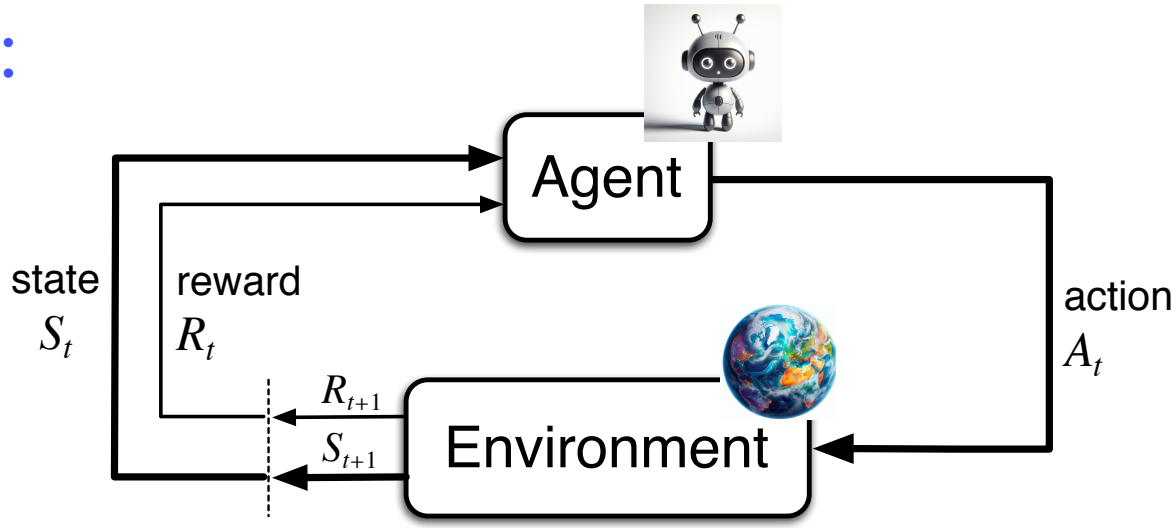
Lecture 17: Review

RL Setting:



Agent and environment interact at discrete time steps: $t = 0, 1, 2, 3, \ldots$

Agent observes state at step t: $S_t \in \mathcal{S}$

produces action at step t : $A_t \in \mathcal{A}(S_t)$

gets resulting reward: $R_{t+1} \in \mathcal{R} \subset \mathbb{R}$

and resulting next state: $S_{t+1} \in \mathcal{S}^+$

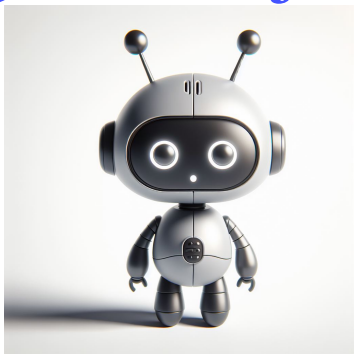
Assumption about the Environment:

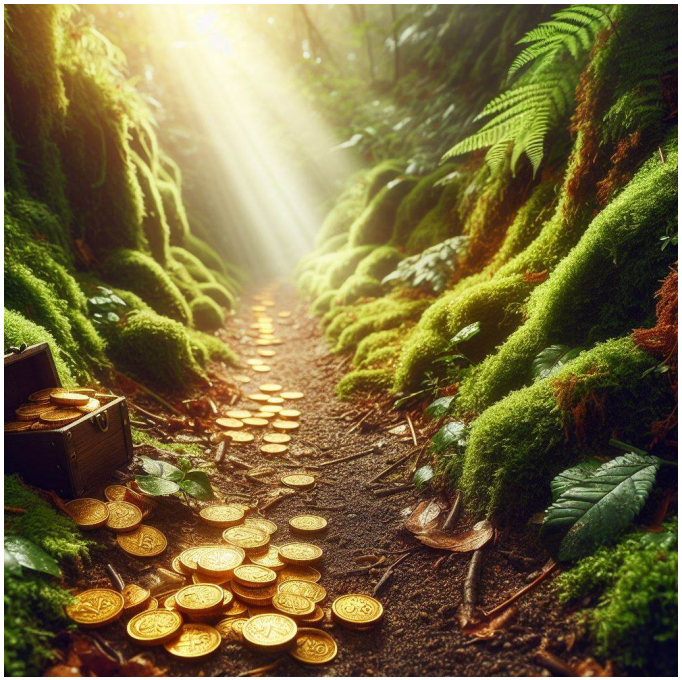
Environment is Markov Decision Process (MDP)

