

Resource Allocation Scheme in Multi-Antenna Systems With Hybrid Energy Supply

Delin Guo, Lan Tang^{ID}, *Member, IEEE*, and Xinggan Zhang^{ID}

Abstract—In this letter, we study the resource allocation problem in a multiuser multi-antenna system, in which the energy supply of the transmitter consists of the grid energy and harvested energy. Our objective is to maximize the long-term sum throughput under the constraint of energy supply by optimizing beamforming vectors and energy allocation. Considering the challenges of imperfect channel state information (CSI) and large action/state spaces, we propose a dimension reduction deep reinforcement learning (RL) method to solve the optimization problem. In the proposed algorithm, the beamforming vectors are first determined based on imperfect CSI, and then policy-based RL is employed to find the optimal mapping between transmit powers and the low dimensional system state. Simulation results demonstrate the superiority of the proposed algorithm over traditional ones in terms of steady-state performance and learning speed.

Index Terms—Energy harvesting, energy allocation, beamforming, deep RL.

I. INTRODUCTION

WITH the development of energy harvesting (EH) technology, some sustainable energies, such as wind energy and solar energy, have been employed to charge the small base stations (BS) and low power terminal nodes. However, the efficiency of energy harvesting is heavily influenced by the environment and typically quite low. Hence, how to utilize the harvested energy fully to effectively improve the performance of communication systems needs to be studied.

Some literatures have focused on the optimization of transmit power taking account of EH constraints. Assuming that harvested energy accounts were known in advance, [1] developed optimal off-line scheduling policies to minimize the transmission time. Reference [2] proposed water-filling energy allocation solution to maximize the throughput when the environment was highly predictable. When the transmitter knew the dynamic characteristic on the channel state and harvested energy, dynamic programming was used to determine the optimal transmission policy. Reference [2] maximized the sum throughput over a finite horizon by choosing a deterministic power allocation policy via dynamic programming. In [3], the throughput over a finite horizon was maximized under a non-fading channel scenario, where the EH transmitter operated on a harvest-use-store model with the non-ideal rechargeable battery. Reference [4] studied the power control for outage

probability minimization when future information of EH rate was unknown prior to the transmission. Reference [5] solved the average rate maximization problem via dynamic programming algorithm when the channel fading and the energy arrival variations are of different timescales. References [6], [7] investigated the energy allocation problem in the energy harvesting system when the transmitter only obtained imperfect channel state information (CSI). When the transmitter did not know any priori information on the environment dynamic process, [8] introduced a learning theoretic approach to optimize the transmission policy in a point-to-point wireless communication system.

In this letter, we apply the learning theoretic approach to the resource allocation in a multiuser multi-antenna system, in which the transmitter is powered by the grid and harvested energy. Our goal is to maximize the long-term sum throughput under the constraint of energy supply by optimizing beamforming vectors and transmit powers. We assume the transmitter can only obtain the imperfect CSI and do not know anything about the dynamic process of the environment. Therefore, this problem is modelled as a model-free reinforcement learning (RL) problem. To improve the performance of RL with large state/action spaces, we propose a dimension reduction deep RL to optimize the beamforming vectors and transmit powers. In the proposed method, we first optimize the beamforming vectors based on estimated CSI. With the given beamforming vectors, the high dimensional system state is transformed into the low dimensional state so that deep RL-Proximal Policy Optimization (PPO) can be used to obtain the optimal mapping between transmit powers and system state. Compared with the traditional method, the proposed method can achieve higher throughput and faster learning speed.

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. System Model

We consider a multiuser communication system, which consists of a transmitter with M_t transmit antennas and K terminals with single receive antenna. The transmitter is equipped with an energy harvesting device so that energy supply of the transmitter can be from grid or battery. In the n th time slot, the received signal vector is given by $\mathbf{y}(n) = \mathbf{H}(n)\mathbf{W}(n)\mathbf{x}(n) + \mathbf{z}(n)$ where $\mathbf{x}(n) = [x_1(n), \dots, x_K(n)]^T$, $\mathbf{W}(n) = [\mathbf{w}_1(n), \dots, \mathbf{w}_K(n)]$, $\mathbf{H}(n) = [\mathbf{h}_1(n), \dots, \mathbf{h}_K(n)]^T$, $x_k(n)$, $\mathbf{w}_k(n)$, $\mathbf{h}_k(n)$ are the transmit signal, beamforming vector and channel vector of terminal k , respectively, and $\mathbf{z}(n) = [z_1(n), \dots, z_K(n)]^T$ is the additive white Gaussian noise (AWGN) vector. Beamforming vectors for K terminals are chosen from the predefined codebook $\mathcal{W}$. Define $\gamma_{kk'}(n) = |\mathbf{h}_k^T(n)\mathbf{w}_{k'}(n)|^2$, $\hat{\gamma}_{kk'}(n) = |\hat{\mathbf{h}}_k^T(n)\mathbf{w}_{k'}(n)|^2$. The signal to interference noise ratio (SINR) of terminal k is

