RIT

Machine Learning as Enabler for Cross-Layer Resource Allocation:

Opportunities and Challenges with Deep Reinforcement Learning

Fatemeh Shah-Mohammadi and Andres Kwasinski

Rochester Institute of Technology

Outline

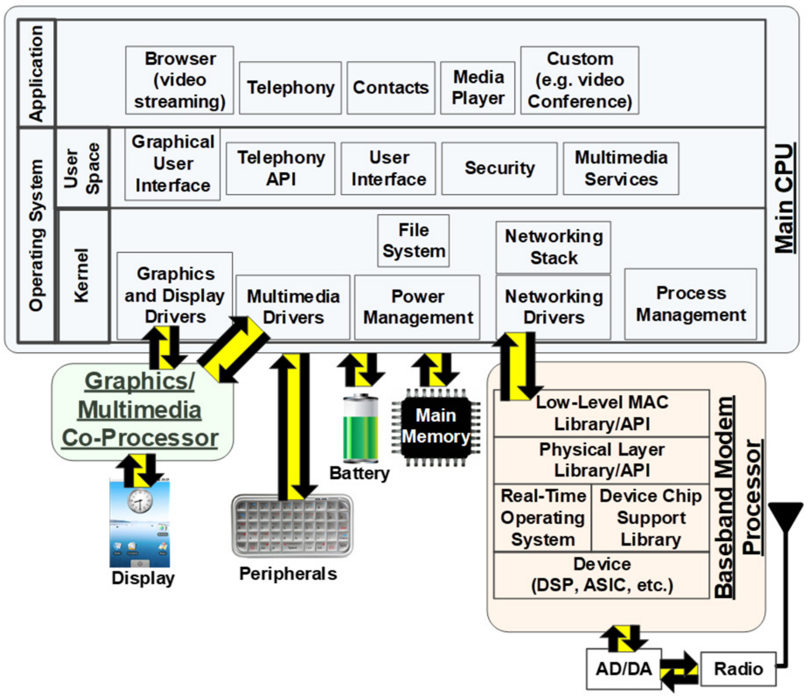
- Benefits for cross-layering.
- Cognitive radios as enablers for cross-layer systems.
- QoE-based resource allocation with Deep Q-learning.
- Transfer learning for accelerated learning of Deep Q-Networks.
- Uncoordinated multi-agent Deep Q-learning with non-stationary environments.

Why Cross-Layer Approach?

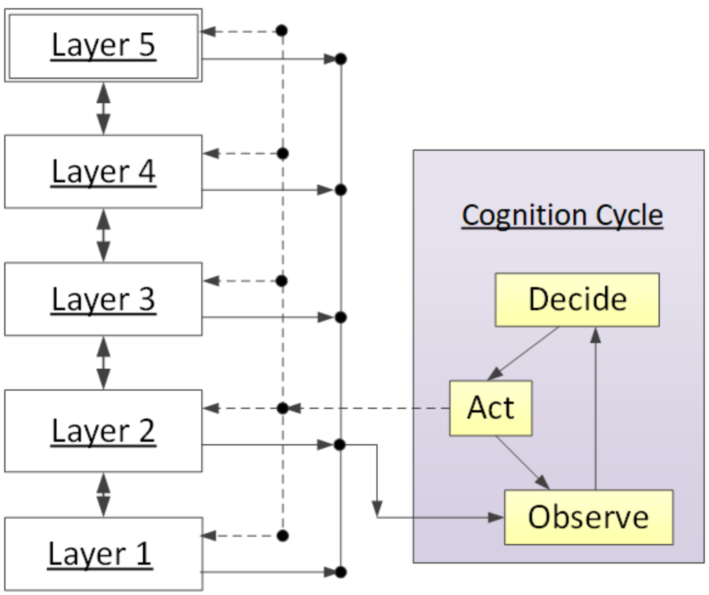
- Ubiquitous computing requires pervasive connectivity,
  - » under different wireless environment,
  - » with heterogeneous network infrastructure and traffic mix.
- User-centric approach translates to QoE metrics:
  - » an end-to-end yardstick.

