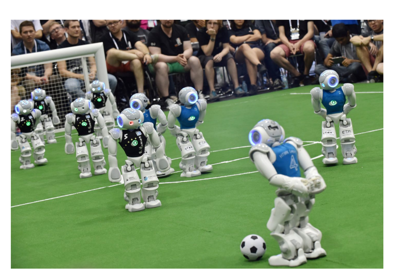
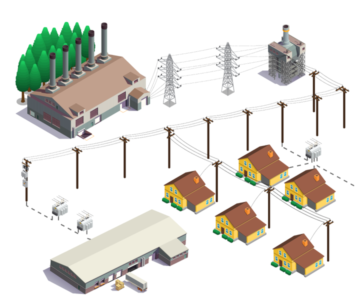
Learning with Safety Constraints: Sample Complexity of Safe Reinforcement Learning

Dileep Kalathil
Assistant Professor, Texas A&M University, College Station, Texas

Reinforcement Learning and Safety

- Reinforcement Learning (RL) addresses the problem of learning to control unknown systems by explicitly considering their inherent dynamical structure
- Standard RL algorithms typically focus only on maximizing a single objective in terms of the value function
- Control policy for any real-world systems should maintain some necessary safety criteria to avoid undesirable outcomes such as avoid collisions, avoid falling down, avoid blackouts

- How do we learn RL algorithms that maximize the objective while satisfying the safety requirements?

Constrained Reinforcement Learning (CRL)

- Constrained Markov Decision Process (CMDP) $M = (S, A, P, r, c, \bar{C}, H)$, where, S: state space, A: action space, P: transition kernel, r: reward c: cost, $\bar{C}$: constraint bound, H: horizon length
- Objective and Constraint Value functions of a policy π

$$V_\pi = \mathbb{E}[\sum_{h=1}^{H} r(s_h, a_h)],\ a_h \sim \pi(s_h), \quad C_\pi = \mathbb{E}[\sum_{h=1}^{H} c(s_h, a_h)],\ a_h \sim \pi(s_h)$$

- Model P is unknown
- CRL Problem

$$\max_\pi \ V_\pi, \quad \text{such that } C_\pi \leq \bar{C}$$

- How do we learn the optimal constrained policy, which is the solution of the CRL problem?
- How do we characterize the performance of the learned policy? How do we characterize the learning efficiency?

Generative Model-Based CRL

- Generative model is a sampling device that gives the next state sample given the current state and action as the input
- Generative model can be used to estimate the underlying transition kernel P, and this model estimate can be used to solve the CMDP problem
- Estimating P: get n_o next state samples from each state-action pair (s,a). Get the maximum likelihood estimate $\hat{P}$
- The CMDP problem may not be feasible for $\hat{P}$
- Solution: solve an optimistic CMDP problem with $\hat{P}$. This can be achieved by an extended linear programming approach [HasanzadeZonuzy et al, 2020a]

Theorem 1 (Sample complexity of Generative Model-Based CRL). [HasanzadeZonuzy et al, 2020a]
Let π_{safe} be the policy obtained from the Generative Model-Based CRL Algorithm with

$$n_o \geq \frac{256}{\epsilon^2} |S| H^3 \log \frac{24|S||A|H}{\delta}$$

Then, $\mathbb{P}(V_{\pi_{safe}} \geq V_{\pi^*} - \epsilon$ *and* $C_{\pi_{safe}} \leq \bar{C} + \epsilon) \geq 1 - \delta.$

Online Model-Based CRL

- RL algorithms often have to collect data in an online way by generating sequential trajectories
- Exploration vs exploitation trade-off is a fundamental problem in any online learning algorithm
  - Exploration: gather data to learn P
  - Exploitation: select actions that maximize the objective without violating the constraints
- Objective: Learn a safe policy with minimum number of online samples, with provable guarantees on performance

Algorithm Online Model-Based CRL

1: Input: problem paramaters (ϵ, δ). Initial policy π_1
2: Set $n(s, a) = n(s', s, a) = 0 \ \forall s, s' \in S, a \in A$.
3: Fix algorithm parameter $m(\epsilon, \delta)$
4: **while** there is (s, a) with $n(s, a) < |S|Hm(\epsilon, \delta)$ **do**
5: **for** episode $k = 1, 2, \ldots$ **do**
6: **for** $t = 1, \ldots, H$ **do**
7: Collect samples using the exploration policy π_k: $a_t \sim \pi_k(s_t), s_{t+1} \sim P(\cdot|s_t, a_t)$
8: Update the counts: $n(s_t, a_t) + +,\ n(s_{t+1}, s_t, a_t) + +$
9: Estimate $\hat{P}$: $\hat{P}(s'|s, a) = \frac{n(s', s, a)}{n(s, a) \wedge 1}, \ \forall (s, a)$
10: Solve the optimistic CMDP problem using $\hat{P}$ and confidence estimate to get π_{k+1}
11: Output $\pi_{\text{safe}} = \pi_k$

Theorem 2 (Sample complexity of Online Model-Based CRL). [HasanzadeZonuzy et al, 2020a] *Under Online Model-Based CRL algorithm, with appropriate $m(\epsilon, \delta)$, we get* $\mathbb{P}(V_{\pi_k} \geq V_{\pi^*} - \epsilon$ *and* $C_{\pi_k} \leq \bar{C} + \epsilon) \geq 1 - \delta$*, for all but at most* $\tilde{O}(\frac{|S|^2|A|H^2}{\epsilon^2} \log \frac{|S||A|}{\delta})$ *episodes.*

Boarder Impact

- Outreach lecture at the Texas A&M University Physics and Engineering Festival on the topic of "A Path to Artificial Intelligence Through Reinforcement Learning"
- Mentoring undergraduate students through Louis Stokes Alliance For Minority Participation (LSAMP), TAMU

Publications

1. A. HasanzadeZonuzy, A., Bura, D. Kalathil, S. Shakkottai, ``Learning with safety constraints: Sample complexity of reinforcement learning for constrained MDPS'', AAAI Conference on Artificial Intelligence, February, 2021
2. A. HasanzadeZonuzy, A., Bura, D. Kalathil, S. Shakkottai, ``Model-Based Reinforcement Learning for Infinite-Horizon Discounted Constrained Markov Decision Processes'', International Joint Conference on Artificial Intelligence (IJCAI), August, 2021

2021 NSF Cyber-Physical Systems Principal Investigators' Meeting
June 2-4, 2021

Award ID#:1850206: CRII: CPS: Towards a Model-Based Reinforcement Learning Approach for Safe Operation of Distributed Energy Systems
PI: Dileep Kalathil (Texas A&M University)