$$\text{SINR}_k(n) = \frac{\gamma_{kk}(n)p_k(n)}{\sigma^2 + \sum_{k' \neq k} \gamma_{kk'}(n)p_{k'}(n)} \quad (1)$$

Manuscript received October 26, 2020; accepted November 12, 2020. Date of publication November 17, 2020; date of current version March 9, 2021. This work was supported by the Natural Science Foundation of China under Grant 61976113, Grant 62072229, and Grant U1936201. The associate editor coordinating the review of this article and approving it for publication was C. Shen. (Corresponding author: Lan Tang.)

The authors are with the School of Electronic Science and Engineering, Nanjing University, Nanjing 210093, China (e-mail: tanglan@nju.edu.cn).

Digital Object Identifier 10.1109/LWC.2020.3038418

2162-2345 © 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

where $p_k(n)$ is the transmit power allocated to terminal k and σ^2 is the noise power. The transmit power $p_k(n) = p_k^b(n) + p_k^g(n)$, where $p_k^b(n)$ and $p_k^g(n)$ are from battery and grid, respectively. When the energy in battery is used, the battery energy changes according to

$$B(n+1) = \min([0, B(n) - T \sum_k p_k^b(n)]^+ + G(n), B_{max}) \quad (2)$$

where $[0, x]^+ = \max(x, 0)$, $B(n)$ and $G(n)$ denote the battery energy and harvested energy at time n , respectively, T is the duration of transmission and B_{max} is the battery storage capacity. We assume that the energy harvesting dynamic process is only affected by environmental conditions and independent of channel dynamic processes.

In time division duplexing (TDD) systems, the transmitter can estimate $\mathbf{H}(n)$ utilizing the uplink pilot signals. With minimum mean squared error (MMSE) estimator, the estimated channel matrix is given by

$$\hat{\mathbf{H}}(n) = \sqrt{1 - \rho} \mathbf{H}(n) + \sqrt{\rho} \Phi(n) \quad (3)$$

where ρ denotes correlation coefficient of $\hat{\mathbf{H}}(t)$ and $\mathbf{H}(t)$, and Φ is the observation error. We assume that $\mathbf{H}(n)$ and $\Phi(n)$ are independent of each other, and elements in $\mathbf{H}(n)$ and $\Phi(n)$ are complex Gaussian variables with zero mean and unit variance.

In each time slot, the transmitter decides the transmit powers and beamforming vectors $\{p_k^b(n), p_k^g(n), \mathbf{w}_k(n)\}_{k=1}^K$ based on the system state $s(n) = (B(n), \hat{\mathbf{H}}(n), G(n))$. Since the transmitter only knows $\hat{\mathbf{H}}(n)$, once $\{p_k^b(n), p_k^g(n), \mathbf{w}_k(n)\}_{k=1}^K$ are determined, the reward of terminal k is defined as the following average rate conditioned on $\{\hat{\gamma}_{kk'}(n)\}_{k'=1}^K$

$$I_k(n) = \mathbb{E}_{\{\gamma_{kk'}\}_{k'=1}^K | \{\hat{\gamma}_{kk'}\}_{k'=1}^K} [\log(1 + \text{SINR}_k(n))]. \quad (4)$$

