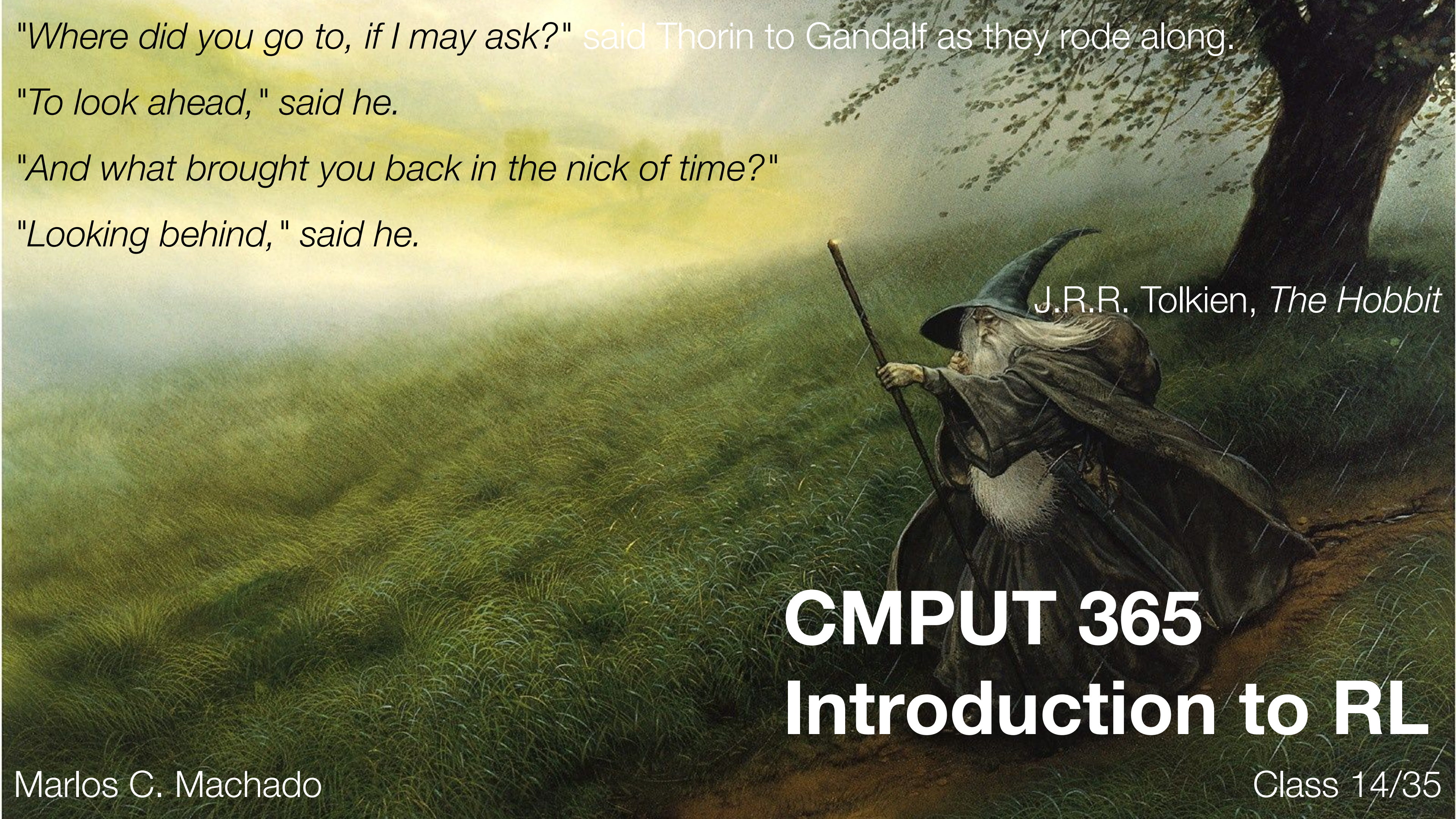


*"Where did you go to, if I may ask?"* said Thorin to Gandalf as they rode along.

*"To look ahead," said he.*

*"And what brought you back in the nick of time?"*

*"Looking behind," said he.*

J.R.R. Tolkien, *The Hobbit*

# CMPUT 365 Introduction to RL

Marlos C. Machado

Class 14/35

# Reminder

You **should be enrolled in the private session** we created in Coursera for CMPUT 365.

