

# Apprentissage par renforcement

## Cours1: MDP et Bandits

Sylvain Lamprier

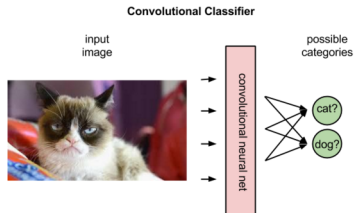
UE RLD - Master DAC

2021

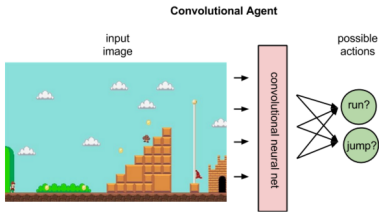
# Apprentissage par renforcement

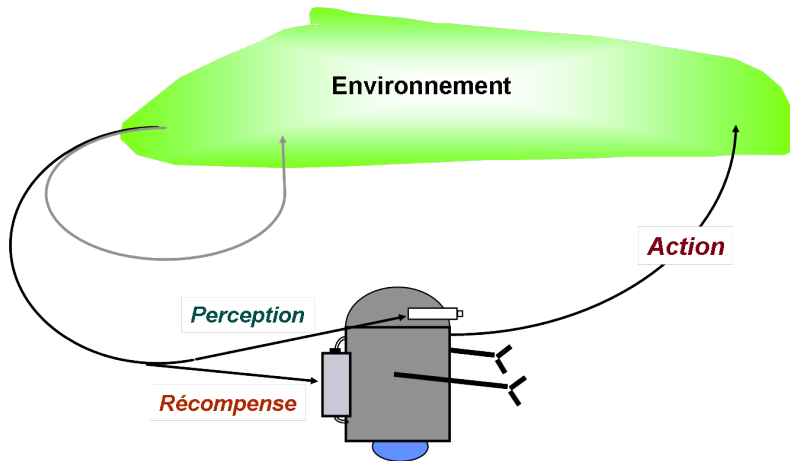
- Apprentissage supervisé
  - On dispose d'une vérité terrain permettant de juger chaque décision
    - ⇒ Minimiser les erreurs par rapport à cette vérité terrain
- Apprentissage non-supervisé
  - On ne dispose que de données non-étiquetées
    - ⇒ Maximiser un critère sur la structure des données
- Apprentissage semi-supervisé
  - On ne dispose que d'une partie des labels
    - ⇒ Mix entre supervisé et non-supervisé
- Apprentissage par renforcement
  - Apprentissage faiblement supervisé : on ne dispose que d'indicateurs de l'utilité des décisions prises
    - ⇒ Maximiser le reward cumulé

## Apprentissage Supervisé (classification) :



## Apprentissage par renforcement (prise de décision) :





Ce contexte décrit des situations très générales et diverses :

- Un robot qui se déplace sur la lune
- Un logiciel qui joue aux échecs (ou au go, au backgammon, au poker,...)
- Un hélicoptère qui apprend à voler
- Un joueur de foot (multi-agent)
- ...

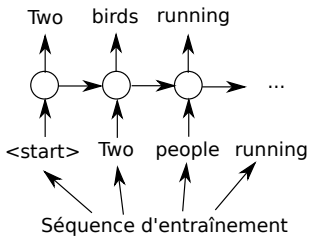
Ce contexte décrit des situations très générales et diverses :

- Un robot qui se déplace sur la lune
- Un logiciel qui joue aux échecs (ou au go, au backgammon, au poker,...)
- Un hélicoptère qui apprend à voler
- Un joueur de foot (multi-agent)
- ...
- Et même pour des tâches de génération séquentielle !  
(e.g., Génération de langue)

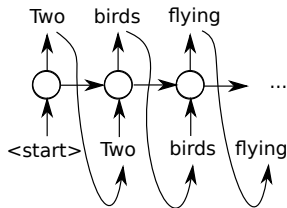
# Apprentissage par renforcement

Génération de séquence textuelle (génération de langue) :

## Apprentissage Supervisé



## Simulation

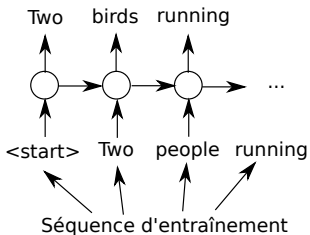


$$\Rightarrow \max_{\theta} \sum_{\omega} \sum_t \log P_{\theta}(\omega_t | \omega_1, \dots, \omega_{t-1})$$

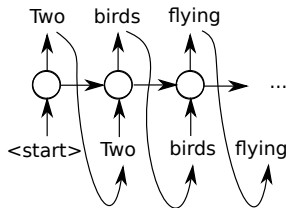
# Apprentissage par renforcement

Génération de séquence textuelle (génération de langue) :