Obstacle to Cross-Layer Realization

- Wireless devices development is divided into different teams, each specialized in implementing one layer or sub-layer in a specific processor (e.g. main CPU or baseband radio processor).



Cognitive Radios as Cross-Layer Enablers

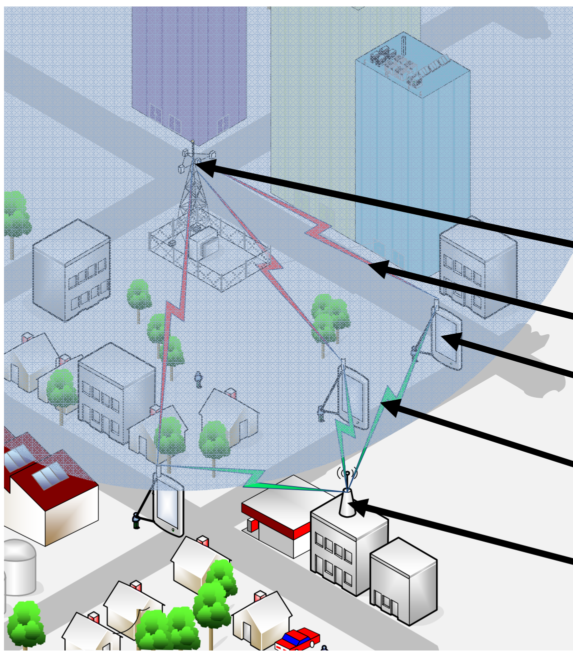


- Wireless network environment as a multi-layer entity.
- Cognitive engine in a cognitive radio senses and interacts with the environment through measuring and acting on the multi-layered environment.

Study Case: Underlay DSA

- A primary network (PN) owns a portion of the spectrum.
- A secondary network (SN) simultaneously transmits over the same portion of the spectrum.
- Transmissions in secondary network are at a power such that the interference they create on the primary network remains below a tolerable threshold.

$$SINR^{(p)} = \frac{G_0^{(p)} P_0}{\sigma^2 + \sum_{j=1}^{N} G_j^{(s)} P_j} \geq \beta_0 \qquad SINR_i^{(s)} = \frac{G_i^{(s)} P_i}{\sigma^2 + G_0^{(s)} P_0 + \sum_{j \neq i} G_j^{(s)} P_j} \geq \beta_i$$



Primary network access point.

Interference from secondary to primary.

Secondary network terminal - SU (cognitive radio).

Transmission in secondary network.

Secondary network access point.

User-Centric Secondary Network

- Heterogeneous traffic mix: interactive video streams (high bandwidth demand, delay constraint) and regular data (FTP).
- Performance: measured as Quality of Experience (QoE)
  - following the user-centric approach to network design and management advocated in 5G systems
- Chosen QoE metric: Mean Opinion Score MOS