I **cannot** use marks from the public repository for your course marks.

You **need** to **check**, **every time**, if you are in the private session and if you are submitting quizzes and assignments to the private section.

There were **17 pending invitations** last time I checked!

If you have any questions or concerns, **talk with the TAs** or email us `cmput365@ualberta.ca`.

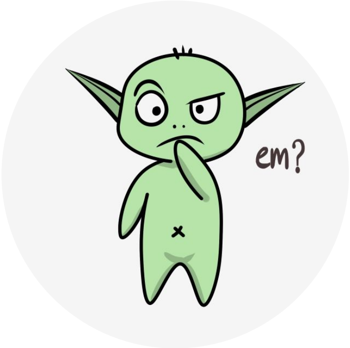
# Reminders and Notes

- On the midterm:
  - The marks are now available on eClass. I will share instructions for the exam viewing soon.
- What **I** plan to do today:
  - Continue overview of Monte Carlo Methods for Prediction & Control (Chapter 5 of the textbook).
- What I recommend **YOU** to do:
  - Graded Quiz (Off-policy Monte Carlo) is due today.
  - Programming Assignment is not graded this week.

# SPOT: Mid-term Course Evaluation

https://go.blueja.io/MlqAHuUezE-my_PTHx9IEg

# Please, interrupt me at any time!





https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

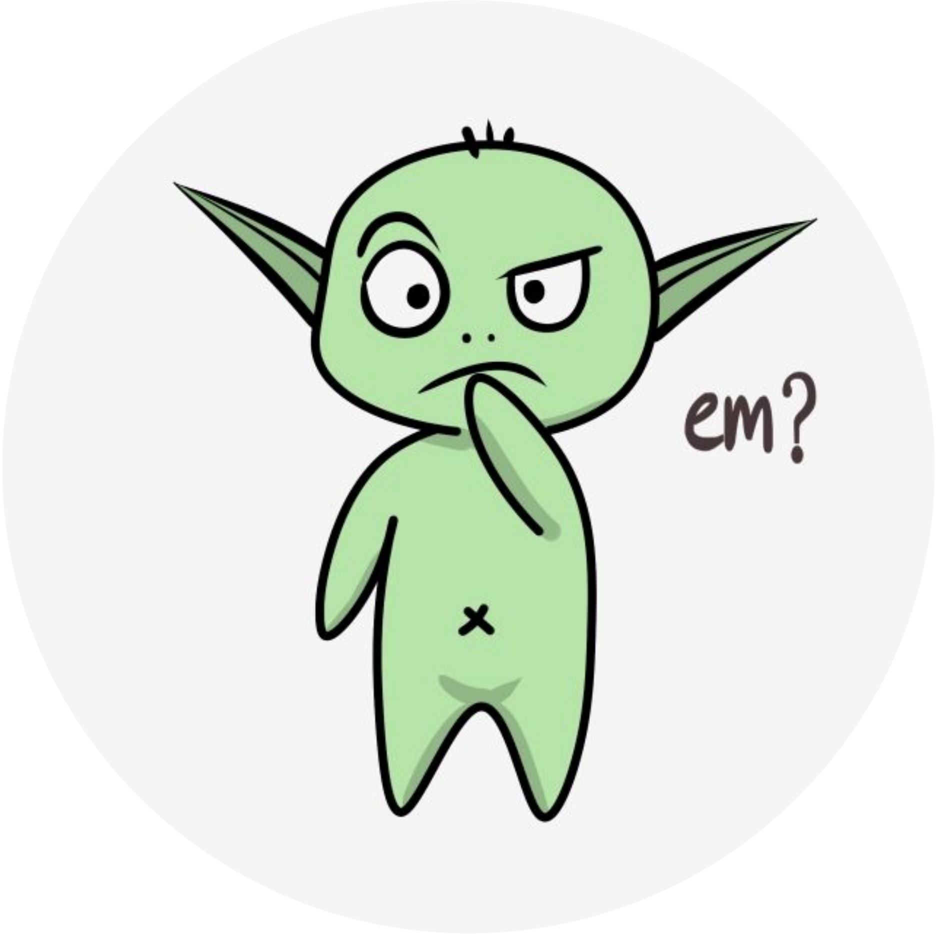
# Last Class: Monte Carlo Prediction

**First-visit MC prediction, for estimating $V \approx v_\pi$**

Input: a policy $\pi$ to be evaluated

Initialize:
- $V(s) \in \mathbb{R}$, arbitrarily, for all $s \in \mathcal{S}$
- $Returns(s) \leftarrow$ an empty list, for all $s \in \mathcal{S}$

