Deep Reinforcement Learning Based Mobility Load Balancing Under Multiple Behavior Policies

Yue Xu[1,2,3], Wenjun Xu[1], Zhi Wang[3], Jiaru Lin[1], Shuguang Cui[3,2]

[1]Beijing University of Posts and Telecommunications

[2]SRIBD and The Chinese University of Hong Kong, Shenzhen

[3]University of California, Davis

xuy@bupt.edu.cn

May 22nd, 2019

Outline

1. Background
2. System Model
3. Experiment Results
4. Scalability

Outline

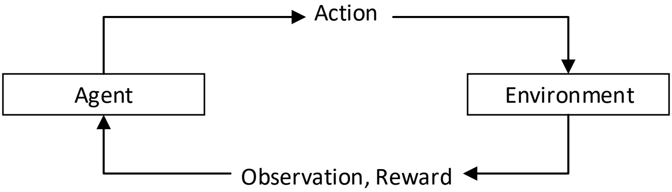
1 Background

2 System Model

3 Experiment Results

4 Scalability

What is reinforcement learning



- Reinforcement learning: aims at maximizing a **cumulative reward** by selecting **a sequence of optimal actions** to **interact with** a stochastic **unknown environment**, where the dynamics is usually modeled as a Markov decision process (MDP)

[1] Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. Cambridge: MIT press, 2018.

Main components of RL

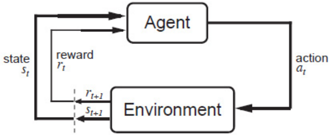
- State & Action: RL input & output
- Reward: immediate feedback, indicates the good and bad events
- Value function : the accumulated reward over the future

$$V^{\pi}(\boldsymbol{s}) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} r(\boldsymbol{s}_t, \boldsymbol{a}_t) \middle| \boldsymbol{s}_t = \boldsymbol{s}\right],$$

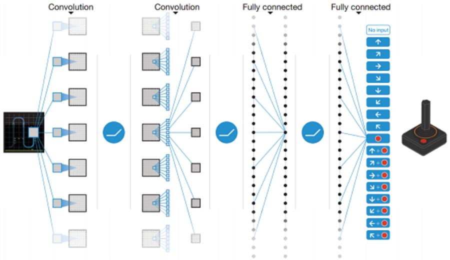
- Policy function: behavior of the agent, the mapping between state and action

$$\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$$

What is deep reinforcement learning?



AlphaGo



Using deep neural networks to approximate

- Value function: model-free, e.g., deep Q-network and its variants
- Policy function: model-free, e.g., DDPG, A3C, PPO

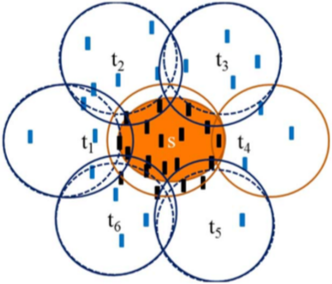
Why using (deep) RL to solve MLB?

- Challenges: Traditional rule-based or model-based mobility load balancing (MLB) models can hardly adapt to complicated and changing wireless networks, e.g.,
  - random network topology
  - scalability to large-scale networks
- Contributions:
  - We propose a RL-based MLB **model**
    - autonomous learning, good adaptability to unknown environments
    - more far-sighted optimization goal
  - We proposed an off-policy DRL-based **algorithm** for MLB
    - learning under multiple behavior policies
    - asynchronous parallel learning framework

Outline

1. Background
2. System Model
3. Experiment Results
4. Scalability

Mobility Load Balancing



- Mobility load balancing (MLB): load balancing in self-organizing networks (SONs), controls logical cell boundaries by tuning the cell individual offset (CIO) to control user handovers

Handover in SONs

- Handovers between a serving cell i and a target cell j are triggered according to the A3 condition by 3GPP:

$$F_j - F_i > O_{i,j} + Hys$$

  - F_i, F_j: users reference signal received power (RSRP)
  - $O_{i,j}$: the HO difference between cell i and j
  - Hys: the Handover Hysteresis (Hys), prevent frequent handovers
- Our aim is to optimally adjust the O_{ij} of each cell pair so as to balance the load distribution

Learning Context of RL

- Cell load is defined as the ratio of users' required physical resource blocks (PRBs) versus the available PRBs
- State: i) the load derived from the averaged load $\tilde{\rho}_i = \rho_i - \rho_g$, where $\rho_g = \frac{1}{N}\sum_{i=1}^{N}\rho_i$ with N the number of SBS; ii) the fraction of the edge users E_i.

$$\boldsymbol{s}_t = \left[\tilde{\rho}_1^t, \tilde{\rho}_2^t, \cdots, \tilde{\rho}_N^t, E_1^t, E_2^t, \cdots, E_N^t\right]^\top.$$

