Episodic SARSA with Function approximation

Initialize:

- Parameters θ (randomly or zero)
- Policy $\pi_\theta(a|s)$ (e.g., ϵ-greedy w.r.t. $\hat{q}(s, a; \theta)$)

For each episode:

1. Initialize state S_0, choose action $A_0 \sim \pi_\theta(\cdot\,|S_0)$
2. For each time step t:
    - Take action A_t, observe R_{t+1}, S_{t+1}
    - Choose $A_{t+1} \sim \pi_\theta(\cdot\,|S_{t+1})$
    - Compute **TD error**:

$$\delta_t = R_{t+1} + \gamma \hat{q}\,(S_{t+1}, A_{t+1}; \theta) - \hat{q}\,(S_t, A_t; \theta)$$

    - Update parameters:

$$\theta \leftarrow \theta + \alpha \cdot \delta_t \cdot \nabla_\theta \hat{q}(s, a; \theta)$$

3. Repeat until episode ends

Exploration under function approximation

1. **ε-greedy**: Select a random action with probability ε, otherwise pick greedy
2. **Softmax / Boltzmann Exploration**: Assign probabilities to actions using a temperature parameter:

$$\pi(a|s) = \frac{e^{\hat{q}(s,a)/r}}{\sum_b e^{\hat{q}(s,b)/r}}$$

3. **Entropy Regularization**:
Encourage the policy to remain stochastic (used in Actor-Critic, PPO):

$$J(\theta) = \mathbb{E}[\log \pi(a|s; \theta) A(s, a)] + \beta \cdot \mathcal{H}[\pi(\cdot\,|S)]$$

Average Reward

For a **continuing task** under a policy π\piπ, the **average reward** is:

$$r(\pi) = \lim_{h \to \infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}_{\pi}[R_t]$$

- Measures **long-run average rate of reward** (per time step)
- Does **not** require a discount factor γ
- Often more natural for **steady-state systems** (e.g., server management, process control)

Relative Value Function

Instead of traditional $v_{\pi}(s)$we define the **differential value function**:

$$h_{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty}(R_{t+1} - r(\pi))|S_0 = s\right]$$

- Measures **relative desirability** of a state compared to average performance
- Focuses learning on **what's better or worse** than the norm

TD Learning with Average Reward

You can extend **SARSA**, **Actor-Critic**, and other methods to use average reward instead of discounted return.

Example: Average Reward SARSA Update

Let $\bar{R}$ be the estimate of average reward. Then:

$$\delta_t = R_{t+1} - \bar{R} + \hat{q}(S_{t+1}, A_{t+1}; \theta) - \hat{q}(S_t, A_t; \theta)$$

- Update θ with TD error δ_t
- Update average reward estimate:

$$\bar{R} \leftarrow \bar{R} + \beta \cdot \delta_t$$