$$p(S_{t+1}, R_{t+1} \mid A_t, S_t, A_{t-1}, S_{t-1}, \ldots, S_0)$$

$$= p(S_{t+1}, R_{t+1} \mid A_t, S_t)$$

Agent's objective:



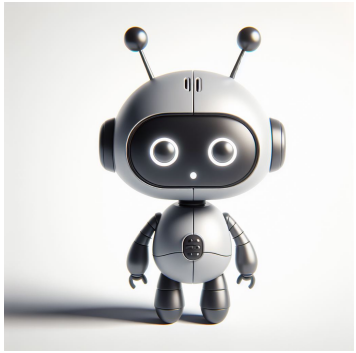


Maximize:

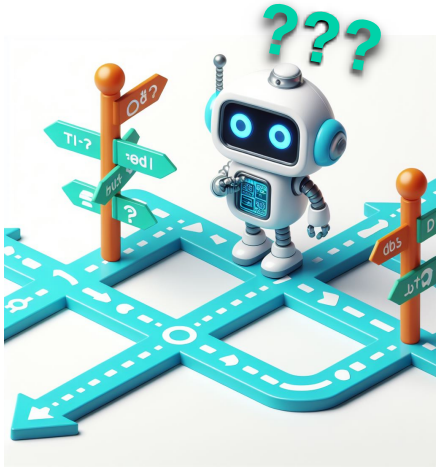
$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1},$$

where γ, $0 \leq \gamma \leq 1$, is the **discount rate**.

Agent does 2 things:

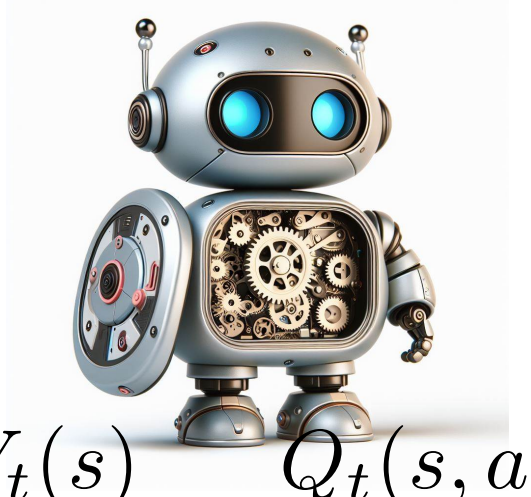


Choose actions:



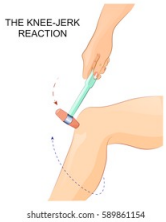
Learn how to choose better actions:

Different Parts of an Agent:



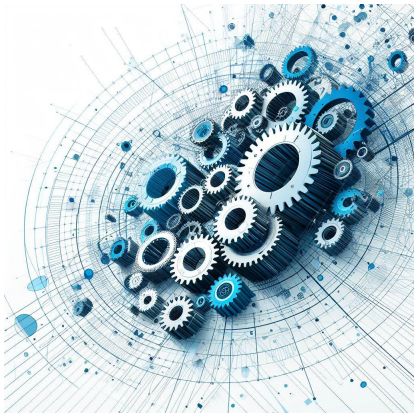
- Value Functions :  $V_t(s)$ $Q_t(s,a)$
- World Model:  $S_{t+1}, R_{t+1} = M(S_t, A_t, \theta)$
- Policy:  $A_t = \pi(S_t, \theta)$
- (Replay Buffer of past experience)

TWO TYPES OF POLICY $\pi(S_t, \theta)$:



Stochastic:

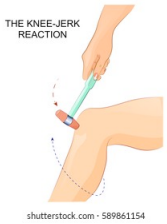
Deterministic:



$\pi(A_t, S_t, \theta)$ is probability.