Data MOS: $Q_D = a \log_{10}(b\, r_i^{(s)}(1 - p_{e2e}))$

Video MOS: $Q_V = \dfrac{c}{1 + exp(d\;(PSNR - h))}$

A common yardstick

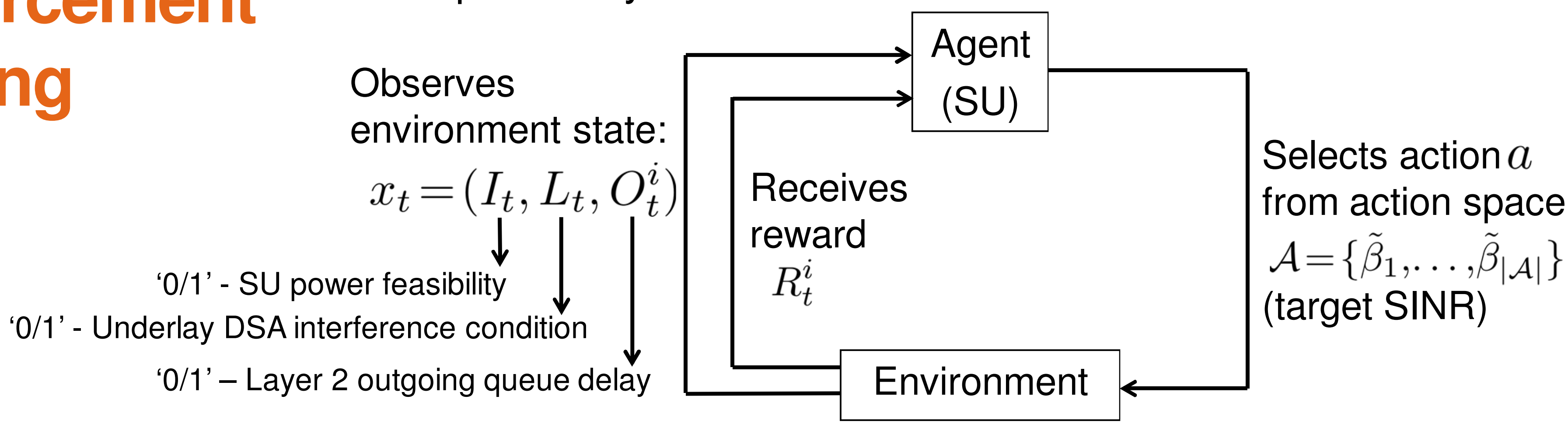
MOS	Quality	Impairment
5	Excellent	Imperceptible
4	Good	Perceptible but not annoying
3	Fair	Slightly annoying
2	Poor	Annoying
1	Bad	Very annoying

Problem Setup

- Cross-layer resource allocation problem.
- For an underlay DSA SN,
- choose:
  - transmitted bit rate (i.e. source compression for video),
  - transmit power,
- such that the QoE for end users is maximized

Solution Based on Deep Reinforcement Learning

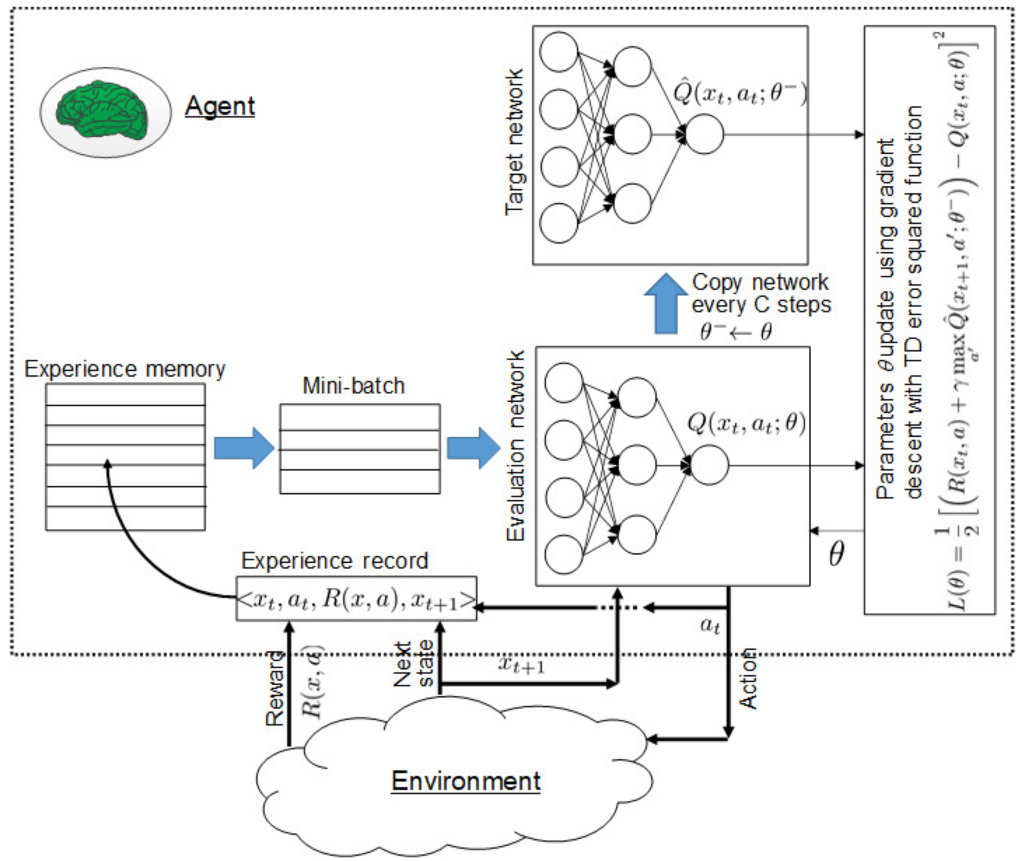
- Use multi-agent Deep Q-Network (DQN) to solve problem.
- An efficient realization of Reinforcement Learning (RL).
- An SU learns the actions (parameters setting) by following a repetitive cycle:



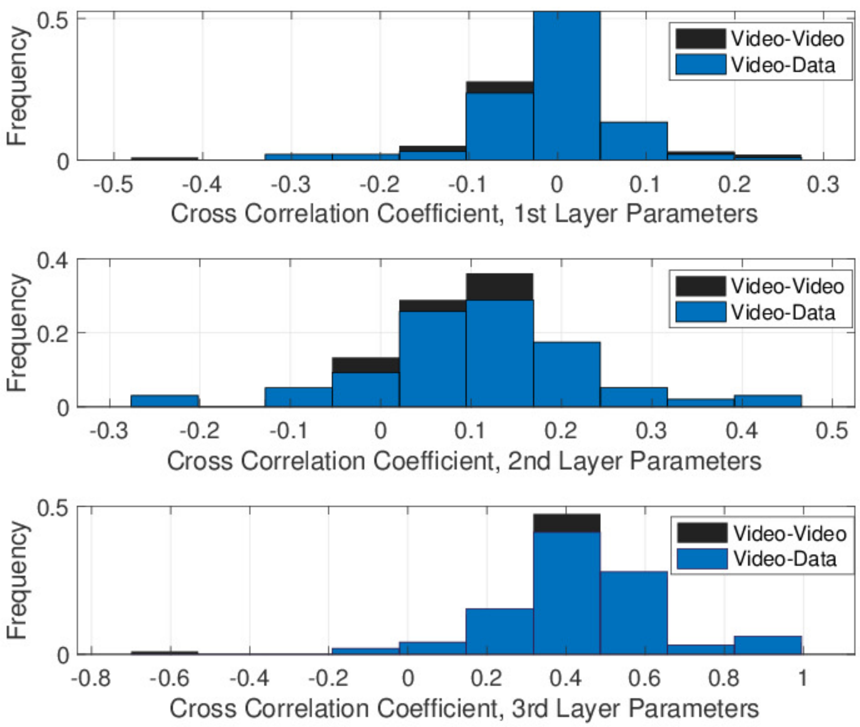
$$R_t^i = \begin{cases} J, & \text{if } I_{t+1} + L_{t+1} + O_{t+1}^i > 0 \\ w_1 \underbrace{r_1^i(a_t, s_t)}_{\text{Layer 2 Delay}} + w_2 \underbrace{r_2^i(a_t, s_t)}_{\text{MOS}}, & \text{otherwise,} \end{cases}$$

Deep Q-Network

- Estimate the Q-action value function – calculation of the expected discounted reward to be received when taking action a when the environment is in state x_t at time t:



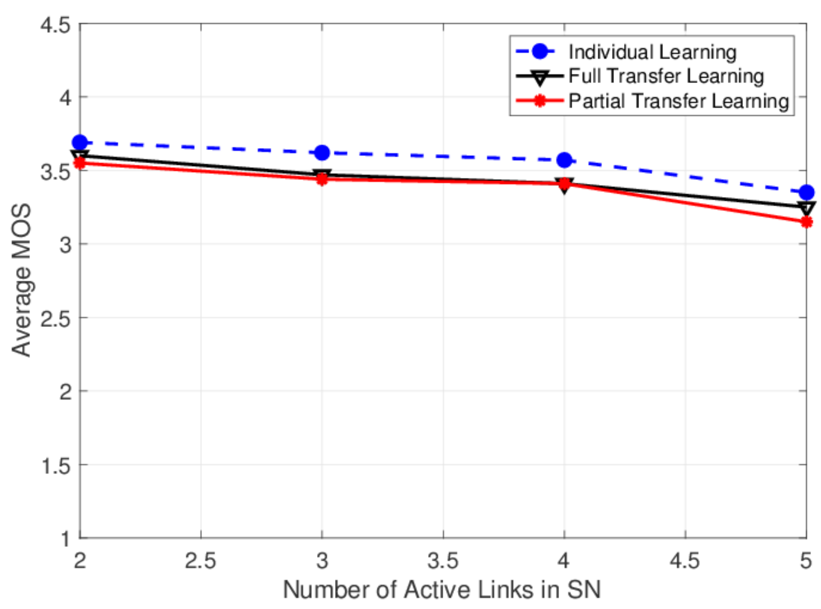
Sharing Experience

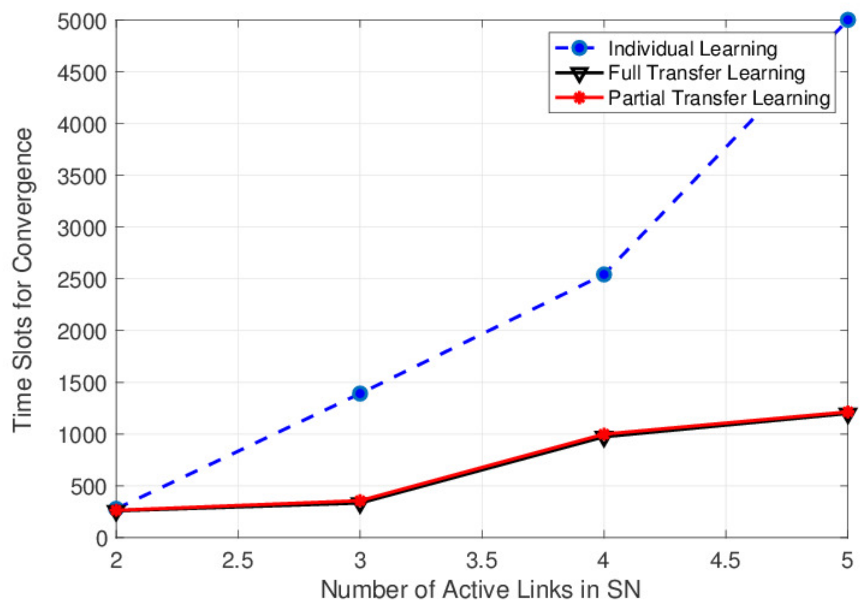


- Limited changes in wireless environment when a newcomer SU joins an already operating network.
- Awareness of the environment (reflected in action-value parameters encoded in DQN weights) of expert SUs can be transferred to the newcomer SU.
- Technique called “Transfer Learning”.

