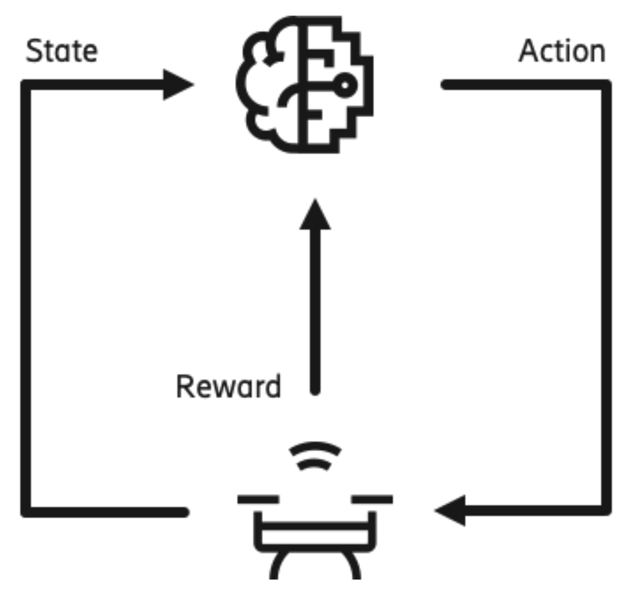
# Multi-Agent Reinforcement Learning with Partial Knowledge over Networks

STEFANOS ANTARIS. AMARU CUBA GYLLENSTEN, MARTIN ISAKSSON, SARIT KHIRIRAT, KLAS SEGELJAKT

*Part of the course* EP3260: Fundamentals of Machine Learning Over Networks

### Reinforcement learning

Reinforcement learning is an area of machine learning inspired by behaviorist psychology, concerned with how software *agents* learn to take *actions* in an *environment* by interacting with it to maximize some notion of cumulative *reward*.




### SARL optimization reformulation

- **Goal: A single agent** determines the policy to maximize the long-term reward, which can be solved by the **Bellman optimality equation**

$$V^\pi = R^\pi + \gamma P^\pi V^\pi.$$

- This can formulated into the equivalent regularized saddle-point optimization

$$\min_\theta \max_w \frac{1}{n}\sum_{t=1}^n \mathcal{L}_t(w,\theta),$$

where

$$\mathcal{L}_t(w,\theta) = \frac{1}{2}w^T(A_t\theta - b_t) - \frac{1}{2}\|w\|^2_{C_t} + \rho\|\theta\|^2$$

$$A_t = \phi_t(\phi_t - \gamma\phi_t')^T$$

$$b_t = \phi_t r_t$$

$$C_t = \phi_t\phi_t^T$$

$\phi_t$ is current feature vector, and

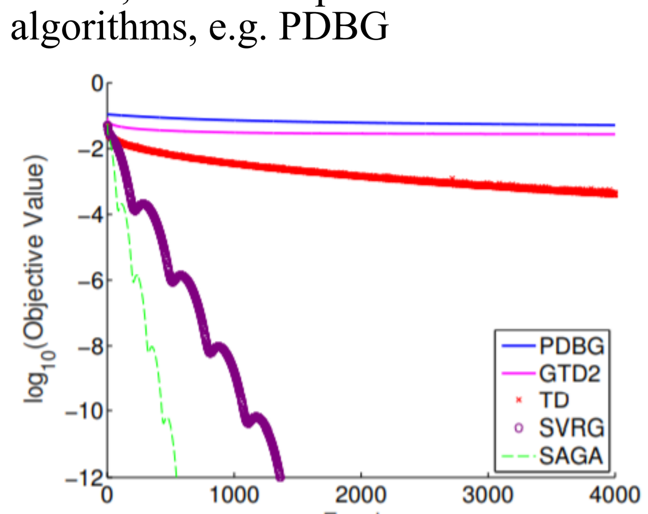
$r_t$ is current reward.

- The saddle-point problem can be solved by classical algorithms, e.g. **gradient descent**, **SVRG**, **SAGA** etc.

$$\theta \leftarrow \theta - \gamma_\theta \frac{1}{n}\sum_{t=1}^n \nabla_\theta \mathcal{L}_t(w,\theta)$$
$$w \leftarrow w + \gamma_w \frac{1}{n}\sum_{t=1}^n \nabla_w \mathcal{L}_t(w,\theta)$$