It is difficult to obtain the closed-form expression of $I_k(n)$ due to the complicated conditional probability density function (pdf) of $\{\gamma_{kk'}\}_{k'=1}^K$ and $\{\hat{\gamma}_{kk'}\}_{k'=1}^K$. However, we can obtain the lower bound function of $I_k(n)$ by the following inequality $I_k(n) > \mathbb{E}_{\gamma_{kk} | \hat{\gamma}_{kk}} [\log(\gamma_{kk}(n) p_k(n))] - \log(\sigma^2 + \sum_{k' \neq k} \mathbb{E}_{\gamma_{kk'} | \hat{\gamma}_{kk'}} [\gamma_{kk'}(n) p_{k'}(n)])$. When $\mathbf{w}_k(n)$ is given, $\gamma_{kk}(n)$ and $\hat{\gamma}_{kk}(n)$ obey standard chi-square distribution $\chi^2(2)$ and $\hat{\gamma}_{kk} | \gamma_{kk}$ obeys noncentral chi-square distribution. Then, on the basis of the Bayesian formula, we can obtain the pdf of γ_{kk} , given that $\hat{\gamma}_{kk} = \hat{\gamma}$

$$f_{\gamma_{kk} | \hat{\gamma}_{kk}}(\gamma | \hat{\gamma}) = \frac{1}{\rho} \exp\left(\frac{\hat{\gamma}(\rho - 1) - \gamma}{\rho}\right) I_0\left(\frac{\sqrt{(1 - \rho)\hat{\gamma}\gamma}}{\rho/2}\right) \quad (5)$$

where $I_0(\cdot)$ is the zeroth-order modified Bessel function of the first kind. $f_{\gamma_{kk'} | \hat{\gamma}_{kk'}}(\gamma | \hat{\gamma})$ ($k \neq k'$) has similar form to that in (5). Hence, $\mathbb{E}_{\gamma_{kk'} | \hat{\gamma}_{kk'}}[\gamma_{kk'}] = \rho + (1 - \rho)\hat{\gamma}_{kk'}$. Based on the above analysis, the lower bound of $I_k(n)$ is given by

$$I_k^l(n) = \int_0^{+\infty} \log(\gamma p_k(n)) f_{\gamma_{kk} | \hat{\gamma}_{kk}}(\gamma | \hat{\gamma}) d\gamma - \log\left(\sigma^2 + \sum_{k' \neq k} (\rho + (1 - \rho)\hat{\gamma}_{kk'}(n)) p_{k'}(n)\right). \quad (6)$$

In the infinite horizon, the discounted sum throughput and consumption of grid energy can be calculated as

$$\bar{I} = \mathbb{E} \left[\sum_{n=0}^{\infty} \beta^n \left(\sum_{k=1}^K I_k^l(n) \right) | s(0) \sim f_0(s(0)) \right] \quad (7)$$

and

$$\bar{D} = \mathbb{E} \left[\sum_{n=0}^{\infty} \beta^n \left(\sum_{k=1}^K p_k^g(n) \right) | s(0) \sim f_0(s(0)) \right] \quad (8)$$

respectively, where $\beta \in [0, 1)$ is the discount factor, $f_0(\cdot)$ is the initial state distribution, and the expectation is taken over all states $\{s(n)\}_{n=0}^{\infty}$.

Our objective is to maximize $\bar{I}$ under the constraint of energy supply by optimizing $\{p_k^b(n), p_k^g(n), \mathbf{w}_k(n)\}_{k=1}^K$. The problem can be formulated as

$$(P0): \max_{\{p_k^b(n), p_k^g(n), \mathbf{w}_k(n)\}_{k=1}^K} \bar{I} \quad (9a)$$

$$s.t. \quad \bar{D} \leq D_0, \quad (9b)$$

$$T \sum_{k=1}^K p_k^b(n) \leq B(n), \quad \forall n \quad (9c)$$

$$\sum_{k=1}^K p_k^g(n) \leq p_p, \quad \forall n \quad (9d)$$

$$\mathbf{w}_k(n) \in \mathcal{W}, \quad \forall n \quad (9e)$$

where D_0 is the average power constraint from grid and p_p is the peak power constraint from the grid.

III. SOLUTION TO (P0)

In this section, we will study the solution to (P0) by assuming unknown system dynamic. Firstly, we construct a partial Lagrangian function

$$J = \bar{I} - \lambda \bar{D} = \mathbb{E} \left[\sum_{n=0}^{\infty} \beta^n R(n) | s(0) \sim f_0(s(0)) \right] \quad (10)$$

where λ is the Lagrange multiplier and $R(n) = \sum_{k=1}^K (I_k^l(n) - \lambda p_k^g(n))$ is immediate reward function. For a given λ , the problem in (P0) can be transformed into the following problem

$$(P1): \min_{\{p_k^b(n), p_k^g(n), \mathbf{w}_k(n)\}_{k=1}^K} J \quad s.t. \quad (9c) - (9e). \quad (11)$$