## Apprentissage Supervisé



## Simulation



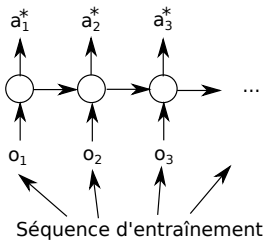
→ Biais d'exposition



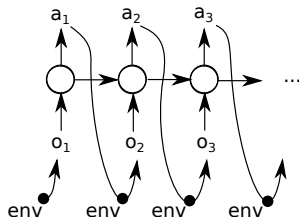
# Apprentissage par renforcement

Génération de séquence d'actions (tâche de contrôle) :

Apprentissage  
Supervisé



Simulation

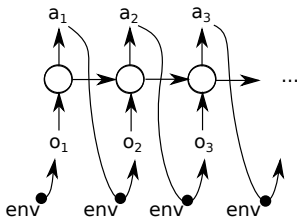


$$\Rightarrow \max_{\theta} \sum_o \sum_t \log P_{\theta}(a_t^* | o_1, \dots, o_t)$$

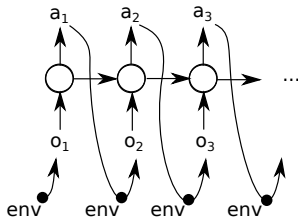
# Apprentissage par renforcement

Génération de séquence d'actions (tâche de contrôle) :

Apprentissage  
par renforcement (RL)



Simulation

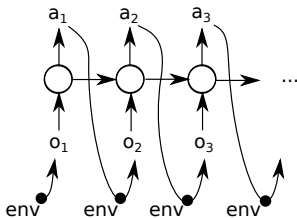


⇒ Apprentissage par expérience

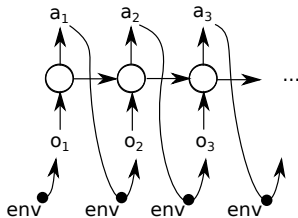
# Apprentissage par renforcement

Génération de séquence d'actions (tâche de contrôle) :

Apprentissage  
par renforcement (RL)



Simulation

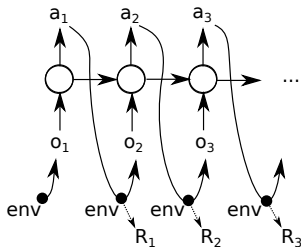


Comment apprendre ? On ne connaît pas les actions optimales  $a^*$ ...

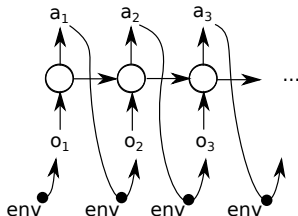
# Apprentissage par renforcement

Génération de séquence d'actions (tâche de contrôle) :

Apprentissage  
par renforcement (RL)



Simulation



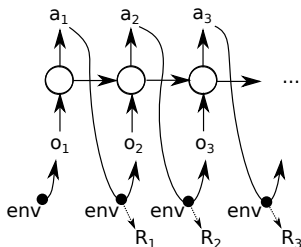
Comment apprendre ? On ne connaît pas les actions optimales  $a^*$ ...

⇒ Récompenses selon la tâche : indicateurs d'utilité sur la séquence

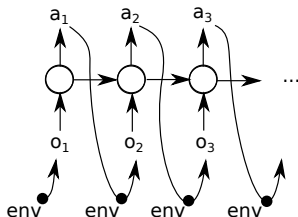
# Apprentissage par renforcement

Génération de séquence d'actions (tâche de contrôle) :

Apprentissage  
par renforcement (RL)



Simulation



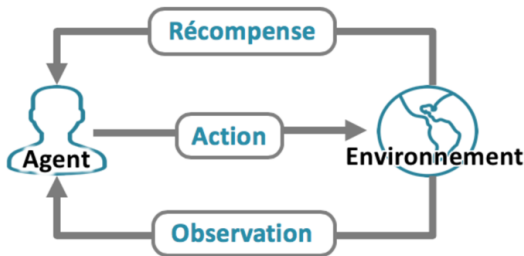
Comment apprendre ? On ne connaît pas les actions optimales  $a^*$ ...

⇒ Récompenses selon la tâche : indicateurs d'utilité sur la séquence

$$\Rightarrow \max_{\pi} \mathbb{E}_{\pi} \sum_t R_t$$

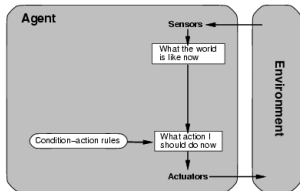
## ⇒ Apprentissage par renforcement

- Tâches : contrôle robotique, problèmes de navigation, conduite autonome, planification, génération textuelle, etc.