### MDP simulations

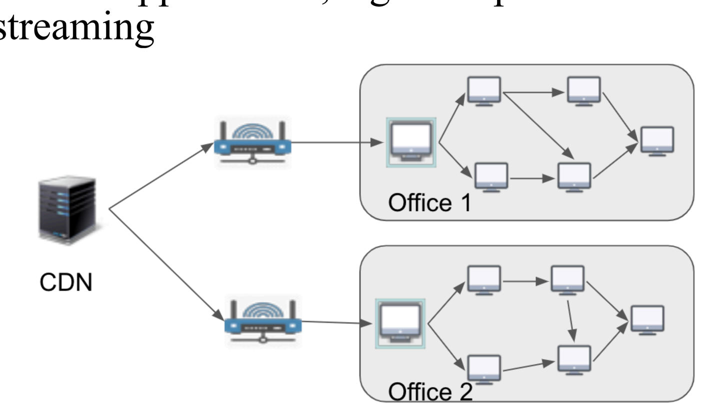
- Generate 400 tasks and 10 actions.
- SVRG, SAGA outperforms traditional algorithms, e.g. PDBG




(a) $\rho = 0$

### Extension to MARL optimization

- MARL applications, e.g. enterprise video streaming




- **Goal: multiple agents** collaboratively determine the policy the maximize long-term reward.
- **Assumptions:** States and actions are broadcasted among agents, while reward is private.
- We can derive multi-agent optimization

$$\min_\theta \max_{w_i, i=1,2,\ldots,N} \frac{1}{n}\frac{1}{N}\sum_{t=1}^n\sum_{i=1}^N \mathcal{L}_t(w_i,\theta),$$

where

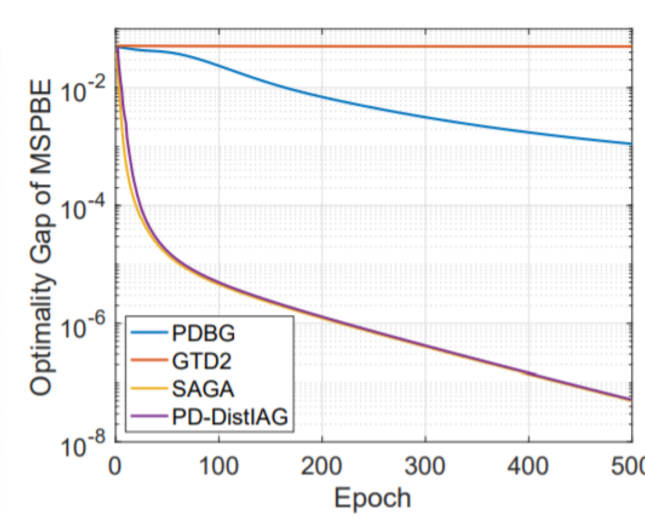
$$\mathcal{L}_t(w_i,\theta) = \frac{1}{2}w_i^T(A_t\theta - b_{t,i}) - \frac{1}{2}\|w_i\|^2_{C_t} + \rho\|\theta\|^2.$$

$$b_{t,i} = \phi_t r_t^i$$

$r_t^i$ is the local reward of agent $i$

- This multi-agent problem can be easily solved by SAGA and consensus-based averaging, `PD-DistIAG` (Wai et.al, 2018).

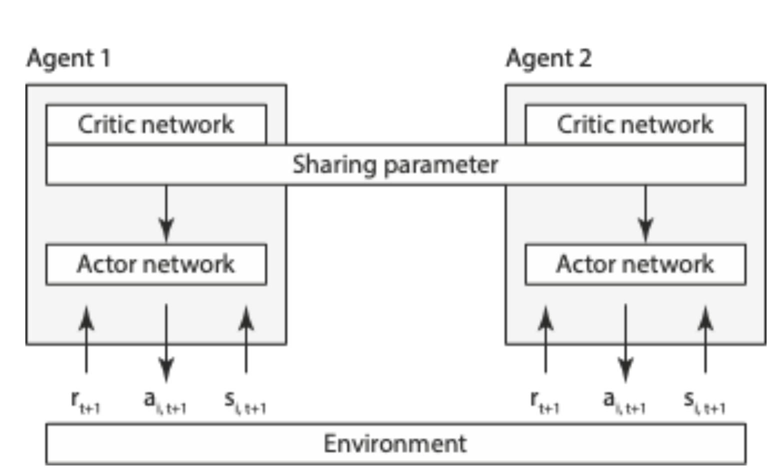
### MDP Simulations on Mountain Car




- `PD-DistIAG` is **comparable** to centralized algorithms (even faster).