$A_t = \pi(S_t, \theta)$

Stochastic: $\pi(A_t, S_t, \theta)$ is probability.



Act by sampling from the distribution:

Discrete Actions:

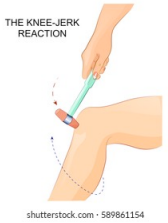
$$\pi(A_i \mid S) = \frac{\exp(\phi(A_i, S))}{\sum_j \exp(\phi(A_j, S))}$$

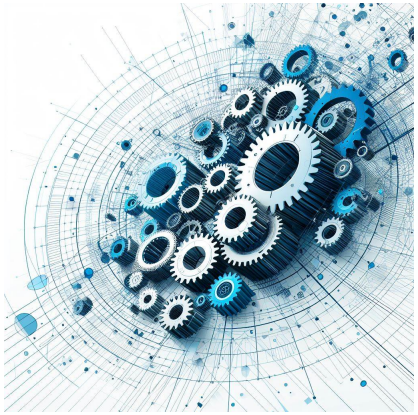
Continuous Actions:

$$\pi(A \mid S) = \mathcal{N}\left(\mu(S), \sigma(S)\right)$$

$$A = \mu(S) + \sigma(S)\epsilon, \qquad \epsilon \sim \mathcal{N}(0,1)$$

Deterministic: $A_t = \pi(S_t, \theta)$





Act by applying π to state: $A_t = \pi(S_t, \theta)$

Value Functions

- The **value of a state** is the expected return starting from that state; depends on the agent's policy:

State - value function for policy π :

$$v_\pi(s) = E_\pi\left\{G_t \mid S_t = s\right\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right\}$$

- The **value of an action (in a state)** is the expected return starting after taking that action from that state; depends on the agent's policy:

Action - value function for policy π :

$$q_\pi(s,a) = E_\pi\left\{G_t \mid S_t = s, A_t = a\right\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right\}$$

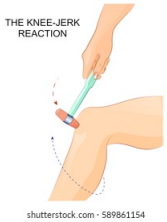
Use $V_t(s)$ $Q_t(s,a)$ to choose action:

Action - value function for policy π :

$$q_\pi(s,a) = E_\pi\left\{G_t \mid S_t = s, A_t = a\right\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right\}$$

Act by taking the action which maximizes the expected return according to the estimate Q:

$$A_t = \text{argmax}_a Q(a, S_t)$$

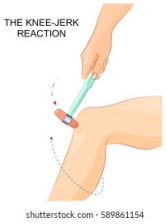


Learning $\pi(S_t, \theta)$:

Action - value function for policy π :

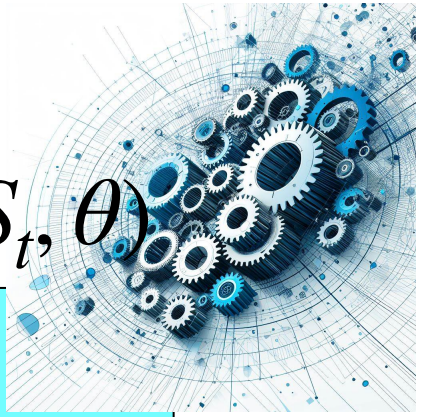
$$q_\pi(s,a) = E_\pi\left\{G_t \mid S_t = s, A_t = a\right\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right\}$$

$$\theta_{t+1} = \theta_t + \alpha \underbrace{\nabla_\theta q_\pi}_{\nabla_\theta J}$$



Learning $\pi(S_t, \theta)$:

Deterministic and Continuous: $A_t = \pi(S_t, \theta)$



Action - value function for policy π :