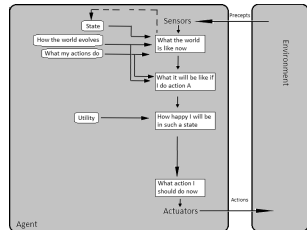


- ⇒ Apprentissage par expérience
- ⇒ Supervision faible : indicateurs d'utilité des décisions prises
  - Problèmes d'exploration
- ⇒ Feedback différé (pas nécessairement à chaque action)
  - Problèmes d'assignation de crédit

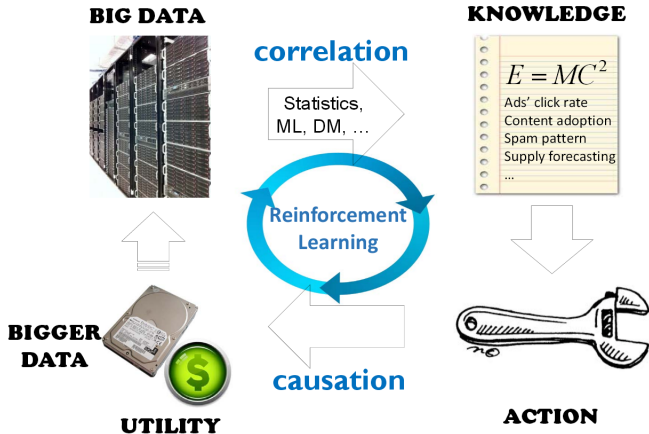
## Agent à réflexes simples



## Agent basé sur un modèle de l'utilité espérée

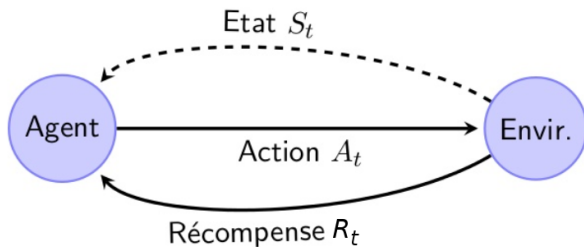


# Apprentissage par renforcement





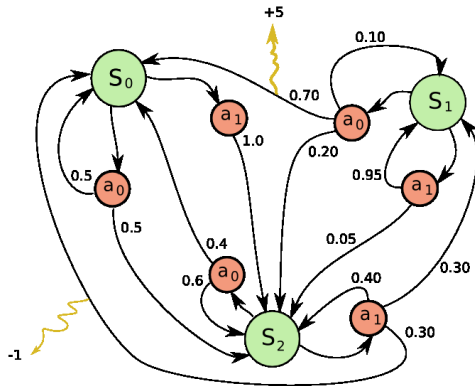
# MDP : Processus de décision de Markov



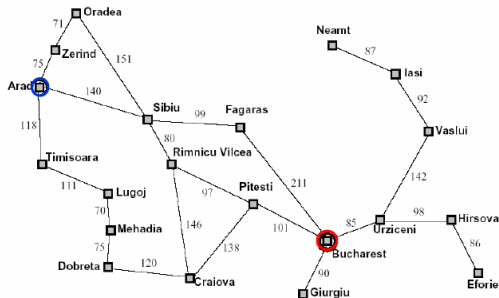
# Markov Decision Process

Nous définissons un MDP comme un quadruplet  $\{\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}\}$  :

- $\mathcal{S}$  est l'ensemble d'états (*states*)
- $\mathcal{A}$  est l'ensemble des actions. On note  $\mathcal{A}(s)$  l'ensemble des actions dans l'état  $s$  tel que  $\mathcal{A}(s) \in \mathcal{A}$
- $\mathcal{P}$  est la fonction de transition :  $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow [0; 1]$ .  
Cette fonction définit une distribution de probabilité sur les transitions  $\mathcal{P}(s, a, s') = P(s' = s_{t+1} | s = s_t, a_t = a)$ .
- $\mathcal{R}$  est la fonction de récompense (*reward*).  
 $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow R$  telle que  
 $R(s, a, s') = E[r_t | s_t = s, a_t = a, s_{t+1} = s']$



# Exemple de problème : Recherche d'itinéraire



## Cas Simple

- Recherche arborescente
- Méthodes Heuristiques
- $A^*$ , etc.

Transitions stochastiques, états non observables, etc.