Loop forever (for each episode):
- Generate an episode following $\pi$: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$
- $G \leftarrow 0$
- Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
  - $G \leftarrow \gamma G + R_{t+1}$
  - Unless $S_t$ appears in $S_0, S_1, \ldots, S_{t-1}$:
    - Append $G$ to $Returns(S_t)$
    - $V(S_t) \leftarrow \text{average}(Returns(S_t))$

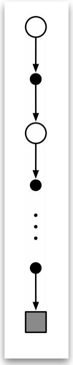




https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

# Some useful information / reminders about MC Methods

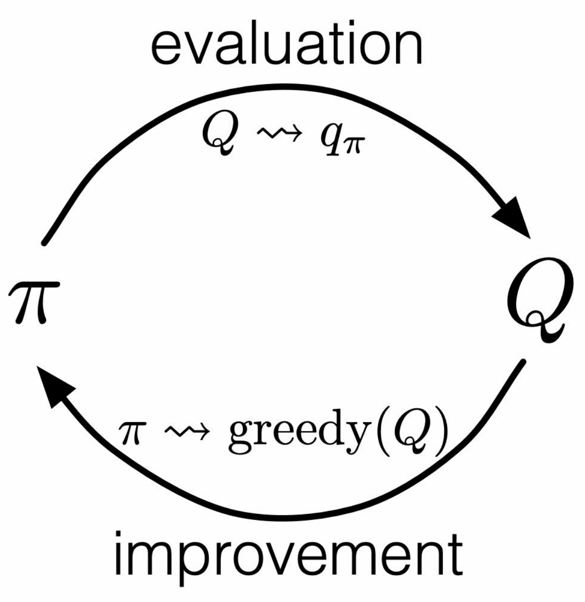
- Often it is much easier to get samples than to get the distribution of next events. Recall the Blackjack example in the textbook.
- Monte Carlo methods do not *bootstrap* (the estimate for one state does not build upon the estimate of any other state).
- First/every-visit MC converge to $v_\pi(s)$ as the number of visits to s goes to infinity. In first-visit MC, each return is i.i.d. and has finite variance ¯_(ツ)_/¯
- The computational cost of estimating the value of a single state is independent of the number of states.



# Monte Carlo Estimation of Action Values

- If we don't have access to a model, we need to estimate *action* values.
- Same as before, but now we visit state-action pairs ¯_(ツ)_/¯
  But to estimate $q_*$ we need to estimate the value of *all* actions from each state.
  Solution? Exploration! … or exploring starts 😒

# Monte Carlo Control




$$\pi_0 \xrightarrow{\text{E}} q_{\pi_0} \xrightarrow{\text{I}} \pi_1 \xrightarrow{\text{E}} q_{\pi_1} \xrightarrow{\text{I}} \pi_2 \xrightarrow{\text{E}} \cdots \xrightarrow{\text{I}} \pi_* \xrightarrow{\text{E}} q_*$$