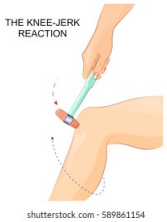
$$q_\pi(s,a) = E_\pi\left\{G_t \mid S_t = s, A_t = a\right\} = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s, A_t = a\right\}$$

$$J_\theta(\pi \mid S_0 = S) = q_\pi(\pi(S), S) \approx Q_\pi(\pi(S,\theta), S)$$

$$\nabla_\theta J_\theta(\pi \mid S_0 = S) \approx \nabla_\theta Q_\pi(\pi(S,\theta), S)$$

$$= \sum_i^m \frac{\partial Q_\pi(A = \pi(S,\theta), S)}{\partial a_i} \nabla_\theta \pi_i(S,\theta)$$

$$= \nabla_A Q_\pi\left(A = \pi(S,\theta), S\right) \nabla_\theta \pi(S,\theta)$$

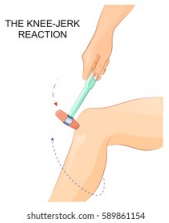


Learning $\pi(S_t, \theta)$:

Stochastic: $\pi(A_t, S_t, \theta)$ is probability.

Policy Gradient Theorem:

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{T} (\gamma^t G_t) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$

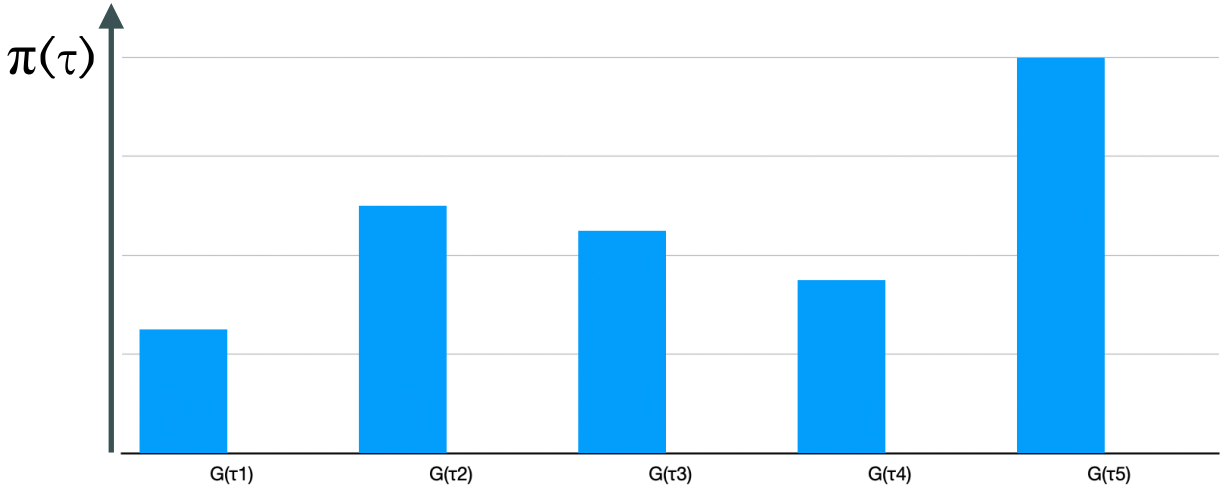


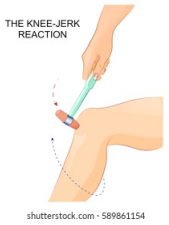
Learning $\pi(S_t, \theta)$:

Stochastic: $\pi(A_t, S_t, \theta)$ is probability.

Policy Gradient Theorem:

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{T} (\gamma^t G_t) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$