- RL, POMDP

# Exemple de problème : Taquin

7	2	4
5		6
8	3	1

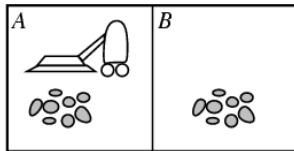
Start State

	1	2
3	4	5
6	7	8

Goal State

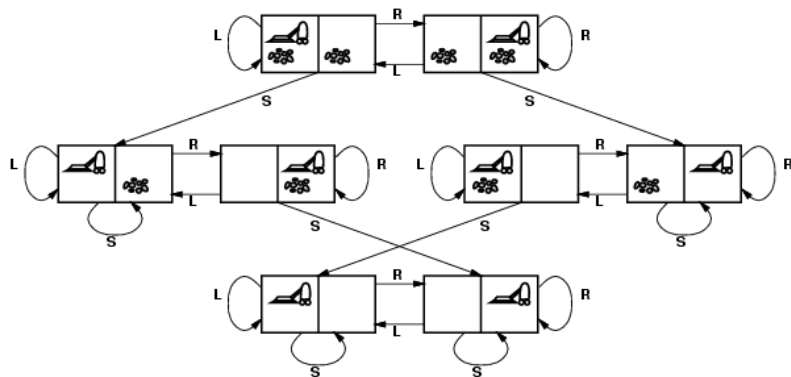
Combien d'états ?

# Exemple de problème : aspirateur

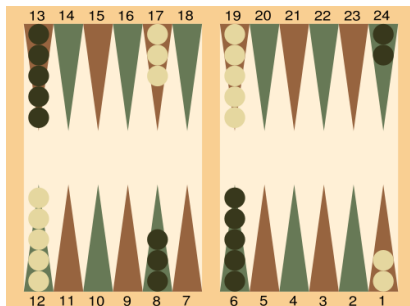


États ? Transitions ?

# Exemple de problème : aspirateur



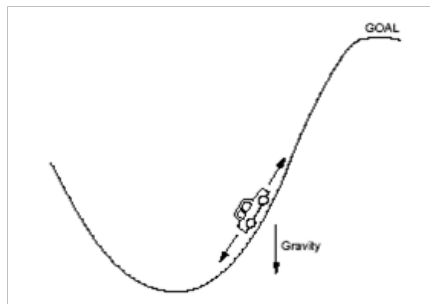
# Exemple de problème : Backgammon



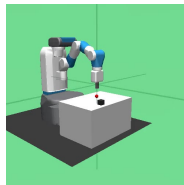
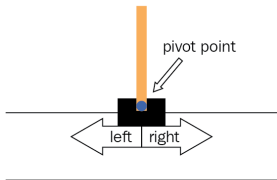
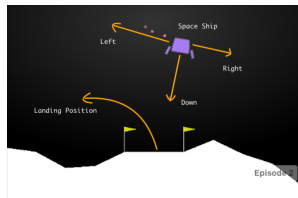
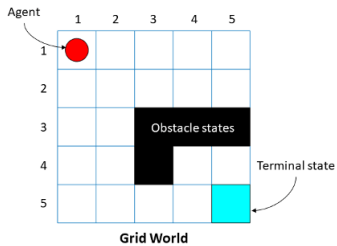
- Etats : état du jeu + état des dés
- Actions : actions de déplacement
- Reward :
  - +100 si gagne
  - -100 si perd
- Entraîné sur 1.5 millions de parties
- Aussi bon que le meilleur joueur du monde



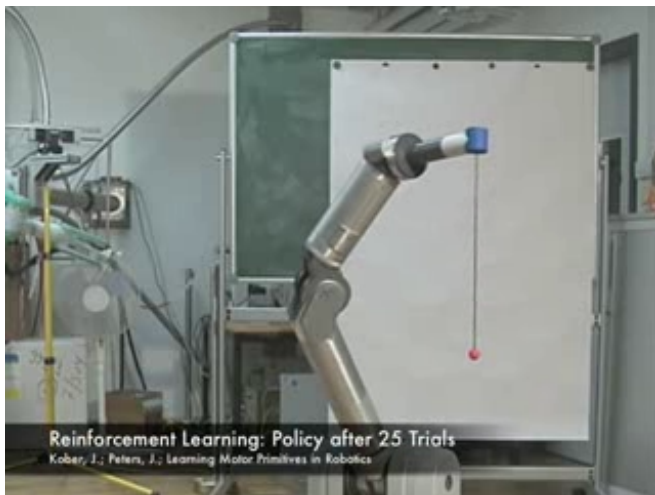
# Exemple de problème : Mountain Car



- Etats : Position de la voiture + vitesse
- Actions : Accélération (avant ou arrière)
- Reward :
  - -1 à chaque "pas"
  - 1 si gagne

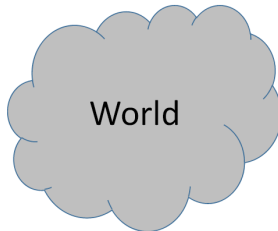


# Exemple de problème : Bilboquet



- Semaine 1 : Bandits
  - $\mathcal{R}$  inconnu a priori
    - ⇒ Exploitation/Exploration
- Semaine 2 : MDPs connus
  - ⇒ Recherche de politiques optimales
- Semaine 3 : MDPs inconnus
  - Ni  $\mathcal{P}$  ni  $\mathcal{R}$  connus a priori
    - ⇒ Exploration de l'espace
    - ⇒ Model-based vs Model-free
    - ⇒ Méthodes Value-Based
- Semaine 4 et 5 : Policy gradients
  - Apprentissage probabiliste de politiques
- Semaine 6 et 7 : Off-Policy Gradients et Actions continues
  - DPG, DDPG, etc.
- Semaine 8 : Curriculum learning & Imitation Learning
  - Apprentissage progressif
  - Imitation Learning : Utilisation de démonstrations humaines
  - Inverse Reinforcement, GANs

# Reinforcement vs Bandits



Actions affect  
world state?