Let $\mathcal{A}$ denote the set of feasible actions satisfying (9c)–(9e). That is, $\{p_k^b(n), p_k^g(n), \mathbf{w}_k(n)\}_{k=1}^K \in \mathcal{A}$. Considering the continuous and high-dimensional state and action spaces, the value-based RL algorithms, such as Q-learning or deep Q-learning, are not suitable for solving (P1). Therefore, we adopt the policy-based RL algorithm to solve (P1). For example, PPO can be applied to obtain the optimal mapping between high dimensional state $s(n)$ and actions in $\mathcal{A}$. However, in the scenario with growing number of antennas or users, the explicit application of PPO to this problem will be inefficient due to the ‘‘curse of high-dimensionality’’, i.e., the learning performance will deteriorate with the increasing of M_t or K .

Based on the above observation and analysis, a dimension reduction deep RL is proposed to solve (P1). In this method, we first determine $\{\mathbf{w}_k(n)\}_{k=1}^K$ based on $\hat{\mathbf{H}}(n)$ to reduce the dimensions of both state and action spaces, and then optimize transmit powers by PPO algorithm. From (6), it is observed that when $\{\mathbf{w}_k(n)\}_{k=1}^K$ are given, $\{p_k^g(n), p_k^b(n)\}_{k=1}^K$ depend on simplified system state $\hat{s}(n) = (B(n), \{\hat{\gamma}_{kk'}(n)\}_{k,k'=1}^K, G(n))$. However, the optimal $\{\mathbf{w}_k(n)\}_{k=1}^K$ which maximize (6) are related

to $\{p_k^g(n), p_k^b(n)\}_{k=1}^K$. To decouple the optimization of $\{\mathbf{w}_k(n)\}_{k=1}^K$ and $\{p_k^g(n), p_k^b(n)\}_{k=1}^K$, we first optimize $\mathbf{w}_k(n)$ based on a heuristic metric $\text{SIR}_k(n) = \frac{\gamma_{kk}(n)}{\sum_{k' \neq k} \gamma_{k'k}(n)}$, which denotes the ratio of signal power $\gamma_{kk}(n)p_k(n)$ to the interference of terminal k to other terminals $\sum_{k' \neq k} |\mathbf{h}_{k'}^T(n)\mathbf{w}_k(n)|^2 p_{k'}(n)$. Replacing $\text{SINR}_k(n)$ in (4) with $\text{SIR}_k(n)$, we have $\mathbb{E}_{\{\gamma_{k'k}\}_{k'=1}^K | \{\hat{\gamma}_{k'k}\}_{k'=1}^K} [\log(1 + \frac{\gamma_{kk}}{\sum_{k' \neq k} \gamma_{k'k}})] \geq \mathbb{E}_{\gamma_{kk} | \hat{\gamma}_{kk}} [\log(\gamma_{kk})] - \log(\sum_{k' \neq k} (\rho + (1 - \rho)\hat{\gamma}_{k'k}))$. Therefore, $\mathbf{w}_k(n)$ is determined by

$$\mathbf{w}_k(n) = \arg \max_{\mathbf{w}_k(n) \in \mathcal{W}} \int_0^{+\infty} \log(\gamma) f_{\gamma_{kk} | \hat{\gamma}_{kk}}(\gamma | \hat{\gamma}) d\gamma - \log \left(\sum_{k' \neq k} (\rho + (1 - \rho)\hat{\gamma}_{k'k}(n)) \right). \quad (12)$$

After getting $\{\mathbf{w}_k(n)\}_{k=1}^K$, the optimal mapping between simplified state $\hat{s}(n)$ and action $a_k(n) = \{p_k^b(n), p_k^g(n)\}_{k=1}^K$ can be found by PPO. First, the action-value function and state-value function are respectively defined as

$$Q^\pi(\hat{s}(n), a(n)) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \beta^t R(n+t) | \hat{s}(n), a(n) \right], \quad (13)$$

$$V^\pi(\hat{s}(n)) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \beta^t R(n+t) | \hat{s}(n) \right], \quad (14)$$

where $\pi(a(n) | \hat{s}(n))$ represents the probability distribution (or policy) that the action $a(n)$ is subject to.