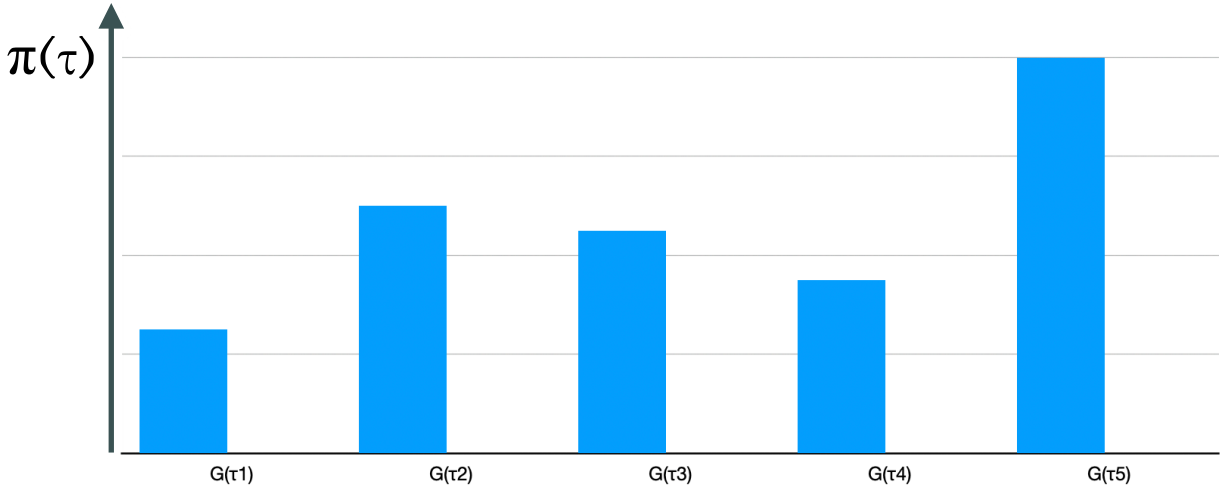
Learning $\pi(S_t, \theta)$:

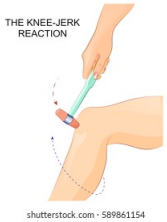
Stochastic: $\pi(A_t, S_t, \theta)$ is probability.

Policy Gradient Theorem:

REINFORCE Estimates G with Monte-Carlo

$$\nabla_{\boldsymbol{\theta}} J_{\boldsymbol{\theta}}(\pi) = \mathbb{E}_{\pi}[\sum_{t=0}^{T} (\gamma^t G_t) \nabla_{\boldsymbol{\theta}} \log \pi(A_t | S_t)]$$





Learning $\pi(S_t, \theta)$:

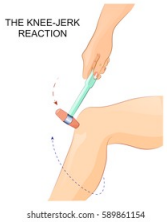
Stochastic: $\pi(A_t, S_t, \theta)$ is probability.

Policy Gradient Theorem:

REINFORCE Estimates G with Monte-Carlo

$$\nabla_\theta J_\theta(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \left(q_\pi(S_t, A_t) - v_\pi(S_t)\right) \nabla_\theta \log(\pi)\right]$$

Advantage



Learning $\pi(S_t, \theta)$:

Stochastic: $\pi(A_t, S_t, \theta)$ is probability.

Policy Gradient Theorem:

$$\nabla_\theta J_\theta(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \underbrace{\left(q_\pi(S_t, A_t) - v_\pi(S_t)\right)}_{\text{Advantage}} \nabla_\theta \log(\pi)\right]$$

Actor-Critic: use V and/or Q to estimate G or Advantage , e.g. TD(λ)

Learn $V_t(s)$ $Q_t(s,a)$:

$$q_\pi(s,a) = \mathbb{E}\{G_t \mid S_t = s, A_t = a, A_{t+1:\infty} \sim \pi\} \qquad q_\pi : \mathcal{S} \times \mathcal{A} \to \Re$$

Monte-Carlo Estimate :

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1},$$

where $\gamma, 0 \le \gamma \le 1$, is the **discount rate**.

- *Every-Visit MC:* average returns for *every* time *s* is visited in an episode
- *First-visit MC:* average returns only for *first* time *s* is visited in an episode
- Both converge asymptotically

Learn $V_t(s)$ $Q_t(s,a)$:

$$q_\pi(s,a) = \mathbb{E}\{G_t \mid S_t = s, A_t = a, A_{t+1:\infty} \sim \pi\} \qquad q_\pi : \mathcal{S} \times \mathcal{A} \to \Re$$

Bootstrapping :

