2.2.3 Cas des transitions et récompenses inconnues

On suppose maintenant que les probabilités de transition et les récompenses sont inconnues.

❒ **Monte-Carlo basé sur modèle** – La méthode de Monte-Carlo basée sur modèle (en anglais *model-based Monte Carlo*) vise à estimer $T(s,a,s')$ et $\text{Reward}(s,a,s')$ en utilisant des simulations de Monte-Carlo avec :

$$\boxed{\widehat{T}(s,a,s') = \frac{\#\text{ de fois où } (s,a,s') \text{ se produit}}{\#\text{ de fois où } (s,a) \text{ se produit}}}$$

and

$$\boxed{\widehat{\text{Reward}}(s,a,s') = r \text{ dans } (s,a,r,s')}$$

Ces estimations sont ensuite utilisées pour trouver les Q-values, ainsi que Q_π et Q_{opt}.

Remarque : la méthode de Monte-Carlo basée sur modèle est dite "hors politique" (en anglais "off-policy") car l'estimation produite ne dépend pas de la politique utilisée.

❒ **Monte-Carlo sans modèle** – La méthode de Monte-Carlo sans modèle (en anglais *model-free Monte Carlo*) vise à directement estimer Q_π de la manière suivante :

$$\boxed{\widehat{Q}_\pi(s,a) = \text{moyenne de } u_t \text{ où } s_{t-1} = s, a_t = a}$$

où u_t désigne l'utilité à partir de l'étape t d'un épisode donné.

Remarque : la méthode de Monte-Carlo sans modèle est dite "sur politique" (en anglais "on-policy") car l'estimation produite dépend de la politique π utilisée pour générer les données.

❒ **Formulation équivalente** – En introduisant la constante $\eta = \frac{1}{1+(\#\text{mises à jour } (s,a))}$ et pour chaque triplet (s,a,u) de la base d'apprentissage, la formule de récurrence de la méthode de Monte-Carlo sans modèle s'écrit à l'aide de la combinaison convexe :

$$\boxed{\widehat{Q}_\pi(s,a) \leftarrow (1-\eta)\widehat{Q}_\pi(s,a) + \eta u}$$

ainsi qu'une formulation mettant en valeur une sorte de gradient :

$$\boxed{\widehat{Q}_\pi(s,a) \leftarrow \widehat{Q}_\pi(s,a) - \eta(\widehat{Q}_\pi(s,a) - u)}$$

❒ **SARSA** – État-action-récompense-état-action (en anglais *state-action-reward-state-action* ou *SARSA*) est une méthode de bootstrap qui estime Q_π en utilisant à la fois des données réelles et estimées dans sa formule de mise à jour. Pour chaque (s,a,r,s',a'), on a :

$$\boxed{\widehat{Q}_\pi(s,a) \longleftarrow (1-\eta)\widehat{Q}_\pi(s,a) + \eta\left[r + \gamma\widehat{Q}_\pi(s',a')\right]}$$

Remarque : l'estimation donnée par SARSA est mise à jour à la volée contrairement à celle donnée par la méthode de Monte-Carlo sans modèle où la mise à jour est uniquement effectuée à la fin de l'épisode.

❒ **Q-learning** – Le Q-apprentissage (en anglais *Q-learning*) est un algorithme hors politique (en anglais *off-policy*) donnant une estimation de Q_{opt}. Pour chaque (s,a,r,s',a'), on a :

$$\boxed{\widehat{Q}_{\text{opt}}(s,a) \leftarrow (1-\eta)\widehat{Q}_{\text{opt}}(s,a) + \eta\left[r + \gamma \max_{a' \in \text{ Actions}(s')} \widehat{Q}_{\text{opt}}(s',a')\right]}$$