Comment y parvenir ?

- Ne perdons pas de vue l’objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)

Comment y parvenir ?

- Ne perdons pas de vue l'objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)
  - Notons ces ou cette stratégie π_*
  - Il peut y en avoir plusieurs !
    - Exemple

Comment y parvenir ?

- Ne perdons pas de vue l'objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)
  - On parle de tâche de « Control »
  - Notons ces ou cette stratégies π_*
  - Il peut y en avoir plusieurs !
    - Exemple
  - Cependant, elle ont toute la même « value function » optimale associée
  - Notons cette dernière v_*

Comment y parvenir ?

- Ne perdons pas de vue l'objectif !
  - Trouver la meilleure des « policies » ! (a.k.a. celle qui maximise en moyenne le reward long terme cumulé)
  - On parle de tâche de « Control »
  - Notons ces ou cette stratégies π_*
  - Il peut y en avoir plusieurs !
    - Exemple
  - Cependant, elle ont toute la même « value function » optimale associée
  - Notons cette dernière v_*
- En effet :
  - $$v_*(s) \doteq \max_\pi v_\pi(s), \tag{3.15}$$

Comment y parvenir ?

- Cependant, elle ont toute la même « value function » optimale associée
  - Notons cette dernière v_*
- En effet :
  - $$v_*(s) \doteq \max_\pi v_\pi(s), \tag{3.15}$$
- Il en va de même pour l' « action-value function » optimale
  - Notons cette dernière q_*
- En effet :
  - $$q_*(s,a) \doteq \max_\pi q_\pi(s,a), \tag{3.16}$$

Comment y parvenir ?

- Si l'on trouve v_* et que l'on connait $p(s', r|s, a)$ alors nous pouvons en déduire une des π_* !

- Si l'on trouve q_* alors nous pouvons en déduire une des π_* !

- Comment trouver v_* ou q_* ?

Comment y parvenir ?

- Partons de deux des équations d'optimalité de Bellman :

$$
\begin{aligned}
v_*(s) &= \max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a) \\
&= \max_a \mathbb{E}_{\pi_*}[G_t \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_t = s, A_t = a] \\
&= \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1}) \mid S_t = s, A_t = a] \\
&= \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma v_*(s')\big].
\end{aligned}
$$

$$
\begin{aligned}
q_*(s,a) &= \mathbb{E}\Big[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \;\Big|\; S_t = s, A_t = a\Big] \\
&= \sum_{s',r} p(s', r \mid s, a)\Big[r + \gamma \max_{a'} q_*(s', a')\Big].
\end{aligned}
$$

Comment y parvenir ?

- Pseudo code pour évaluer puis améliorer en boucle un stratégie a.k.a. « Policy Iteration »:

Policy Iteration (using iterative policy evaluation) for estimating $\pi \approx \pi_*$

1. Initialization
   $V(s) \in \mathbb{R}$ and $\pi(s) \in \mathcal{A}(s)$ arbitrarily for all $s \in \mathcal{S}$

2. Policy Evaluation
   Loop:
       $\Delta \leftarrow 0$
       Loop for each $s \in \mathcal{S}$:
           $v \leftarrow V(s)$
           $V(s) \leftarrow \sum_{s',r} p(s', r \mid s, \pi(s)) \big[r + \gamma V(s') \big]$
           $\Delta \leftarrow \max(\Delta, |v - V(s)|)$
   until $\Delta < \theta$ (a small positive number determining the accuracy of estimation)

3. Policy Improvement
   policy-stable $\leftarrow$ *true*
   For each $s \in \mathcal{S}$:
       old-action $\leftarrow \pi(s)$
       $\pi(s) \leftarrow \operatorname{argmax}_a \sum_{s',r} p(s', r \mid s, a) \big[r + \gamma V(s') \big]$
       If *old-action* $\neq \pi(s)$, then *policy-stable* $\leftarrow$ *false*
   If *policy-stable*, then stop and return $V \approx v_*$ and $\pi \approx \pi_*$; else go to 2

Comment y parvenir ?