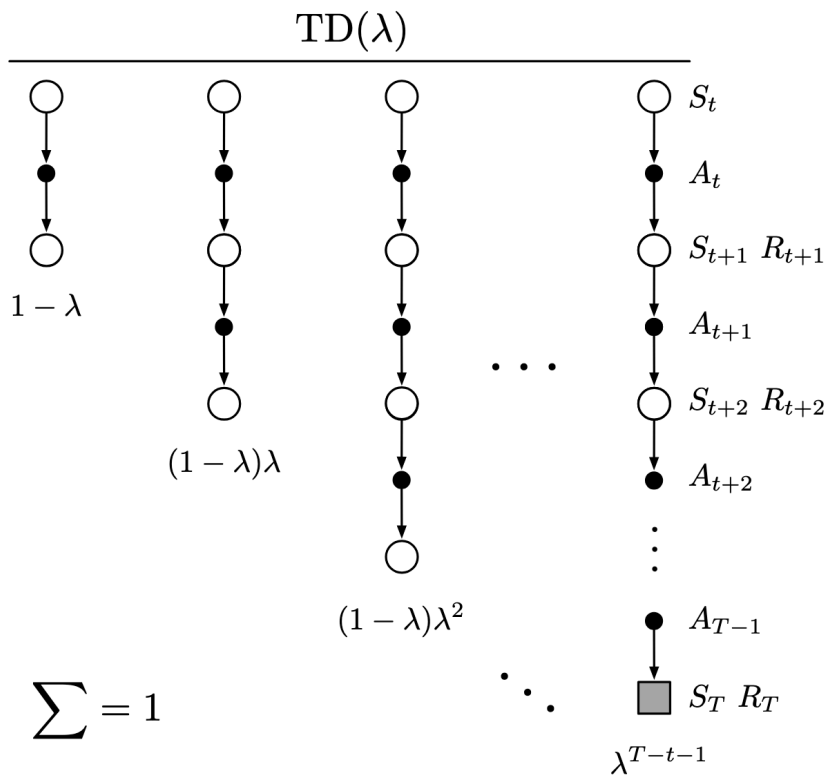
- TD: $G_t^{(1)} \doteq R_{t+1} + \gamma V_t(S_{t+1})$
  - Use V_t to estimate remaining return
- *n*-step TD:
  - 2 step return: $G_t^{(2)} \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 V_t(S_{t+2})$
  - *n*-step return: $G_t^{(n)} \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V_t(S_{t+n})$

    with $G_t^{(n)} \doteq G_t$ if $t + n \geq T$

The λ-return is a compound update target

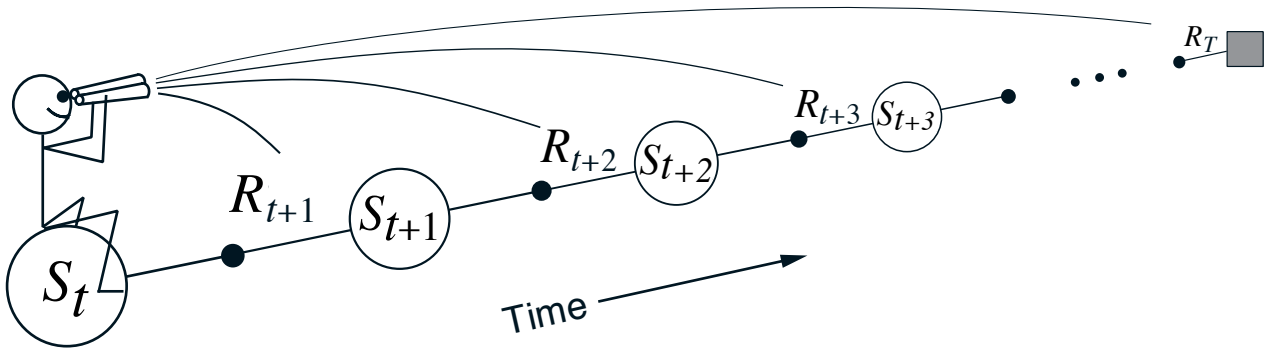
- The λ-return a target that averages all *n*-step targets
- Each weighted by λ^{n-1}

$$G_t^\lambda = (1-\lambda)\sum_{n=1}^{T-t-1}\lambda^{n-1}G_{t:t+n} + \lambda^{T-t-1}G_t,$$



