Apprentissage par Renforcement Hors Ligne (BATCH)

Alexandre Gerussi, Léo Pérard, Lucas Seguinot

M2 MOCAD - IIR

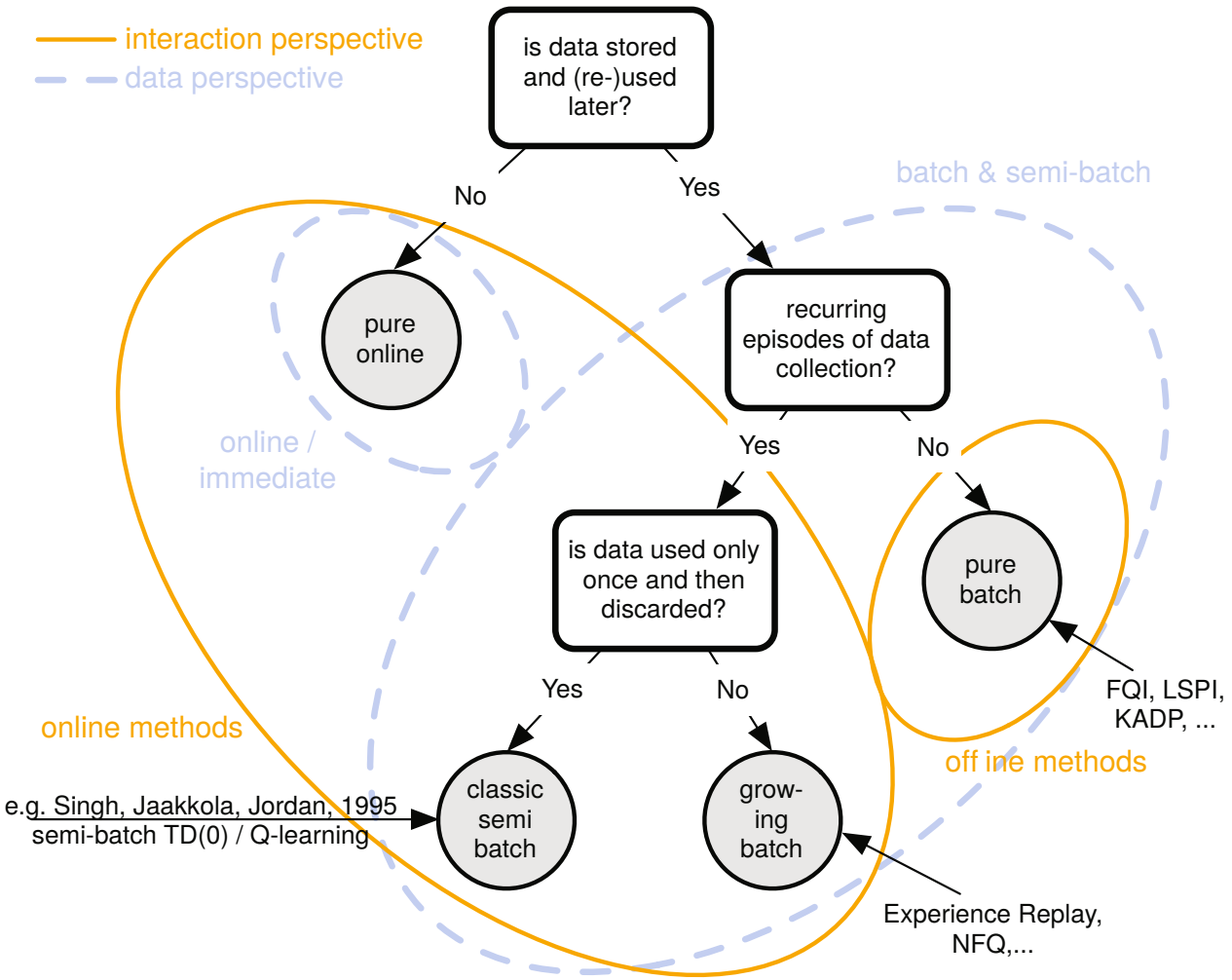
15 décembre 2015

Plan de l'exposé

Pourquoi batch ?