In PPO, the state-value function $V^\pi(\hat{s}(n))$ is estimated by a critic – a parameterized value function $V_\omega(\hat{s}(n))$, where ω s are parameters, and the action-value function $Q^\pi(\hat{s}(n), a(n))$ is estimated by temporal-difference (TD) method,

$$\hat{Q}(n) = R(n) + \beta V_\omega(\hat{s}(n+1)). \quad (15)$$

Besides the state-value function, the policy is also parameterized. The given current policy is denoted as $\pi_{\theta_{old}}$, and the policy to be updated is π_θ , where θ_{old} and θ are their parameters respectively. Let the estimated advantage function $\hat{A}(n) = \hat{Q}(n) - V_\omega(\hat{s}(n))$, and $\text{clip}^\varepsilon(x)$ denote the function which limits x to interval $[1-x, 1+x]$. According to [9], for given $\pi_{\theta_{old}}$, the policy is updated by maximizing the following objective with the gradient ascent method,

$$L(\theta) = \hat{\mathbb{E}}_n [\min(r_\theta(n)\hat{A}(n), \text{clip}^\varepsilon(r_\theta(n))\hat{A}(n))], \quad (16)$$

where the policy ratio $r_\theta(n) = \frac{\pi_\theta(a(n) | \hat{s}(n))}{\pi_{\theta_{old}}(a(n) | \hat{s}(n))}$, and $\hat{\mathbb{E}}_n[\dots]$ is the sample average, where the sample data are obtained by implementing $\pi_{\theta_{old}}$ in the practical system. The parameters in the critic are updated by minimizing the following TD loss function,

$$L(\omega) = \hat{\mathbb{E}}_n [(\hat{Q}(n) - V_\omega(\hat{s}(n)))^2]. \quad (17)$$

By executing (12) and PPO algorithm, we can solve the problem (P1) and get the suboptimal policy for given λ . The whole procedure for solving (P1) is given by Algorithm 1. However, we can not get $\bar{D}$ and $\bar{I}$ in an offline method due to the unknown state transition probability. Therefore, we use the TD prediction method to evaluate $\bar{D}$ and $\bar{I}$ in the practical system. In (P1), to search for the optimal λ which makes

Algorithm 1 Dimension Reduction PPO Algorithm for Solving (P1)

Initiate $V_\omega, \pi_\theta, \pi_{\theta_{old}}, \alpha_\theta, \alpha_\omega$, number of iteration K , $\theta_{old} = \theta$.
 for iteration = 1, 2, ..., K
 for an episode $n=1, 2, \dots, N$
 Determine $\{\mathbf{w}_k(n)\}_{k=1}^K$ according to (12),
 Choose an action according to $\pi_{\theta_{old}}(a(n) | \hat{s}(n))$,
 Get the reward $R(n)$, and the next observation $\hat{s}(n+1)$.
 end for
 Get a trajectory: $\tau = \{\hat{s}(n), a(n), R(n)\}_{n=1}^N$.
 Compute $\{\hat{A}(n), \hat{Q}(n)\}_{n=1}^N$.
 Update parameters using the obtained data:
 $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta L(\theta)$
 $\omega \leftarrow \omega + \alpha_\omega \nabla_\omega L(\omega)$
 update $\theta_{old} \leftarrow \theta$
 end for

(9b) hold with dichotomy, we have to redo the Algorithm 1 for each λ . To avoid searching for the optimal λ and improve the learning efficiency, we can modify $R(n)$ in Algorithm 1 to $R(n) = \sum_k I_k^l(n) - c(\sum_k p_k^g(n) - D_0)^2$, where c is a positive penalty parameter and $c(\sum_k p_k^g(n) - D_0)^2$ is the penalty term. By selecting an appropriate c , we can obtain the solution close to that of the Lagrangian multiplier method.