NO

Multi-armed Bandits

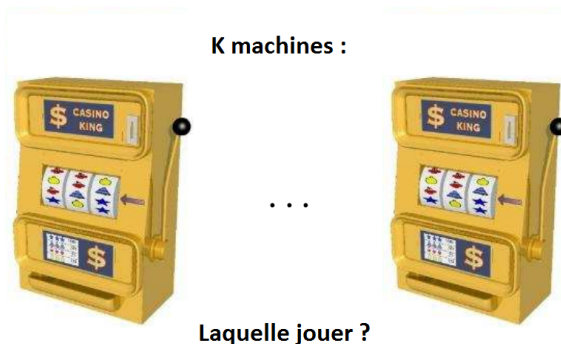
YES

Reinforcement Learning

# Problèmes de bandits

- Prise de décision en ligne
  - Apprentissage en continu
  - Décision Temps réel
  - Pas ou peu d'informations sur les entités manipulées

⇒ Problèmes de bandits-manchots multi-bras



- Problèmes de bandits = Problèmes d'apprentissage par renforcement
  - Reward immédiat
  - Etat courant ne dépend pas des actions passées
- Stochastique vs Adverse
- Stationnaire vs Non-stationnaire
- Bras inter-dépendents ou indépendents
- Prise en compte du contexte décisionnel ?

- Problèmes de bandits = Problèmes d'apprentissage par renforcement
  - Reward immédiat
  - Etat courant ne dépend pas des actions passées
- **Stochastique** vs Adverse
- **Stationnaire** vs Non-stationnaire
- Bras inter-dépendents ou **indépendents**
- Prise en compte du contexte décisionnel ?



- Problèmes de bandits multi-bras

- $K$  actions (bras) possibles à chaque pas de temps  $t$ , une seule effectuée :  $I_t$
- Resultat de l'action  $i$  au temps  $t$  :  $\omega_{i,t} \in \Omega$   
*Seul le resultat du bras joué au temps  $t$  est observé :  $\omega_{I_t,t}$*
- Fonction de reward  $g : \Omega \rightarrow [0; 1]$  définie pour estimer l'utilité du resultat d'une action
- Hypothèse (cas stochastique) : les rewards obtenus pour chaque action sont i.i.d. et suivent une distribution inconnue  $\nu_i$  d'espérance  $\mu_i$
- Une stratégie de décision (ou politique)  $\pi$  détermine, en fonction des actions passées  $I_1 \dots I_{t-1}$ , l'action  $I_t = \pi_t$  à effectuer à l'instant  $t$
- Objectif : Maximiser le reward cumulé sur la période d'actions  $1..T$  :

$$\pi^* = \arg \max_{\pi} \sum_{t=1}^T g(\omega_{\pi_t,t})$$

- Notion centrale de regret :
  - Regret  $\rho_n$  d'avoir effectué les actions  $\pi_1.. \pi_n$  dans les  $n$  premiers pas de temps plutôt que l'action  $i^* = \arg \max_i \mu_i$  de meilleure espérance :

$$\rho_n = \sum_{t=1}^n g(\omega_{i^*,t}) - \sum_{t=1}^n g(\omega_{\pi_t,t})$$

- Espérance de Regret  $\mathbb{E}(\rho_n)$  :

$$\mathbb{E}(\rho_n) = n \times \mu_{i^*} - \mathbb{E}\left(\sum_{t=1}^n \mu_{\pi_t}\right)$$

- Espérance empirique des rewards de  $i$  après  $x$  essais de  $i$  :

$$\hat{\mu}_{i,x} = \frac{1}{x} \sum_{s=1}^x g_{i,s}$$

Avec  $g_{i,s}$  le  $s$ -ième reward obtenu par le bras  $i$ .

- Plus on joue un bras, meilleure est l'estimation de son espérance de reward :

$$\lim_{x \rightarrow \infty} \hat{\mu}_{i,x} = \mu_i$$

- Proposition de politique  $\pi$  :
  - $\pi_t = \arg \max_{i \in K} \hat{\mu}_{i, T_i(t)}$   
avec  $T_i(t)$  le nombre de fois que  $i$  a été joué au temps  $t$
  - Qu'en pensez-vous ?