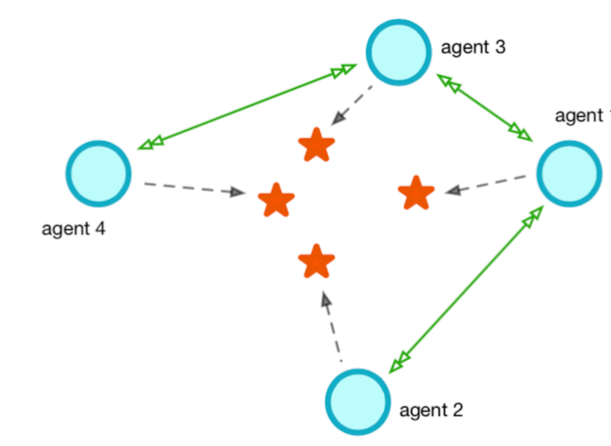
### Actor-critic MARL algorithms

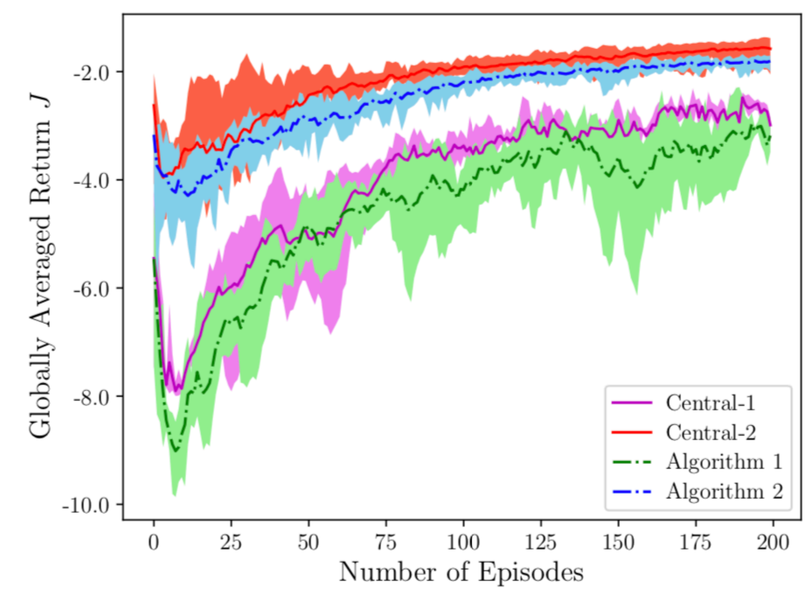
- Actor-critic algorithms can reduce variance but guarantee fast convergence.
- Extended for MARL (Zhang et.al, 2018).




### Safe cooperative navigation simulations

- Each agent's local reward represents a distance to its targeted landmark and a penalty depending on distance to other agents.




- Distributed actor-critic algorithms are **comparable** to centralized counterparts.

### Acknowledgment

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

### References

[1] M. Lanctot *et al.*, "A Unified Game-Theoretic Approach to Multiagent Reinforcement Learning," *Adv. Neural Inf. Process. Syst. 30*, no. Nips, 2017.

[2] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents," 2018.

[3] H.-T. Wai, Z. Yang, Z. Wang, and M. Hong, "Multi-Agent Reinforcement Learning via Double Averaging Primal-Dual Optimization," 2018.

[4] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep Decentralized Multi-task Multi-Agent Reinforcement Learning under Partial Observability," 2017.

[5] L. Matignon, G. J. Laurent, and N. Le Fort-Piat, "Hysteretic Q-Learning : An algorithm for decentralized reinforcement learning in cooperative multi-agent teams," *IEEE Int. Conf. Intell. Robot. Syst.*, pp. 64–69, 2007.

[6] S. Kapoor, "Multi-Agent Reinforcement Learning: A Report on Challenges and Approaches," pp. 1–24, 2018.

[7] Y. Li, "Deep Reinforcement Learning: An Overview," pp. 1–70, 2017.

[8] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, 2017.

[9] D. Lee, H. Yoon, and N. Hovakimyan, "Primal-Dual Algorithm for Distributed Reinforcement Learning: Distributed GTD2," 2018.

[10] Du, Simon S., Jianshu Chen, Lihong Li, Lin Xiao, and Dengyong Zhou. "Stochastic variance reduction methods for policy evaluation." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1049-1058. JMLR. org, 2017.

**IN COOPERATION WITH:**

KTH ROYAL INSTITUTE OF TECHNOLOGY

Stefanos Antaris antaris@kth.se
Amaru Cuba Gyllensten amaru.cuba.gyllensten@ri.se
Martin Isaksson martisak@kth.se
Sarit Khirirat sarit@kth.se
Klas Segeljakt klasseg@kth.se