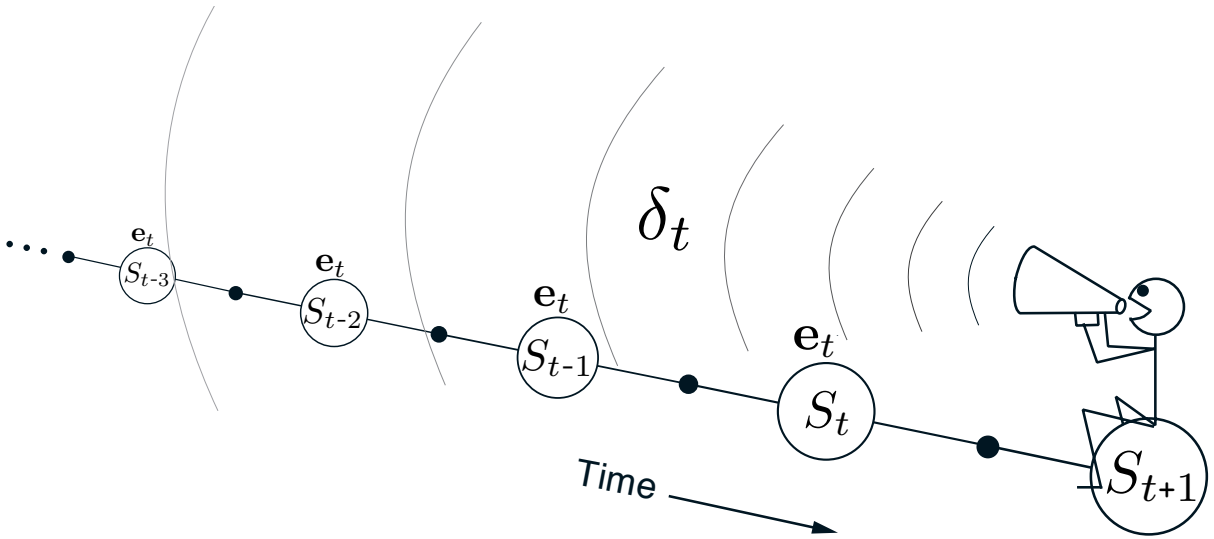
The forward view

- Look forward from each state to determine update from future states and rewards:



The backward view

- Shout the TD error backwards
- The traces fade with temporal distance by γλ



Eligibility traces (mechanism)

- The forward view was for theory
- The backward view is for *mechanism*

- New memory vector called *eligibility trace* $\mathbf{e}_t \in \mathbb{R}^n \geq \mathbf{0}$ (same shape as $\boldsymbol{\theta}$)
  - On each step, decay each component by $\gamma\lambda$ and increment the trace for the current state by 1
  - *Accumulating trace*

$$\mathbf{e}_0 \doteq \mathbf{0},$$
$$\mathbf{e}_t \doteq \nabla \hat{v}(S_t, \boldsymbol{\theta}_t) + \gamma\lambda \mathbf{e}_{t-1}$$

  - *Replacing trace:* trace becomes 1 when state is visited

Learn $V_t(s)$ $Q_t(s,a)$:

$$q_\pi(s,a) = \mathbb{E}\{G_t \mid S_t = s, A_t = a, A_{t+1:\infty} \sim \pi\} \qquad q_\pi : \mathcal{S} \times \mathcal{A} \to \Re$$

(Expected) SARSA (Bellman Eqn) :

$$\begin{aligned} Q(S_t,A_t) &\leftarrow Q(S_t,A_t) + \alpha\Big[R_{t+1} + \gamma\mathbb{E}[Q(S_{t+1},A_{t+1}) \mid S_{t+1}] - Q(S_t,A_t)\Big] \\ &\leftarrow Q(S_t,A_t) + \alpha\Big[R_{t+1} + \gamma\sum \pi(a|S_{t+1})Q(S_{t+1},a) - Q(S_t,A_t)\Big] \end{aligned}$$

Learn $V_t(s)$ $Q_t(s,a)$:

$$q_*(s,a) = \max_\pi q_\pi(s,a) \qquad q_* : \mathcal{S} \times \mathcal{A} \to \Re$$

Q-Learning (Bellman Optimality Eqn):

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