- Proposition de politique  $\pi$  :

- $\pi_t = \arg \max_{i \in K} \hat{\mu}_{i, T_i(t)}$

avec  $T_i(t)$  le nombre de fois que  $i$  a été joué au temps  $t$

- Qu'en pensez-vous ?

⇒ Pas d'exploration

⇒ Risque de rester "bloqué" sur un bras sous-optimal

- Proposition de politique  $\pi$  :

- $\pi_t = \arg \max_{i \in K} \hat{\mu}_{i, T_i(t)}$

avec  $T_i(t)$  le nombre de fois que  $i$  a été joué au temps  $t$

- Qu'en pensez-vous ?

⇒ Pas d'exploration

⇒ Risque de rester "bloqué" sur un bras sous-optimal

⇒ Définir un compromis entre :

- Exploitation :

- Récupération des gains fournis par le meilleur bras actuel

- Exploration :

- Découverte de nouveaux bras
    - Raffinement de l'estimation de bras  $\neq \arg \max_{i \in K} \hat{\mu}_{i, T_i(t)}$

## Théorème [Lai & Robbins, 1985]

Il est possible de définir des stratégies tel que :

$$\mathbb{E}(\rho_n) \leq cK \ln(n)$$

Avec  $c \approx \frac{1}{\Delta^*}$ , où  $\Delta^* = \mu^* - \max_{j: \mu_j < \mu^*} \mu_j$

- Un premier algo : Epsilon-greedy
  - A chaque itération  $t$  :
    - Avec une probabilité de  $1 - \epsilon_t$ ,  $\pi_t = \arg \max_{i \in K} \hat{\mu}_{i, T_i(t)}$  (bras de meilleure espérance empirique)
    - Avec une probabilité de  $\epsilon_t$ ,  $\pi_t =$  bras choisi au hasard
  - Compromis exploitation-exploration défini par  $\epsilon_t$ 
    - Performances très dépendantes de  $\epsilon_t$
    - $\epsilon_t$  généralement décroissant en fonction de  $t$
- ⇒ De nombreuses variantes existent



- Un premier algo : Epsilon-greedy
- A chaque itération  $t$  :
  - Avec une probabilité de  $1 - \epsilon_t$ ,  $\pi_t = \arg \max_{i \in K} \hat{\mu}_{i, T_i(t)}$  (bras de meilleure espérance empirique)
  - Avec une probabilité de  $\epsilon_t$ ,  $\pi_t =$  bras choisi au hasard
- Compromis exploitation-exploration défini par  $\epsilon_t$ 
  - Performances très dépendantes de  $\epsilon_t$
  - $\epsilon_t$  généralement décroissant en fonction de  $t$

⇒ De nombreuses variantes existent
- Est-il possible de spécifier  $\epsilon_t$  de manière à garantir un regret logarithmique ?

## Tuned Epsilon-greedy

### Théorème [Auer et al., 2002]

Si  $\epsilon_t = \min\{1; \frac{12}{d^2 t}\}$  avec  $d \in ]0; \Delta^*]$ , alors le regret instantané au temps  $n$  de la stratégie epsilon-greedy est dans le pire des cas en  $O(\frac{K}{dn})$

## Tuned Epsilon-greedy

### Théorème [Auer et al., 2002]

Si  $\epsilon_t = \min\{1; \frac{12}{d^2 t}\}$  avec  $d \in ]0; \Delta^*]$ , alors le regret instantané au temps  $n$  de la stratégie epsilon-greedy est dans le pire des cas en  $O(\frac{K}{dn})$

⇒ Si on connaît  $\Delta^*$  alors il est possible de définir une stratégie epsilon-greedy où  $\mathbb{E}(\rho_n) \leq \frac{K}{\Delta^*} \ln(n) + C$  (avec  $C$  une constante)

## Tuned Epsilon-greedy

### Théorème [Auer et al., 2002]

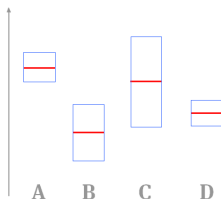
Si  $\epsilon_t = \min\{1; \frac{12}{d^2 t}\}$  avec  $d \in ]0; \Delta^*]$ , alors le regret instantané au temps  $n$  de la stratégie epsilon-greedy est dans le pire des cas en  $O(\frac{K}{dn})$

- ⇒ Si on connaît  $\Delta^*$  alors il est possible de définir une stratégie epsilon-greedy où  $\mathbb{E}(\rho_n) \leq \frac{K}{\Delta^*} \ln(n) + C$  (avec  $C$  une constante)
- Pb :  $\Delta^*$  n'est pas connu a priori ⇒ Définition d'un paramètre  $d$  efficace difficile

