna selection algorithm for massive MIMO," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 239–242, Feb 2018.

[5] M. Olyaee, M. Eslami, and J. Haghighat, "An energy-efficient joint antenna and user selection algorithm for multi-user massive MIMO downlink," *IET Communications*, vol. 12, no. 3, pp. 255–260, 2018.

[6] M. Hanif, H. Yang, G. Boudreau, E. Sich, and H. Seyedmehdi, "Antenna subset selection for massive MIMO systems: A trace-based sequential approach for sum rate maximization," *Journal of Communications and Networks*, vol. 20, no. 2, pp. 144–155, April 2018.

[7] A. Konar and N. D. Sidiropoulos, "A simple and effective approach for transmit antenna selection in multiuser massive MIMO leveraging submodularity," *IEEE Transactions on Signal Processing*, vol. 66, no. 18, pp. 4869–4883, Sep. 2018.

[8] H. Li, J. Cheng, Z. Wang, and H. Wang, "Joint antenna selection and power allocation for an energy-efficient massive MIMO system," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 257–260, Feb 2019.

[9] D. Park, "Sum rate maximisation with transmit antenna selection in massive MIMO broadcast channels," *Electronics Letters*, vol. 54, no. 21, pp. 1245–1247, 2018.

[10] S. Asaad, A. M. Rabiei, and R. R. MÃŒller, "Massive MIMO with antenna selection: Fundamental limits and applications," *IEEE Transactions on Wireless Communications*, vol. 17, no. 12, pp. 8502–8516, Dec 2018.

[11] W. A. Al-Hussaibi and F. H. Ali, "Efficient user clustering, receive antenna selection, and power allocation algorithms for massive MIMO-NOMA systems," *IEEE Access*, vol. 7, pp. 31 865–31 882, 2019.

[21] N. Garg and G. Sharma, "Analog precoder feedback schemes with interference alignment," *IEEE Transactions on Wireless Communications*, vol. 17, no. 8, pp. 5382–5396, 2018.

[12] H. Huang, Y. Song, J. Yang, G. Gui, and F. Adachi, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3027–3032, March 2019.

[13] S. Zhang, C. Xiang, S. Cao, S. Xu, and J. Zhu, "Dynamic carrier to MCPA allocation for energy efficient communication: Convex relaxation versus deep learning," *IEEE Transactions on Green Communications and Networking*, vol. 3, no. 3, pp. 628–640, Sep. 2019.

[14] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2200–2210, Oct 2019.

[15] K. Singh, K. Wang, S. Biswas, Z. Ding, F. A. Khan, and T. Ratnarajah, "Resource optimization in full duplex non-orthogonal multiple access systems," *IEEE Transactions on Wireless Communications*, vol. 18, no. 9, pp. 4312–4325, 2019.

[16] L. Li, F. Khan, M. Pesavento, T. Ratnarajah, and S. Prakriya, "Sequential search based power allocation and beamforming design in overlay cognitive radio networks," *Elsevier Signal Process.*, vol. 97, no. C, pp. 221–231, Apr. 2014.

[17] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Large-scale mimo transmitters in fixed physical spaces: The effect of transmit correlation and mutual coupling," *IEEE Transactions on Communications*, vol. 61, no. 7, pp. 2794–2804, 2013.

[18] S. Payami, M. Ghoraishi, M. Dianati, and M. Sellathurai, "Hybrid beamforming with a reduced number of phase shifters for massive mimo systems," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 6, pp. 4843–4851, 2018.

[19] X. Liu, Z. Qin, Y. Gao, and J. A. McCann, "Resource allocation in wireless powered iot networks," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 4935–4945, June 2019.

[20] F. Sangare, D. H. N. Nguyen, and Z. Han, "Learning frameworks for dynamic joint RF energy harvesting and channel access," *IEEE Access*, vol. 7, pp. 84 524–84 535, 2019.

[22] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Transactions on Automatic Control*, vol. 42, no. 5, pp. 674–690, 1997.