IV. NUMERICAL RESULTS

In this section, we provide some simulation results to demonstrate the performance of the proposed algorithm. We consider a multi-antenna system with $M_t = 4$ or 6. $\sigma^2 = 1$. The practical channel vector $\mathbf{h}_k(n), \forall k$ is modeled as first order Markov model, i.e., $\mathbf{h}_k(n+1) = \sqrt{\alpha^h} \mathbf{h}_k(n) + \sqrt{1 - \alpha^h} \boldsymbol{\phi}_k^h(n+1)$, where $\alpha^h = 0.9$, $\mathbf{h}_k(n), \boldsymbol{\phi}_k^h(n)$ are independent and $\{\mathbf{h}_k(n), \boldsymbol{\phi}_k^h(n)\} \sim \mathcal{CN}(0, \mathbf{I})$. Considering that the environment usually doesn't change abruptly, the energy harvesting process is also modeled as first order Markov model. That is, $G(n+1) = \sqrt{1 - \alpha^G} G(n) + \sqrt{\alpha^G} \phi^G(n)$, where $\alpha^G = 0.9$ and $\{G(n), \phi^G(n)\} \sim \mathcal{N}(3, 1)$. Let $|\mathcal{W}|$ denote the number of beamforming vectors in $\mathcal{W}$. When $M_t = 4$, $|\mathcal{W}| = 32$, and when $M_t = 6$, $|\mathcal{W}| = 64$. The code-words in $\mathcal{W}$ are generated by finding $|\mathcal{W}|$ central points from sufficient channel samples with K-means. The battery capacity $B_{\max} = 10$, and $p_p = 4$. In PPO algorithm, we use the 2-layer Rectified Linear Unit (ReLU) Multi-Layer Perceptron (MLP) as the approximator for policy π_θ and value function V_ω . In simulations, we find that 64 neurons per layer is enough for the proposed scheme, while for the scheme without dimension reduction, 256 neurons per layer is needed. The Adam optimizer with learning rate 4×10^{-4} is adopted. The clip range $\varepsilon = 0.2$, and the discount factor $\beta = 0.99$.

As stated in Section III, for (P1), we can directly apply PPO to find the mapping between high dimensional state $s(n)$ and action $\{p_k^b(n), p_k^g(n), \mathbf{w}_k(n)\}_{k=1}^K$ ('high' in Fig. 1), or first determine $\{\mathbf{w}_k(n)\}_{k=1}^K$ using (12) and then get the low dimension action $\{p_k^b(n), p_k^g(n)\}_{k=1}^K$ by using PPO ('low' in Fig. 1). In Fig. 1, we compare the steady-state performances of schemes mentioned above with $\rho = 1 \times 10^{-3}$, $K = 2$, $M_t \in \{4, 6\}$. We can see that the performance of low dimensional scheme is significantly superior to that of high dimensional scheme. Although jointly optimize the beamforming vectors and power allocation using the high dimensional scheme is theoretically optimal, it is hard to attain a good performance

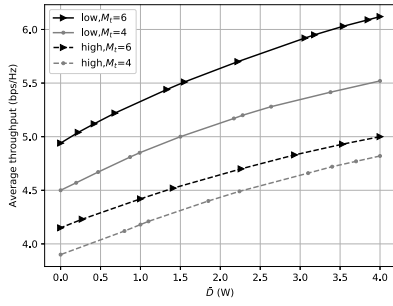
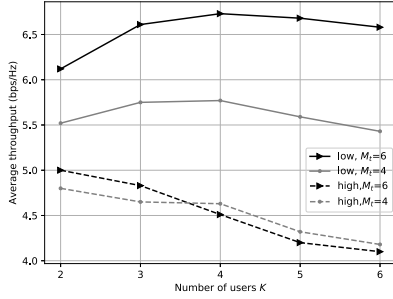
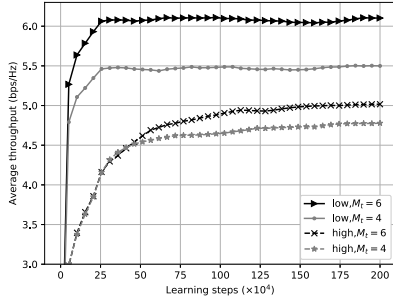
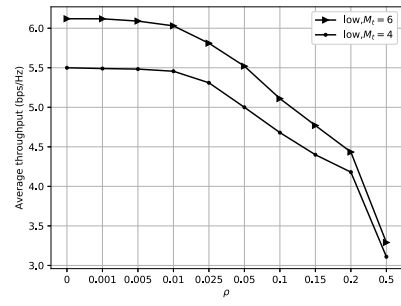
Fig. 1. Average throughput vs. consumption of grid energy with different M_t .Fig. 2. Average throughput vs. number of users with different M_t .

Fig. 3. Learning curves of high and low dimensional schemes.

due to the high dimensional action space. Specially, when $\lambda = 0$, we find that the high dimensional scheme has learned a time division multiple access (TDMA) strategy. That is, allocating all power to the user with the best channel condition, which is typically not the optimal in MIMO multiuser systems.