- Nous pouvons être plus rapide en itérant directement sur v a.k.a. « Value Iteration »:

Value Iteration, for estimating $\pi \approx \pi_*$

Algorithm parameter: a small threshold $\theta > 0$ determining accuracy of estimation
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

Loop:
| $\Delta \leftarrow 0$
| Loop for each $s \in \mathcal{S}$:
| $\quad v \leftarrow V(s)$
| $\quad V(s) \leftarrow \max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$
| $\quad \Delta \leftarrow \max(\Delta, |v - V(s)|)$
until $\Delta < \theta$

Output a deterministic policy, $\pi \approx \pi_*$, such that
$\quad \pi(s) = \arg\max_a \sum_{s',r} p(s', r \mid s, a)\big[r + \gamma V(s')\big]$

Et si l’on n’a pas de modèle ?

- Première hypothèse :
  - Cas épisodique
- Techniques dites de Monte Carlo
  - a.k.a faisons plein de tests !

Et si l'on n'a pas de modèle ?

- Monte Carlo Prediction :

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

Input: a policy π 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 π: $S_0, A_0, R_1, S_1, A_1, R_2, \ldots, S_{T-1}, A_{T-1}, R_T$

  $G \leftarrow 0$

  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 $Returns(S_t)$

      $V(S_t) \leftarrow \text{average}(Returns(S_t))$

Et si l’on n’a pas de modèle ?

- Mais si nous n’avons pas de modèle, nous préférons obtenir q plutôt que v, pour pouvoir améliorer π !

- Problématique de l’exploration …

- Si nous pouvons démarrer dans un état aléatoire et jouer une action aléatoire alors …

Et si l'on n'a pas de modèle ?

- Mais si nous n'avons pas de modèle, nous préférons obtenir q plutôt que v, pour pouvoir améliorer π !

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 π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  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 $Returns(S_t, A_t)$
      $Q(S_t, A_t) \leftarrow \text{average}(Returns(S_t, A_t))$
      $\pi(S_t) \leftarrow \arg\max_a Q(S_t, a)$

Et si l’on n’a pas de modèle ?

- Mais si nous n’avons pas de modèle, nous préférons obtenir q plutôt que v, pour pouvoir améliorer π !

- Problématique de l’exploration …

- Sinon nous pouvons utiliser une stratégie dite « ε-greedy » …

Et si l'on n'a pas de modèle ?

- Mais si nous n'avons pas de modèle, nous préférons obtenir q plutôt que v, pour pouvoir améliorer π !

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

Algorithm parameter: small $\varepsilon > 0$
Initialize:
  $\pi \leftarrow$ an arbitrary ε-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 π: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  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 $a \in \mathcal{A}(S_t)$:

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

Et si l'on n'a pas de modèle ?

- Mais si nous n'avons pas de modèle, nous préférons obtenir q plutôt que v, pour pouvoir améliorer π !

- Attention, nous n'apprenons plus q_* !

Et si l'on n'a pas de modèle ?

Off-policy MC control, for estimating $\pi \approx \pi_*$

Initialize, for all $s \in \mathcal{S}$, $a \in \mathcal{A}(s)$:
  $Q(s,a) \in \mathbb{R}$ (arbitrarily)
  $C(s,a) \leftarrow 0$
  $\pi(s) \leftarrow \operatorname{argmax}_a Q(s,a)$ (with ties broken consistently)

Loop forever (for each episode):
  $b \leftarrow$ any soft policy
  Generate an episode using b: $S_0, A_0, R_1, \ldots, S_{T-1}, A_{T-1}, R_T$
  $G \leftarrow 0$
  $W \leftarrow 1$
  Loop for each step of episode, $t = T-1, T-2, \ldots, 0$:
    $G \leftarrow \gamma G + R_{t+1}$
    $C(S_t, A_t) \leftarrow C(S_t, A_t) + W$
    $Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \frac{W}{C(S_t, A_t)} [G - Q(S_t, A_t)]$
    $\pi(S_t) \leftarrow \operatorname{argmax}_a Q(S_t, a)$ (with ties broken consistently)
    If $A_t \neq \pi(S_t)$ then exit inner Loop (proceed to next episode)
    $W \leftarrow W \frac{1}{b(A_t|S_t)}$