Intro to RL: Notes

Aga Slowik

July 2020

1 Lecture 1

Link: TD-Gammon

Link: Summary of Chapter 2 from the RL book. The transition from multi-armed bandits to full RL through contextual bandits, ways of balancing exploration and exploitation: ϵ-greedy, UCB

Action-Value Methods: e.g. sample average $Q_t(a)$

The sample average converges to the optimal value for an action a if a has been taken an infinite number of times.

Standard form for the learning/update rules: $NewEstimate = OldEstimate + StepSize[Reward - OldEstimate]$

In a non-stationary env (non-stationary bandit): exponential, recency-weighted average

$Q_{n+1} = (1-\alpha)^n Q_1 + \sum_{i=1}^{n} \alpha(1-\alpha)^{n-i} R_i$

$\alpha \in (0, 1]$

Q function - estimate value of (s, a) under the policy π at the timestep t. (future return starting at $t + 1$) (*action-value function*)

V - *state-value function*

(V, Q - random variables)

We are interested in maximizing expected (future) return starting from the timestep t.

In episodic task, this is usually a sum of the future rewards in an episode.

In continuing tasks (no natural episodes), use a discount factor $\gamma \in [0, 1]$ (typically $\gamma = 0.9$ which means we are rather farsighted).

Mean Square Value Error in RL ($\mu(s)$ - the fraction of timesteps spent in the state s/distribution). Similar to regression but: the IID input assumption does not work (returns and inputs are correlated as they lie on the same trajectory). Gradient Monte Carlo algorithm. State aggregation.