In Fig. 2, we compare the performances of low and high dimensional schemes with different K when $M_t = 4$ and 6. In the low dimensional scheme, the throughput increases with the increasing of K due to the multiplexing gain. However, the multiuser interference cannot be fully eliminated because of the use of discrete codebook. Hence, the throughput starts to decline when $K > 4$ due to the multiuser interference. It is observed that the throughput of the high dimensional scheme is lower than that of low dimension scheme and decreases as K increases. Besides, when $K > 3$, the high dimension scheme with $M_t = 6$ even performs worse than that with $M_t = 4$. This is because the action space increases dramatically with the increasing of K or M_t , which makes the learning very difficult. Fig. 2 verifies that the proposed scheme can significantly ease the learning difficulty so that the performance is improved.

Fig. 3 shows the learning curves of high and low dimensional schemes for $M_t = 4$ or 6 and $K = 2$. For a fair comparison, we consider the case with $\lambda = 0$ in which $\bar{D} = 4$. From Fig. 3, we can see that PPO in low dimension scheme takes only 2.5×10^5 steps to reach the steady state. While in

Fig. 4. Effect of ρ on throughput.

high dimensional scheme, it takes more than 1.5 million steps, which is 7 times of that in the low dimension one. Besides, because 256 neurons per layer is used in the high dimensional scheme, it spends more time on each learning step. In our computer configuration, when $M_t = 4$, each learning step takes 1.2ms in the high dimensional scheme while 0.7ms in the low dimensional scheme.

The curves in Fig. 4 demonstrate the effect of estimation error factor ρ on the throughput performance when $M_t = 4$ or 6. We set $D_0 = 4$, $K = 2$. When $\rho < 10^{-2}$, the estimation error has negligible effect on the throughput and when $\rho > 10^{-2}$, the throughput will deteriorate with the increasing of ρ .

V. CONCLUSION

In this letter, we investigate that in the multiuser multi-antenna system, how to maximize the long-term sum throughput under the constraint of energy supply when the transmitter is powered with hybrid energy and only has imperfect CSI. The throughput maximization problem is solved by a dimension reduction deep RL method, in which beamforming vectors are firstly determined based on estimated CSI, and then PPO is applied to find the optimal mapping between simplified system state and transmit powers. Simulation results demonstrate the advantage of proposed method over traditional method, and the effect of estimation error on the throughput performance.

REFERENCES

- [1] Y. Jing and S. Ulukus, "Optimal packet scheduling in an energy harvesting communication system," *IEEE Trans. Commun.*, vol. 60, no. 1, pp. 220–230, Jan. 2012.
- [2] C. K. Ho and R. Zhang, "Optimal energy allocation for wireless communications with energy harvesting constraints," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4808–4818, Sep. 2011.
- [3] F. Yuan, K. Q. T. Zhang, S. Jin, and H. Zhu, "A harvest-use-store mode for energy harvesting communication systems with optimal power policy," in *Proc. IEEE Int. Conf. Commun.*, 2014, pp. 5366–5371.
- [4] C. Huang, R. Zhang, and S. Cui, "Optimal power allocation for outage probability minimization in fading channels with energy harvesting constraints," *IEEE Trans. Wireless Commun.*, vol. 13, no. 2, pp. 1074–1087, Feb. 2014.
- [5] J. Gong, Z. Zhou, and S. Zhou, "Analysis and optimization of wireless transmissions over fast fading channels with slow time-varying energy arrival," in *Proc. IEEE Int. Conf. Commun.*, 2017, pp. 1–6.
- [6] A. Aprem, C. R. Murthy, and N. B. Mehta, "Transmit power control policies for energy harvesting sensors with retransmissions," *IEEE J. Sel. Topics Signal Process.*, vol. 7, no. 5, pp. 895–906, Oct. 2013.
- [7] M. R. Zennaidi, Z. Rezki, and M. S. Alouini, "Performance limits of online energy harvesting communications with noisy channel state information at the transmitter," *IEEE Access*, vol. 5, pp. 1239–1249, 2017.
- [8] P. Blasco, D. Gunduz, and M. Dohler, "A learning theoretic approach to energy harvesting communication system optimization," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1872–1882, Apr. 2013.
- [9] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017. [Online]. Available: arXiv:1707.06347.