- Algorithmes en ligne : interactions libres avec l'environnement
- Pas toujours possible :
  - Sondages
  - Conduite de vélo, bras robotique... Nécessite un opérateur humain

Types de batch



Principes généraux

- nombreux algorithmes, différant essentiellement sur des détails
- itérations dans SARSA ou Q-learning: exploration + convergence
- $\Rightarrow$ **experience replay**: faire converger sans explorer
- itérations en-ligne: locales, propagation aléatoire par les itérations
- $\Rightarrow$ **fitting**: accélérer et stabiliser la propagation en globalisant les MAJ

Kernel Based Approximate Dynamic Programming

Phase 1

$$\hat{Q}_a^{i+1}(\sigma) = \sum_{(s,a,r,s') \in F_a} k(s,\sigma)[r + \gamma \hat{V}^i(s')]$$

Phase 2

$$\hat{V}^{i+1}(s) = \max_{a \in A} \hat{Q}_a^{i+1}(s)$$

- F_a étant l'ensemble des transitions éffectué via a
- k défini une 'distance' (weighted kernel) entre 2 états

Application: Choix du portfolio optimal

- s_t : valeur de l'action à un instant t
- a_t : action d'investissement

Maximisation

$$W_{t+1} = (1 + a_t \frac{s_{t+1} - s_t}{s_t}) W_t$$

Fitted Q-iteration

Itération sur la valeur

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

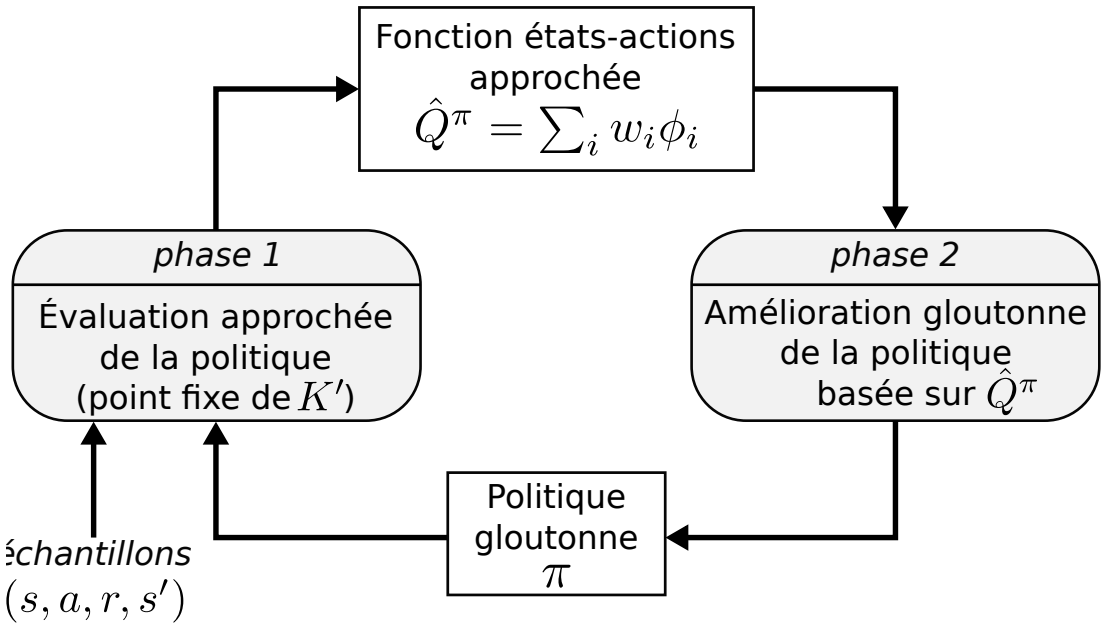
- $\mathcal{F}$ ensemble de transitions (s,a,r,s')
- Base de donnée d'apprentissage P :

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

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

Least-Squares Policy Iteration

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



Application: 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
- pûrement hors-ligne à partir de quelques milliers de trajectoires effectuées aléatoirement
- experience replay: quelques passes de l'ensemble des données font converger