Google slides presentation:
https://docs.google.com/presentation/d/1ChUC3ehOT_pVpq15XWf1JOlLQKS2HJmTaHpJU9zbFMM/edit?usp=sharing

GitHub repository: https://github.com/fewagner/icsc23

dynamics function $p : (A,\, S) \mapsto \{\text{probabilities for } S'\}$

reward function $r : \left(S,\, A,\, S'\right) \mapsto R \in \mathbb{R}$

policy function $\pi : S \mapsto \pi(A \mid S) = \{\text{probabilities for } A \,\text{given state}\, S\}$

returns with discounting factor gamma $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + \dots$

state values and action-state values

$$v_\pi(s) = E_\pi[G_t \mid S_t = s]$$
$$q_\pi(s,a) = E_\pi[G_t \mid S_t = s, A_t = a]$$

Bellman equations for state and action-state values

$$v_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) \mid S_t = s]$$
$$q_\pi(s,a) = \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) \mid S_t = s, A_t = a]$$

“Epsilon-greedy” policy: Take greedy action with probability 1-ε and random action with probability ε. (greedy action = action with highest value)

SARSA value update rule

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha[R_{t+1} + \gamma Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t)]$$

Q-learning value update rule

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha\left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t)\right]$$

Actor-critic gradient update for value function parameters

$$w_{t+1} = w_t + \alpha_w \left(R_{t+1} + \gamma \hat{v}_w(S_{t+1}) - v_w(\hat{S}_t)\right) \nabla_w \hat{v_{w_t}}(S_t)$$

Actor-critic gradient update for policy function parameters

$$\theta_{t+1} = \theta_t + \alpha_\theta \gamma^t \left(R_{t+1} + \gamma \hat{v}_w(S_{t+1}) - v_w(\hat{S}_t)\right) \nabla_\theta \ln \pi_{\theta_t}(A_t|S_t)$$