**Multi-Agent Deep Reinforcement Learning for Distributed Resource Management in Wirelessly Powered Communication Networks**

<https://arxiv.org/pdf/2010.09171.pdf>

|  |
| --- |
| **List of Symbols** |
|  |

|  |
| --- |
| **The model accomplishes…** |
| * **Agent 간의 simple interaction**을 위한 목적의 **분산 학습** (distributed training) * **지역적으로만 관측 가능**한 정보에 대한 **distributed execution** * 전통적인 centralized 알고리즘에 비해 **computational complexity를 줄일** 수 있음 |

|  |
| --- |
| **Figure : System Model** |
|  |

|  |
| --- |
| **System Model** |
| WPCN process는 **Figure 1**과 같이 time-slot 방법으로 이루어진다.   * **HAP i**는 duration 동안 **downlink WET**을 수행한다. * **User i**의 **WIT**는 남은 시간인 동안 **uplink**에서 수행된다.   따라서 **N개의 cell**이 있다면 전체 system block은 **N+1개의 interval**로 구성된다.   * 따라서 HAP들의 ordering 는 으로 정의된다. * 이때 interval n의 duration은 이 된다. **()**   Interval n에서…   |  |  |  |  |  | | --- | --- | --- | --- | --- | | 인 **HAP i**는 사용자로부터 전송된 **WIT 신호를 수신**한다.  인 나머지 HAP들은 **downlink를 통해 에너지를 전송**한다.   * 여기서 **=1이면 HAP i가 WIT 신호를 수신**했다는 것이고, 0이면 그렇지 않다는 것이다.  |  |  | | --- | --- | |  | HAP i는 다른 사용자 j로부터 **WET interference**를 경험함 | |  | HAP i는 다른 사용자 j로부터 **WIT interference**를 경험함 |   여기서 사용자 j로부터 HAP i로의 **instantaneous WIT interference** 와 **cross-link WET interference** 는 다음과 같다.  user i의 uplink data transmission은 **일 때 carry out**되고, 이에 대응되는 **achievable data rate** 는 다음과 같이 쓸 수 있다.  여기서 는 noise power와 같다. |   따라서 **user i의 N+1 interval 동안의 total achievable rate**는 다음과 같다.  여기서 는 주어진 input power x에 대한 **EH circuit의 input-output relationship**을 나타낸다. 여기서 **실제 EH circuit에서의 non-linearity**는 다음과 같이 모델링된다.  따라서 **user i의 모든 interval에서의 total harvested energy**는 다음과 같이 정의된다. |

|  |
| --- |
| **System Model (Cont.)** |
| 여기서는 **time allocation 와 uplink power allocation joint하게 최적화**하여 sum-rate performance를 최대화한다. 따라서 **문제는 다음과 같이 formulate**된다. |

|  |
| --- |
| **Proposed Model Approach : MADAL** |
| Agent i는 **actor DNN과 critic DNN**의 2개의 DNN을 가지며, 이들 각각은 파라미터로 **와** 를 갖는다.  **<Actor DNN>**   |  | | --- | | 주어진 **state 에 대해 action 의 조건부 확률**을 계산한다. 이때 의 stochastic probability 정책을 이용한다. 이때 action 는 로 정의된다. |   **<critic DNN>**   |  | | --- | | Value function 를 **policy 에 대해 모델링**한다.   * 이것은 actor DNN의 최적화를 위해 사용된다. 즉 critic DNN을 feed-forward 구조의 standard fully-connected DNN으로 구성한다. |   **<Action>**  한정된 action set을 만들기 위해서, 여기서는 **연속 변수 와** 를 이용한다. 즉 **time allocation과 power allocation에 대한 action space**를 agent i에 대해서 다음과 같이 각각 정의한다. 따라서 전체 action space의 크기는 이다.  여기서 는 가 되는 문제를 방지하기 위해 사용된다.  Actor DNN의 출력은 **,** 의 element 값에 대한 **probability mass function의 값을 반영**한다.  **<State>**  Time slot t에서 agent i의 state 는 다른 모든 cell에서 탐지된 **internal observation 와 external information 들을 concatenate**하여 구성된다. 따라서 internal information 는 의 집합으로 구성된다. 반면 external information 는 의 집합으로 구성된다. |

|  |
| --- |
| **Proposed Model Approach : MADAL (Cont.)** |
| **<Reward>**  Agent **i**의 time slot **t+1**에서의 reward 는 다음과 같이 나타낼 수 있다.  여기서 는 다음과 같이 나타낸다.  **<Loss function>**  여기서는 MA-A2C structure를 이용하여 학습을 하므로, **loss function**은 다음과 같다.  학습 알고리즘은 다음과 같다. |

|  |
| --- |
| **Simulation Result** |
|  |

|  |
| --- |
| **Simulation Result (Cont.)** |
|  |