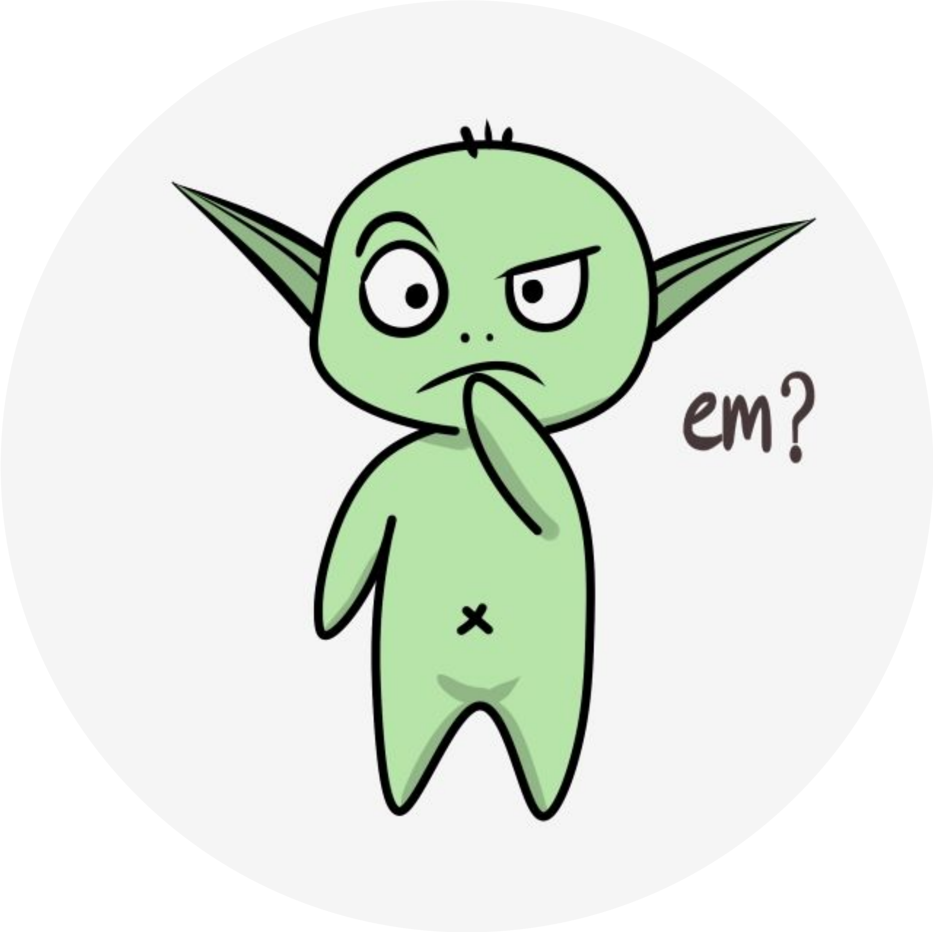
$$\pi(s) \doteq \arg\max_a q(s, a)$$

# Monte Carlo ES

**Monte Carlo ES (Exploring Starts), for estimating $\pi \approx \pi_*$**

Initialize:
  $\pi(s) \in \mathcal{A}(s)$ (arbitrarily), for all $s \in \mathcal{S}$
  $Q(s, a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$
  $Returns(s, a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Loop forever (for each episode):
  Choose $S_0 \in \mathcal{S}$, $A_0 \in \mathcal{A}(S_0)$ randomly such that all pairs have probability $> 0$
  Generate an episode from $S_0, A_0$, following $\pi$: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    Unless the pair $S_t, A_t$ appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:
      Append $G$ to $Returns(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$





https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht

# MC Control without Exploring Starts

**On-policy first-visit MC control (for $\varepsilon$-soft policies), estimates $\pi \approx \pi_*$**

Algorithm parameter: small $\varepsilon > 0$

Initialize:

$\quad \pi \leftarrow$ an arbitrary $\varepsilon$-soft policy

$\quad Q(s,a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

$\quad Returns(s,a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

$\quad$ Generate an episode following $\pi$: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

$\quad G \leftarrow 0$

$\quad$ Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

$\quad\quad G \leftarrow \gamma G + R_{t+1}$

$\quad\quad$ Unless the pair $S_t, A_t$ appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

$\quad\quad\quad$ Append $G$ to $Returns(S_t, A_t)$

$\quad\quad\quad Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

$\quad\quad\quad A^* \leftarrow \arg\max_a Q(S_t, a)$ $\quad$ (with ties broken arbitrarily)

$\quad\quad\quad$ For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$

**We need to ensure that the probability we select each action is not zero.**

# MC Control without Exploring Starts

**On-policy: You learn about the policy you used to make decisions.**

**Off-policy: You learn about a policy that is different from the one you used to make decisions.**

**$\boxed{\text{On-policy}}$ first-visit MC control (for $\varepsilon$-soft policies), estimates $\pi \approx \pi_*$**

Algorithm parameter: small $\varepsilon > 0$

Initialize:

