Apprentissage par Renforcement Hors Ligne (BATCH)

Alexandre Gerussi, Léo Pérard, Lucas Seguinot

M2 MOCAD - IIR

15 décembre 2015

Plan de l'exposé

Pourquoi batch ?

- en ligne: interactions libres, voir illimitées avec l'environnement
- pas toujours possible
  - sondages
  - "conduite de vélo": nécessite un opérateur humain
  - ?? : casse du matériel en cas d'échec

Types de batch

- pure batch
- growing batch
- semi-batch

Principes généraux

- utilisation maximale de l'expérience déjà acquise
- experience replay: faire converger sans explorer
- fitting: accélérer et stabiliser la propagation en globalisant les mises à jour

Kernel Based Approximate Dynamic Programming

- ??

Fitted Q-iteration

Itération sur la valeur

$$Q^{i+1}(s,a) = \sum_{s' \in S} \mathcal{T}^a_{ss'}(\mathcal{R}^a_{ss'} + \gamma \max_{a' \in A} Q^i(s',a'))$$

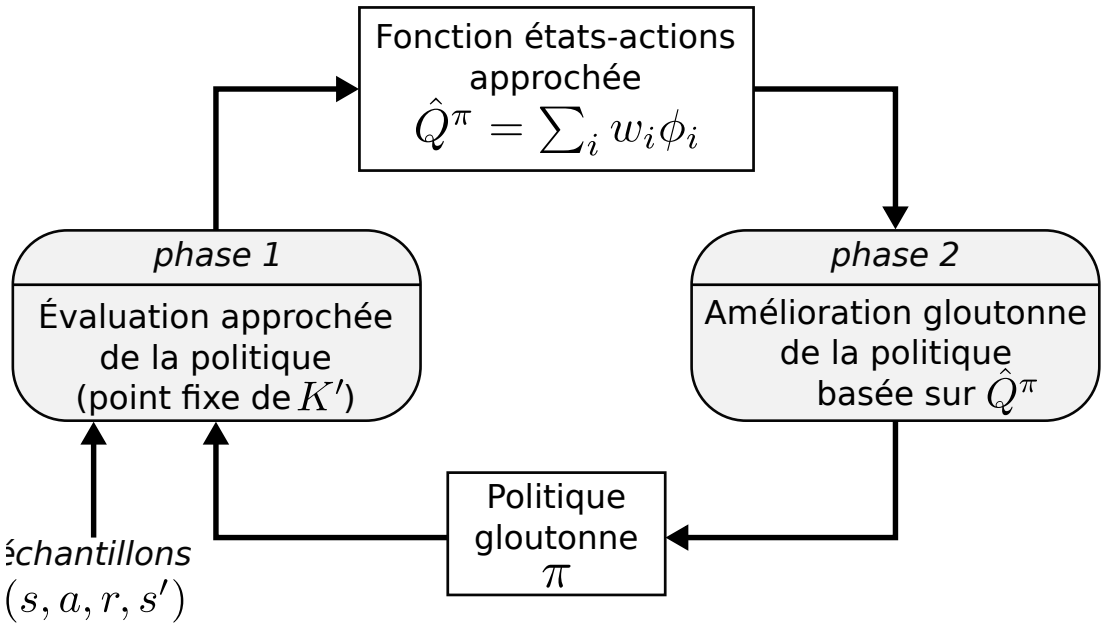
- $\mathcal{F}$ ensemble de transitions (s,a,r,s')
- Base de donnée d'apprentissage P :

$$(s,a) \rightarrow r + \gamma \max_{a' \in A} \hat{Q}^i(s',a')$$

- Apprentissage supervisé $\rightarrow \hat{Q}^{i+1}$

Least-Squares Policy Iteration

$$K'(Q) = \mathcal{P}_{\mathcal{Q}}^{\perp}(K(Q))$$



Équilibre et conduite d'un vélo

- rester debout et atteindre un but en vélo
- valeurs sous contrôle:
  - force rotatoire à appliquer au guidon
  - placement du centre de masse par rapport au vélo
- pure hors-ligne
- dizaine de milliers de trajectoires effectuées aléatoirement
- experience replay: quelques passes de l'ensemble des données font converger