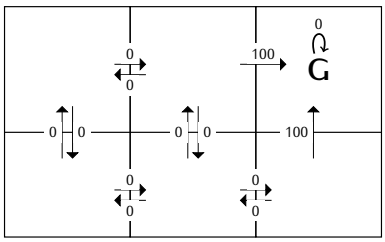
Apprentissage par renforcement

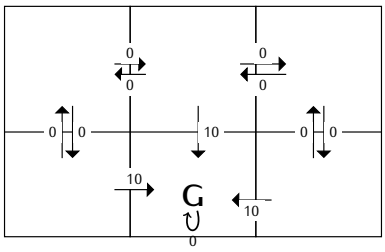
Exercice 1



Fonction de récompenses immédiates r

On utilise un taux de dévaluation $\gamma = 0.9$.
Calculez la fonction $Q(s, a)$.
Calculez la fonction $V^\star$.
Donnez plusieurs stratégies optimales.

Exercice 2



Fonction de récompenses immédiates r

1. On fait de même pour la situation ci-dessus avec $\gamma = 0.8$: donnez $V^\star$ pour chaque état ; $Q(s, a)$ pour chaque transition et une politique optimale.
2. Proposez un changement de la fonction de récompense qui induirait un changement pour la fonction Q mais pas de changement pour la politique optimale
3. Proposez un changement de la fonction de récompense qui induirait un changement pour la fonction Q mais pas de changement pour $V^\star$
4. On suppose que l'agent va utiliser Q-learning en partant de la case en bas à gauche et avec une politique qui lui fait visiter toutes les cases dans le sens des aiguilles d'une montre. Quelles valeurs de $\widehat{Q}$ sont mises à jour durant cet épisode ? Que ce passe-t-il lors d'un second épisode en suivant la même politique ? Lors d'un troisième épisode ?

On rappelle que lorsque l'agent a choisi une action a dans un état s, ce qui amène l'agent dans l'état s', la règle de mise à jour est

$$\widehat{Q}(s, a) \leftarrow r(s, a) + \gamma \max_{a'} \widehat{Q}(s', a')$$