- Action: the CIO value of each cell pair, i.e.,

$$\boldsymbol{a}_t = \{O_{ij}(t) | \forall i, j \in \mathcal{I}\}$$

- Reward: the inverse of the maximum cell load of cell set $\mathcal{I}$

$$r(\boldsymbol{s}_t, \boldsymbol{a}_t) = \frac{1}{\max_{i\in\mathcal{I}} \rho_i(t)},$$

The mathematical problem can be formulated as

$$
\begin{aligned}
\mathcal{P}_0: \quad & \max_{\mu} J(\mu) \\
\text{s.t.} \quad & C_1: \mathcal{X}_{u,i} \in \{0,1\}, \sum_{i \in \mathcal{I}} \mathcal{X}_{u,i} \leq 1, \forall u \in \mathcal{U} \\
& C_2: O_{ij} \in [O_{\min}, O_{\max}], \forall i,j \in \mathcal{I}
\end{aligned}
$$

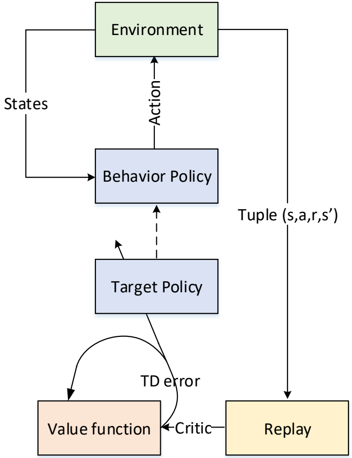
where

$$
J(\mu) = \mathbb{E}(r_0^{\gamma} | \mu), \quad r_t^{\gamma} = \sum_{k=t}^{\infty} \gamma^{k-t} r(s_k, a_k)
$$

- Remark: the RL-based formulation aims at achieving a more far-sighted balanced load distribution

Learning Under Multiple Behavior Policies

- Off-policy RL: the *target policy* is optimized by using the samples generated by following the *behavior policy*
- Multiple behavior policies: improved stability and efficiency, expert guided learning



Problem Formulation Under Multiple Behavior Policies

The mathematical problem can be formulated as

$$\begin{aligned}
\mathcal{P}_0: \quad & \max_{\mu} J(\pi_\theta) = \sum_{m \in \mathcal{M}} J_{\beta_m}(\pi_\theta). \\
\text{s.t.} \quad & C_1: \mathcal{X}_{u,i} \in \{0,1\}, \sum_{i \in \mathcal{I}} \mathcal{X}_{u,i} \leq 1, \forall u \in \mathcal{U}. \\
& C_2: O_{ij} \in [O_{\min}, O_{\max}], \forall i,j \in \mathcal{I}.
\end{aligned}$$

where

$$J_\beta(\pi_\theta) = \mathbb{E}_{s \sim \kappa^\beta}\left[\sum_{k=0}^{\infty} \gamma^k r\left(s, \pi_\theta(s)\right)\right],$$

- Remark: The objective can be viewed as optimizing the value function of the target policy averaged over the state distribution of the behavior policy

Parallel Learning Framework

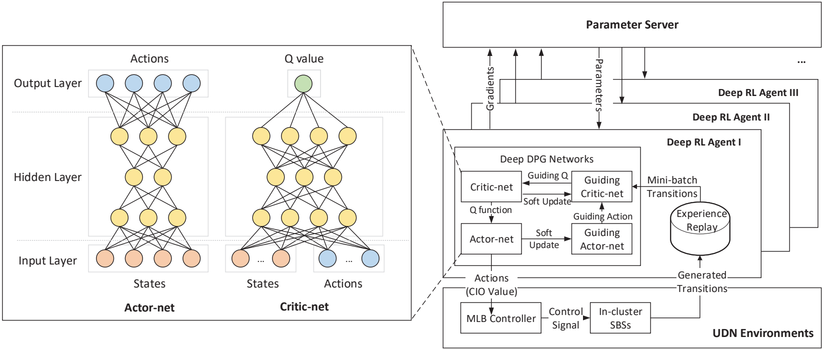


Figure: Overview of the DRL architecture for MLB

Outline

1. Background
2. System Model
3. Experiment Results
4. Scalability

Simulation Setup

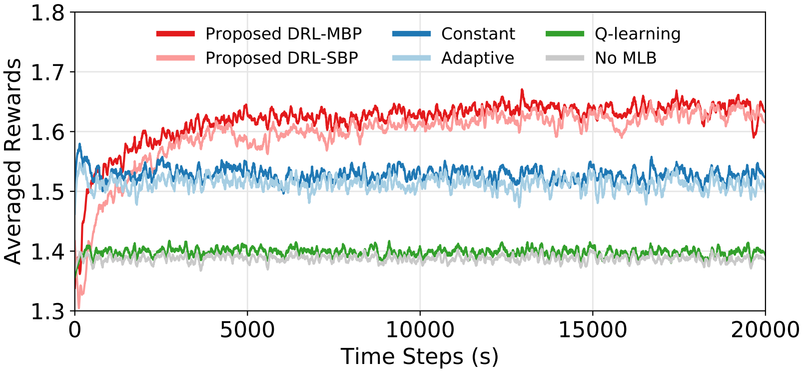
- Environment:
  - six cells randomly distributed in a 1km^2 area
  - 200 users randomly walking at 1m/s
  - constant bit rate (CBR) traffic demand ($32 \sim 80 kbps$)
- Comparing Schemes:
  - rule-based control: constant step size and adaptive step size
  - learning-based control: Q-learning
  - proposed DRL-based control: single behavior policy and multiple behavior policies
- The performances are averaged over 50 different randomized cell topologies to give a fair comparison