- Une stratégie centrale : UCB
  - Upper-Confidence Bound [Auer et al., 2002]

$$\pi_t = \arg \max_i B_{t, T_i(t-1)}(i), \text{ avec } B_{t,s}(i) = \hat{\mu}_{i,s} + \sqrt{\frac{2 \log t}{s}}$$

- Stratégie optimiste :
  - ⇒  $B_{t, T_i(t-1)}(i)$  représente une borne supérieure de  $\hat{\mu}_{i, T_i(t-1)}$  à l'iteration  $t$



- ⇒ On choisit le bras qui serait le meilleur si les valeurs des bras étaient les meilleures possibles selon l'intervale de confiance

- Stratégie optimiste :

- Inégalités de Chernoff-Hoeffding pour des variables aléatoires indépendantes  $X_i \in [0, 1]$  d'espérance  $\mu$  :

$$P\left(\frac{1}{s} \sum_{i=1}^s X_i - \mu \geq \epsilon\right) \leq \exp^{-2s\epsilon^2} \quad \text{et} \quad P\left(\frac{1}{s} \sum_{i=1}^s X_i - \mu \leq -\epsilon\right) \leq \exp^{-2s\epsilon^2}$$

- On a alors pour tout bras  $i$  :

$$P(\hat{\mu}_{i, T_i(t-1)} + \sqrt{\frac{2 \log t}{T_i(t-1)}} \leq \mu_i) \leq t^{-4} \quad \text{et} \quad P(\hat{\mu}_{i, T_i(t-1)} - \sqrt{\frac{2 \log t}{T_i(t-1)}} \geq \mu_i) \leq t^{-4}$$

⇒ Cela définit un intervalle de confiance de niveau  $1 - 2t^{-4}$  :

$$\mu_i - \sqrt{\frac{2 \log t}{T_i(t-1)}} \stackrel{(a)}{\leq} \hat{\mu}_{i, T_i(t-1)} \stackrel{(b)}{\leq} \mu_i + \sqrt{\frac{2 \log t}{T_i(t-1)}}$$

- UCB choisit un bras sous-optimal  $i$ , i.e.  $B_{t,s}(i) \geq B_{t,s^*}(i^*)$ , si :

$$\hat{\mu}_{i, T_i(t-1)} + \sqrt{\frac{2 \log t}{T_i(t-1)}} \geq \hat{\mu}_{i^*, T_{i^*}(t-1)} + \sqrt{\frac{2 \log t}{T_{i^*}(t-1)}}$$

- Si on est dans l'intervale de confiance, on a alors dans ce cas :

$$\mu_i + 2\sqrt{\frac{2 \log t}{T_i(t-1)}} \geq \mu^*, \text{ soit : } T_i(t-1) \leq \frac{8 \log t}{\Delta_i^2}$$

$$\text{car } \hat{\mu}_{i^*, T_{i^*}(t-1)} \geq \mu^* - \sqrt{\frac{2 \log t}{T_{i^*}(t-1)}} \text{ (inégalité (a))}$$

$$\text{et } \hat{\mu}_{i, T_i(t-1)} \leq \mu_i + \sqrt{\frac{2 \log t}{T_i(t-1)}} \text{ (inégalité (b))}$$

- Sinon, c'est que l'une des inégalités (a) ou (b) n'est pas vérifiée

- On pose, pour tout entier  $u \geq 0$  :

$$T_i(n) \leq u + \sum_{t=u+1}^n \mathbb{I}(\exists s : u < s \leq t, \exists s^* : 1 \leq s^* \leq t, B_{t,s}(i) \geq B_{t,s^*}(i^*))$$

- En choisissant  $u = \frac{8 \log n}{\Delta_i^2}$ , on sait alors qu'un bras sous-optimal est choisi seulement si (a) ou (b) n'est pas vérifiée. Or :
  - (a) n'est pas vérifiée avec une proba de  $t^{-4}$
  - (b) n'est pas vérifiée avec une proba de  $t^{-4}$

- Donc :

$$\begin{aligned} \mathbb{E} T_i(n) &\leq \frac{8 \log n}{\Delta_i^2} + \sum_{t=u+1}^n \left[ \sum_{s=u+1}^t t^{-4} + \sum_{s=1}^t t^{-4} \right] \\ &\leq \frac{8 \log n}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \end{aligned}$$



# UCB : Borne supérieure du regret

- On pose, pour tout entier  $u \geq 0$  :

$$T_i(n) \leq u + \sum_{t=u+1}^n \mathbb{I}(\exists s : u < s \leq t, \exists s^* : 1 \leq s^* \leq t, B_{t,s}(i) \geq B_{t,s^*}(i^*))$$

- En choisissant  $u = \frac{8 \log n}{\Delta_i^2}$ , on sait alors qu'un bras sous-optimal est choisi seulement si (a) ou (b) n'est pas vérifiée. Or :
  - (a) n'est pas vérifiée avec une proba de  $t^{-4}$
  - (b) n'est pas vérifiée avec une proba de  $t^{-4}$