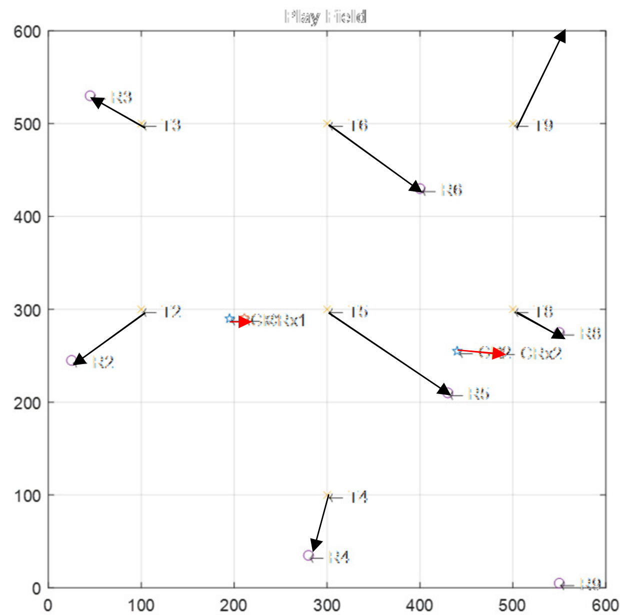
Transfer Learning Results

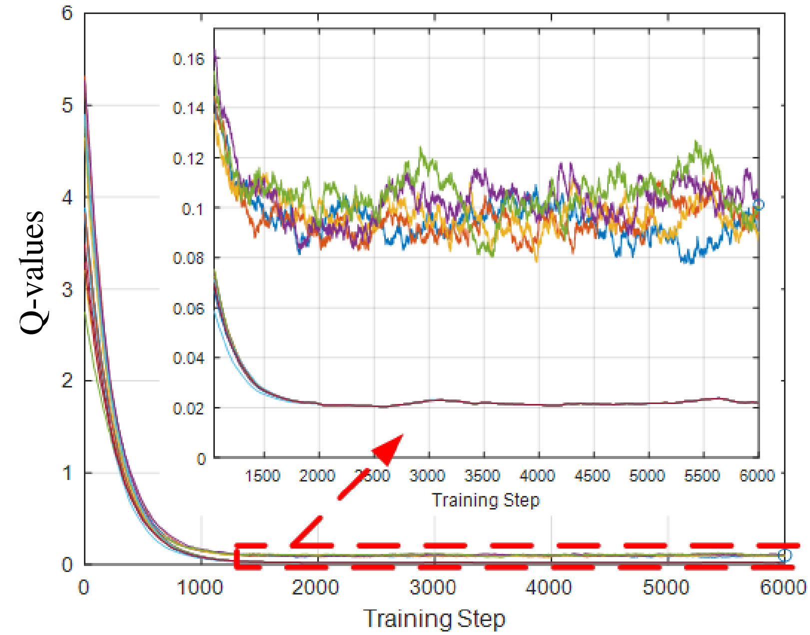
- Accelerated learning without performance penalty.





An Issue With the Standard DQN





- Scenario: Uncoordinated multi-agent power allocation. CRs maximize their throughput while keeping relative throughput change in PN below limit.
- Standard DQN may not converge due to non-stationary environment.

Uncoordinated Multi-Agent DQN

(Acknowledgement to Ankita Tondwalkar)

Algorithm 1 Uncoordinated Distributed Multi-agent DQN

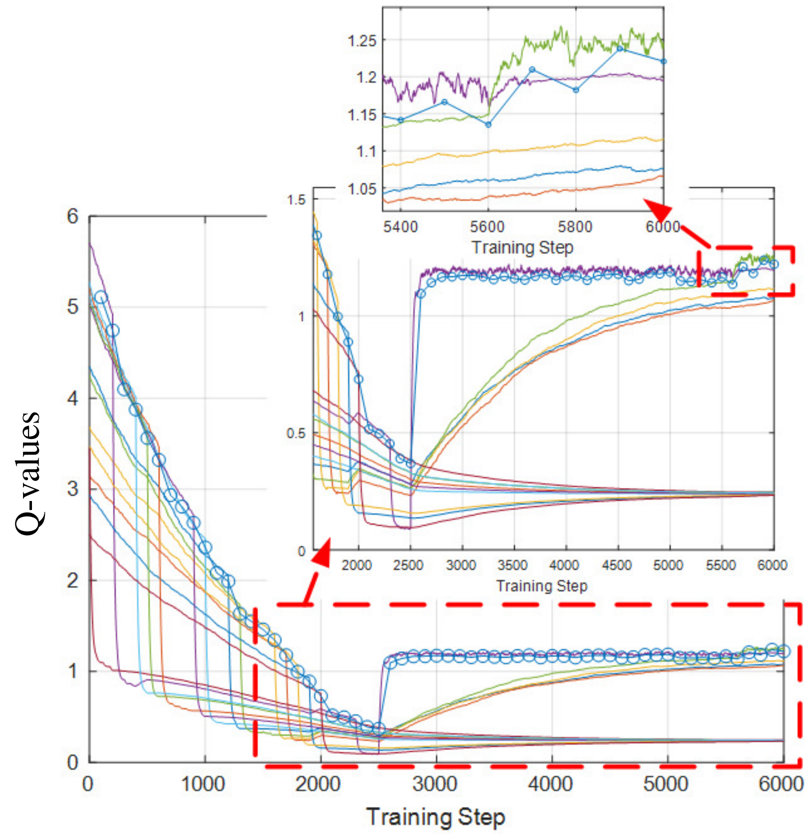
Set parameters
$\rho \in$ (0,1) : Experimentation probability
$\lambda \in$ (0,1) : Inertia
$\gamma \in$ (0,1) : Discount factor
Learning rate $\alpha_n \in= \frac{1}{n^v}$,where $v \in (\frac{1}{2}, 1)$;
Initialize policy $\pi_0 \in \Pi$ (arbitrary)
Sense state x_0
Initialization of the neural network for action-value function Q_i with random weights θ_i
for $0 \le k \le K$ **do**
    for Iterate $t = t_k, t_k + 1, \cdots, t_{k+1} - 1$ **do**
        (kth. exploration phase)
        $a_t = \begin{cases} \pi_k(x_t), & w.p.\ \ 1-\rho \\ \text{any } a \in A, & w.p.\ \ \rho/|A| \end{cases}$
        Receive reward R_t
        Sense state x_{t+1}; n_t = number of visits to (x_t, a_t)
        Update the state $x_{t+1}^{(i)}$ and the reward $R_t^{(i)}$.
        Update parameters (θ) of action-value function $Q(s_t^{(i)}, a_t^{(i)}; \theta_i)$, mini-batch backpropagation
        every c step update array in memory with target action-value function: $\hat{Q}(x, a) \leftarrow Q(x, a; \theta_i), \forall x, a$.
    end for
    $\Pi_{k+1}^i = \{\hat{\pi}^i \in \Pi^i : Q_{t_{k+1}}^i(s, \hat{\pi}^i(x)) \ge \max_{v^i} Q_{t_{k+1}}^i(s, v^i) - \delta^i, \text{for all s}\}$
    $\pi_{k+1}^i = \begin{cases} \pi_k^i, & w.p.\ \ \lambda \\ any\ \pi^i \in \Pi_{k+1}^i, & w.p. \frac{(1-\lambda)}{|\Pi_{k+1}^i|} \end{cases}$
end for