[1] Kwan, Raymond, et al. "On mobility load balancing for LTE systems." Vehicular Technology Conference Fall (VTC 2010-Fall), IEEE, 2010.

[2] Yang, Ying, et al. "A high-efficient algorithm of mobile load balancing in LTE system." Vehicular Technology Conference (VTC Fall), IEEE, 2012.

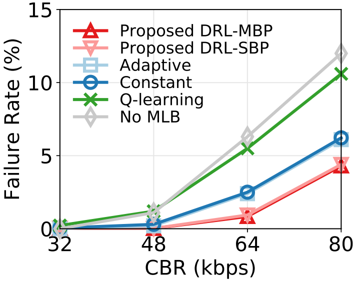
[3] Mwanje, Stephen S., Lars Christoph Schmelz, and Andreas Mitschele-Thiel. "Cognitive Cellular Networks: A Q-Learning Framework for Self-Organizing Networks." IEEE Transactions on Network and Service Management 13.1 (2016): 85-98.

Experiment Result: Rewards

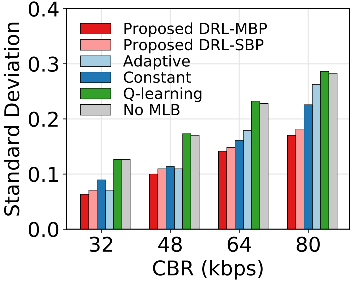


- under a CBR of 80 kbps, averaged every 200 steps
- no-MLB baseline (71%); rule-based control (64%); Q-learning-based control (70%); proposed DRL-based control (60%)

Experiment Result: HFR and LSD



(a) Handover Failure Rate



(b) Load Standard Deviation

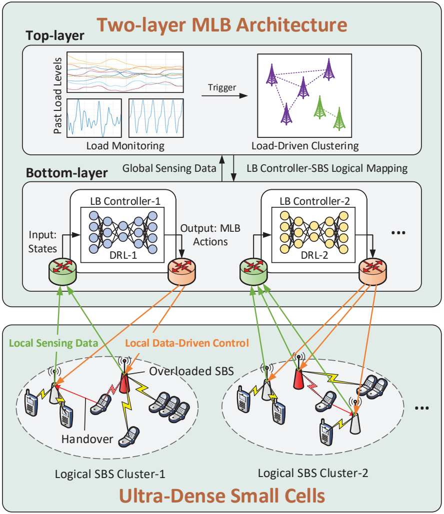
- Handover failure: we block incoming handover attempts to a cell with load exceeding 80% for admission control
- Load standard deviation: an indicator for the load distribution among all the cells

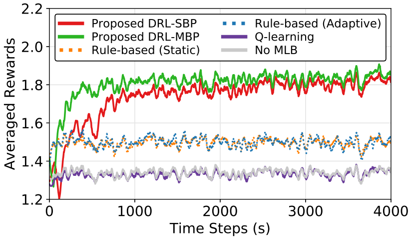
Outline

1. Background
2. System Model
3. Experiment Results
4. Scalability

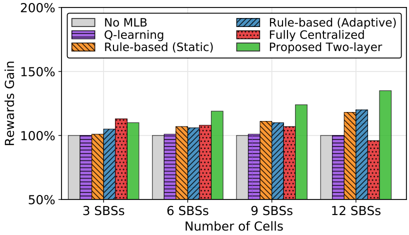
Large-Scale MLB with A Two-Layer Framework

- A two-layer architecture:
  - dynamic load-aware clustering
  - adaptive DRL-based in-cluster MLB
- Main advantages:
  - scalability: self-organized control
  - learning efficiency: break the large non-convex problem into smaller pieces





(c) Performance of learning under multiple behavior policies



(d) Normalized performance gain

Conclusion

- We propose a RL-based MLB **model**
  - autonomous adaptation to unknown environments
  - more far-sighted learning goal
- We proposed an off-policy DRL-based **algorithm** for MLB
  - learning under multiple behavior policies
  - asynchronous parallel learning framework
- The proposed model and algorithm form a general autonomous and intelligent network control framework, which is also promising to solve other large-scale network control problems in the future systems by changing the learning context.