  $\pi \leftarrow$ an arbitrary $\varepsilon$-soft policy

  $Q(s, a) \in \mathbb{R}$ (arbitrarily), for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

  $Returns(s, a) \leftarrow$ empty list, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$

Repeat forever (for each episode):

  Generate an episode following $\pi$: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$

  $G \leftarrow 0$

  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:

    $G \leftarrow \gamma G + R_{t+1}$

    Unless the pair $S_t, A_t$ appears in $S_0, A_0, S_1, A_1 \ldots, S_{t-1}, A_{t-1}$:

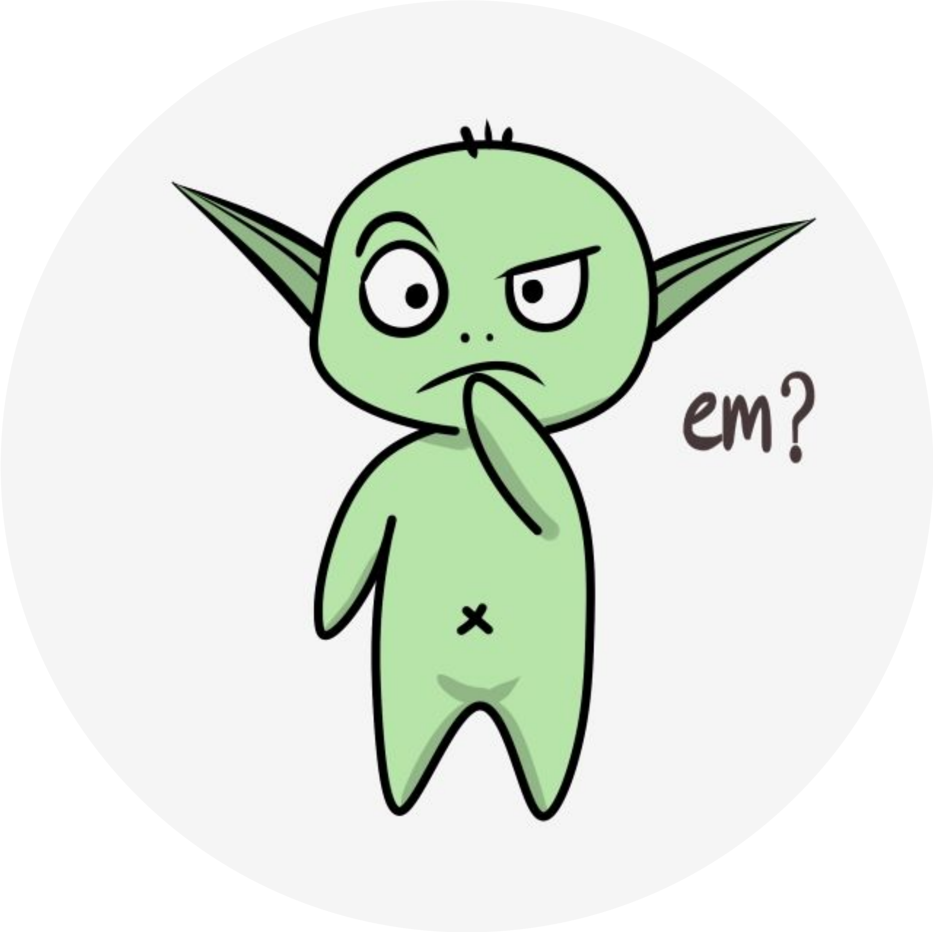
      Append $G$ to $Returns(S_t, A_t)$

      $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$

      $A^* \leftarrow \arg\max_a Q(S_t, a)$    (with ties broken arbitrarily)

      For all $a \in \mathcal{A}(S_t)$:

$$\pi(a|S_t) \leftarrow \begin{cases} 1 - \varepsilon + \varepsilon/|\mathcal{A}(S_t)| & \text{if } a = A^* \\ \varepsilon/|\mathcal{A}(S_t)| & \text{if } a \neq A^* \end{cases}$$






https://pngtree.com/freepng/question-expression-cartoon-illustration_4545209.ht