Optimization-driven Hierarchical Deep Reinforcement Learning for Hybrid Relaying Communications

Yuze Zou¹, Yutong Xie², Canhui Zhang³, Shimin Gong³, Hoang Thai Dinh⁴, and Dusit Niyato⁵
Huazhong University of Science and Technology, China
Shenzhen Institutes of Advanced Technology, China
Sun Yat-sen University, China
University of Technology Sydney, Australia
Nanyang Technological University, Singapore
Outline

• Introduction
• System Model
• Problem Formulation
• Numerical Results
• Conclusions
Introduction

Internet of Things (IoT) is an emerging paradigm that provides the future network of interconnected devices

e.g., smart phones
wearable devices
wireless sensors
consumer devices.

Technical challenges:
✓ ENERGY SUPPLY to billions of IoT devices
✓ SPECTRUM DEMAND for information transmission
Introduction

- Extreme low power consumption, i.e., < 100 uW
- Low data rate/ high delay /vulnerability to channel

- High power consumption, i.e., >10 mW
- High data rate (> 1Mbps), reliability via power control
A. Hybrid Relaying Communications

This is an simplified example with just two relays

- Active RF signals
- Signal backscattering

Complex channel vectors

Beamforming information of HAP
A. Hybrid Relaying Communications
The beamforming information can be received by both the active relay-1 and the target receiver directly.

The passive relay-n can enhance channel $f_0$ and $f_1$ through backscattering.
A. Hybrid Relaying Communications

**System Model**

The active relay-1 amplifies and forwards the received signal to the receiver.

The HAP also beamforms the same information symbol to the receiver.

The passive relay-n can enhance the forward channel $g_1$ from the active relay-1 to the receiver.

- The active relays’ forwarding.
A. Hybrid Relaying Communications

Due to the passive relays’ backscattering, the enhanced channels can be represented as follows:

\[ \hat{f}_0 = f_0 + \sum_{k \in \mathcal{N}} b_k g_k \Gamma_k f_k, \]
\[ \hat{f}_n = f_n + \sum_{k \in \mathcal{N}} b_k z_{kn} \Gamma_k f_k, \forall n \in \mathcal{N}. \]

\( \hat{f}_0 \) and \( \hat{f}_n \) denote the equivalent channels from the HAP to receiver and to the active relay-\( n \).
System Model

B. Signal Model in Two Hops

Received signal at the relay-n

\[ r_n = \sqrt{1 - \rho_n} p_t \hat{f}_n^H w_1 s + \sigma_n \]

Received signal at the receiver

\[ r_d = \sum_{n=1}^{N} \hat{g}_n x_n r_n + \sqrt{p_t} \hat{f}_0^H w_2 s + v_d \]
B. Signal Model in Two Hops

**SNR in the first hop**

\[ \gamma_1 = p_t |\hat{f}_0^H w_1|^2 \]

**SNR in the second hop**

\[ \gamma_2 = \frac{\left| \sum_{n \in \mathcal{N}} x_n y_n \hat{g}_n + \sqrt{p_t} \hat{f}_0^H w_2 \right|^2}{1 + \sum_{n \in \mathcal{N}} |x_n \hat{g}_n|^2} \]
Problem Formulation

Optimization Explanation:

To maximize the overall throughput $\gamma = \gamma_1 + \gamma_2$ in two hops, we aim to optimize the HAP’s beamforming strategies $(w_1, w_2)$, as well as the relays’ radio mode selection $b_n$ and operating parameters:

$$\max_{w_1, w_2, b_n, \rho_n, \theta_n} \gamma_1 + \gamma_2$$

subject to:

- $\|w_1\| \leq 1$ and $\|w_2\| \leq 1$,
- $p_n \leq \eta \rho_n p_t |\hat{f}_n^H w_1|^2$, $\forall n \in \mathcal{N}_a$,
- $\rho_n \in (0, 1)$, $\forall n \in \mathcal{N}_a$,
- $b_n \in \{0, 1\}$, $\forall n \in \mathcal{N}$,
- $\theta_n \in [0, 2\pi]$, $\forall n \in \mathcal{N}_b$. 

- Overall throughput
- HAP’s beamforming strategies
- Transmit power of the n-th active relay
- Active relays’ operating parameter
- Relays’ radio mode selection
- Active relays’ operating parameter
The optimization-driven H-DDPG framework for hybrid relaying communications

- Combining DQN and DDPG in one hierarchical framework
- Better-informed estimate of target value $y_t$ with model-based optimization
Problem Formulation

Deep Q network (DQN) algorithm structure

DQN loss function

Environment

Current Q network

Target Q network

Playback memory unit

$Q(s, a; \theta)$

$\max_a Q(s', a'; \theta^-)$

$\text{arg max}_a Q(s, a; \theta)$

Gradient of loss function

Copy parameters every N steps

$s$

$(s, a)$

$s'$

$r$

$(s, a, r, s')$
We use DQN algorithm to select the relay mode at the outer-loop.
Problem Formulation

Deep Deterministic Policy Gradient (DDPG) algorithm structure
Problem Formulation

Deep Deterministic Policy Gradient (DDPG) algorithm

**Algorithm 1 DDPG algorithm**

Randomly initialize critic network $Q(s, a|\theta^Q)$ and actor $\mu(s|\theta^\mu)$ with weights $\theta^Q$ and $\theta^\mu$.