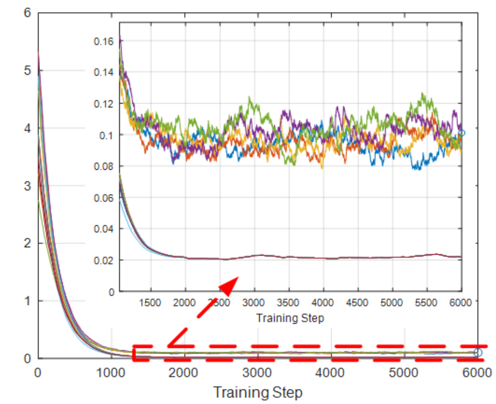
Exploration phase: do action exploration only occasionally – generate near-stationary environment

Near-standard DQN (no replay memory, target action-values stored in array.

Policy update with inertia

Uncoordinated Multi-Agent DQN - Results





Standard DQN (for comparison purposes, same scenario)

- Demonstrable convergence to optimal solution as learning time goes to infinite

Uncoordinated Multi-Agent DQN - Results

Description	Mean Relative Difference	Mean Exp. Phases to Converge	Percent Optimal Policy
Reference setting	0.0249	33.62	69
Standard DQL $c = 1$	0.0945	N/A	56
Standard DQL $c = 60$	0.0744	N/A	55
$c = 1$	0.0331	35.18	66
$c = 60$	0.0096	33.42	72
Mini batch size = 120	0.0375	35.6	67
Mini batch size = 30	0.0241	36.35	69
Uncoordinated per-CR reward	0.0411	32.01	70

- Comparison against optimal solution through exhaustive search – optimality based on maximum sum throughput in SN.

Conclusions

- Discussed the benefits for cross-layered protocols and their practical realization through cognitive radios.
- Presented QoE-based cross-layer resource allocation cognitive engine with Deep Q-learning.
- Explained how learning could be accelerated for a newcomer node by transferring experience from other node.
  - » Learning is accelerated with no discernable performance loss.
- Presented a first-of-its-kind Deep Q-learning technique that converges to optimal resource allocation in uncoordinated interacting multi-agent scenario (non-stationary environment).

Thank You!

Questions?