Written Notes about Paper
Survey of Generalization in Deep RL

1. Reinforcement Learning – section 3.2.

- RL: Framework for learning - how to interact with the environment from Experience.

MDP (Markovian Decision Process) : 4-tuple $(S, A, R, P) = \mathcal{M}$

→ S: set of states

→ A: set of actions

→ R: immediate rewards after going from state s to s' with action a

$$R(s', s, a) = \mathbb{P}\{r_{k+1} \mid s_{k+1} = s',\ s_k = s,\ a_k = a\}$$ (REWARD FUNCTION)

→ P: prob. to get from state s to s' with action a

$$P(s', s, a) = \mathbb{P}\{s_{k+1} = s' \mid s_k = s,\ a_k = a\}$$ STOCHASTIC (MARKOVIAN TRANSITION FUNCTION)

- Policy $\pi(a|s)$: distribution over actions given a state
  ↳ $\mathbb{P}\{A = a \mid S = s\}$

GOAL: Optimize the policy $\pi(a|s)$ such that the cumulative rewards of the policy in the MDP is maximized:

$$\pi^* = \underset{\pi \in \Pi}{\operatorname{argmax}}\ V_\pi(s)$$

⇒ Value Function (with policy π): Total expected reward gained by the policy π from a state s.

$$V_\pi(s) = \underset{s \sim p(s_0)}{\mathbb{E}}\left\{ \sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1}) \mid s_0 = s \right\}$$

↳ initial state distribution. → Discount Rate $0 < \gamma \leq 1$

⇒ Quality Function: (with policy π) Total expected reward gained by the policy π from state s and action a.

$$Q_\pi(s, a) = \underset{\substack{s \sim p(s_0) \\ a \sim \pi(a|s)}}{\mathbb{E}}\left\{ \sum_{\tau=0}^{\infty} \gamma^\tau R(s_t, a_t, s_{t+1}) \mid s_0 = s,\ a_0 = a \right\}$$

$$\Leftrightarrow\ V_\pi(s) = \sum_{a \in A} \pi(a|s)\, Q_\pi(s, a)$$