Initialize target network $Q'$ and $\mu'$ with weights $\theta'^Q$, $\theta'^\mu$.

Initialize replay buffer $R$.

for episode = 1, M do

  Initialize a random process $\mathcal{N}$ for action exploration.

  Receive initial observation state $s_1$.

  for $t = 1, T$ do

    Select action $a_t = \mu(s_t; \theta^\mu) + \mathcal{N}_t$ according to the current policy and exploration noise.

    Execute action $a_t$ and observe reward $r_t$ and observe new state $s_{t+1}$.

    Store transition $(s_t, a_t, r_t, s_{t+1})$ in $R$.

    Sample a random minibatch of $N$ transitions $(s_i, a_i, r_i, s_{i+1})$ from $R$.

    Set $y_i = r_i + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta'^\mu)|\theta'^Q')$.

    Update critic by minimizing the loss: $L = \frac{1}{N} \sum_i (y_i - Q(s_i, a_i|\theta^Q))^2$.

    Update the actor policy using the sampled policy gradient:

    $$\nabla_{\theta^\mu} J \approx \frac{1}{N} \sum_i \nabla_a Q(s, a|\theta^Q)|_{s=s_i, a=\mu(s_i)} \nabla_{\theta^\mu} \mu(s|\theta^\mu)|_{s_i}$$

    Update the target networks:

    $$\theta'^Q \leftarrow \tau \theta^Q + (1 - \tau) \theta'^Q$$

    $$\theta'^\mu \leftarrow \tau \theta^\mu + (1 - \tau) \theta'^\mu$$

  end for

end for

We use DDPG algorithm to optimize the continuous beamforming and relays’ operating parameters.
Problem Formulation

Search lower bound:

**Proposition 1:** Given the radio mode of each relay \( n \in \mathcal{N} \), a feasible lower bound on (5) can be found by the convex reformulation as follows:

\[
\begin{align*}
\max_{\mathbf{w}_1, \mathbf{w}_1 \succeq 0} & \quad p_t |\mathbf{f}_0|^2 + p_t |\mathbf{f}_0^H \mathbf{w}_1|^2 + p_t \sum_{n \in \mathcal{N}_a} s_{n,1} \\
\text{s.t.} & \quad \left[ \kappa_n v_n - \frac{(1 + v_n)s_{n,1}}{\sqrt{p_t s_{n,1}}} \right] \geq 0, \quad \forall n \in \mathcal{N}_a \tag{8a} \\
& \quad \kappa_n \leq \hat{\mathbf{f}}_n^H \mathbf{W}_1 \hat{\mathbf{f}}_n, \quad \forall n \in \mathcal{N}_a \tag{8b} \\
& \quad s_{n,1} = \hat{\mathbf{f}}_n^H \mathbf{W}_1 \mathbf{f}_n - \hat{\mathbf{f}}_n^H \mathbf{W}_1 \hat{\mathbf{f}}_n, \quad \forall n \in \mathcal{N}_a, \tag{8c}
\end{align*}
\]

where \( v_n \triangleq \eta p_t |\hat{\mathbf{g}}_n|^2 |\mathbf{f}_0|^2 \) is a constant. At optimum, the power-splitting ratio is given by \( \rho_n = \frac{\hat{\mathbf{f}}_n^H \mathbf{W}_1 \hat{\mathbf{f}}_n}{\hat{\mathbf{f}}_n^H \mathbf{W}_1 \mathbf{f}_n} \) for \( n \in \mathcal{N}_a \).
Performance comparison of different algorithms with different value of hyper parameter:

- Optimization-driven H-DDPG achieves the highest convergence rate.
- In either case, H-DDPG framework outperforms the conventional DDPG in terms of a higher learning rate, due to the reduced action space.
- Optimization-driven H-DDPG is more robust to different values of the hyper parameter \( \gamma \), which is a very significant advantage.
Numerical Results

Strategy update of the inner-loop DDPG and its dynamics in different DQN episodes:

- Each DQN episode spans over 4000 episodes of DDPG strategy updates to ensure the convergence of the inner-loop DDPG algorithm.
- Within each part, the inner-loop DDPG algorithm can converge to a stable reward value with a fixed radio mode selection, which is generated by the out-loop DQN episode.
- The Optimization-driven H-DDPG has a faster learning rate than the Model-free H-DDPG in the inner loop.
Performance improvement with the increases in the number of relays:

- The convergent reward increases with more relays assisting the information transmission.
- The learning rate becomes slightly reduced with more relays, because more relays provide additional degree of freedom for the HAP to leverage higher diversity for its information transmissions, while at the cost of a lower convergence rate due to increased action space.
Conclusions

Optimization-driven Hierarchical Deep Reinforcement Learning for Hybrid Relaying Communications:

• We proposed a novel optimization-driven hierarchical deep reinforcement learning approach to solve the throughput maximization problem.

• We integrated Deep Q-network and model-based optimization technique into the conventional DDPG algorithm in a hierarchical structure.

• We also proposed a model-based optimization to give a guidance for the target estimation within the learning process, especially in the early stage.

• Simulation results reveal that the proposed algorithm outperforms the conventional DDPG algorithm in terms of robustness to the hyper parameters and higher convergence rate.
Questions & Answers

Thank you!

Contact: gong0012@e.ntu.edu.sg