- Donc :

$$\begin{aligned} \mathbb{E} T_i(n) &\leq \frac{8 \log n}{\Delta_i^2} + \sum_{t=u+1}^n \left[ \sum_{s=u+1}^t t^{-4} + \sum_{s=1}^t t^{-4} \right] \\ &\leq \frac{8 \log n}{\Delta_i^2} + 1 + \frac{\pi^2}{3} \end{aligned}$$

⇒ Borne supérieure logarithmique sur l'espérance du nombre de tirages de chaque bras sous-optimal

- Borne du regret à partir de cette borne du nombre de tirages des bras sous-optimaux ?

$$\mathbb{E}(\rho_n)$$

$$= n \times \mu^* - \mathbb{E}\left(\sum_{t=1}^n \mu_{\pi_t}\right)$$

$$= n \times \mu^* - \sum_{i=1}^K \mathbb{E}T_i(n) \times \mu_i$$

$$= \sum_{i=1}^K \mathbb{E}T_i(n) \times (\mu^* - \mu_i)$$

$$= \sum_{i=1}^K \mathbb{E}T_i(n) \times \Delta_i$$

$$\leq \sum_{i \in \{1..K\} : \mu_i < \mu^*} \frac{8 \log n}{\Delta_i} + \Delta_i \left(1 + \frac{\pi^2}{3}\right)$$

$$\leq K \frac{8 \log n}{\Delta^*} + K \Delta^* \left(1 + \frac{\pi^2}{3}\right)$$

# \* UCB : Application à la publicité sur le Web

Exemple d'application d'UCB sur le Web : la publicité dans les moteurs de recherche [*Pandey&Olston, 2007*]

- Publicités  $A_1..A_k$
- Requêtes (ou mots)  $Q_1..Q_m$
- Revenu par clic  $a_{i,j}$  pour chaque paire publicité  $A_i$ -requête  $Q_j$
- Probabilité (inconnue)  $p_{i,j}$  que les utilisateurs cliquent sur la publicité  $A_i$  pour la requête  $Q_j$

⇒ Objectif : Maximiser les gains du moteur sur l'ensemble des  $n_j$  recherches selon chaque requête  $Q_j$  de la journée :

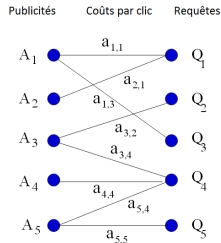
$$\sum_{i=1}^{n_j} \mathbb{I}(\text{clic sur la publicité } A_i \text{ affichée}) \times a_{i,j} \sim \sum_{i=1}^{n_j} p_{i,j} \times a_{i,j}$$

⇒ Choix de la publicité à afficher pour la  $i$ -ième recherche utilisant la requête  $Q_j$  :

$$A_i = \arg \max_{A_x \in A} (\hat{p}_{x,j}(i-1) + \sqrt{\frac{2 \log(i)}{n_{x,j}(i-1)}}) \times a_{x,j}$$

Avec sur les  $i-1$  premières recherches concernant la requête  $Q_j$  :

- $\hat{p}_{x,j}(i-1)$  : l'estimation de la probabilité de clic sur la pub  $A_x$
- $n_{x,j}(i-1)$  : le nombre de fois où  $A_x$  a été affiché



## \* UCB : Application à la collecte de données

Collecte de données temps réel sur les réseaux sociaux  
[Gisselbrecht et al., 2015]

- Plateformes de streaming des réseaux
- Ecoute d'un nombre limité d'utilisateurs en simultané
- Pb : choisir les k utilisateurs avec le meilleur potentiel d'utilité selon la fonction de reward considérée :

$$\pi^* = \arg \max_{\pi} \sum_{t=1}^n \sum_{i \in \pi_t} g(\omega_{i,t})$$

⇒ UCBV appliqué à la sélection de k bras simultanés  
(Combinatorial UCBV)

# \*Problèmes de bandits : une variante d'UCB

- UCB-V [Audibert et al., 2007]

- Intuition

- Certains bras ont une variabilité des rewards plus importante que d'autres
    - Estimation des bras à plus grande variabilité plus difficile
  - ⇒ Meilleure prise en compte de ces bras par considération de la variance empirique des rewards

- Variance Empirique :

$$\hat{\sigma}_{i,x}^2 = \frac{1}{x} \sum_{s=1}^x (g_{i,s} - \hat{\mu}_{i,x})^2$$

- UCB-V = UCB avec borne supérieure de l'intervalle de confiance de la variance

$$\pi_t = \arg \max_i B_{t, T_i(t-1)}(i)$$

Avec

$$B_{t,s}(i) = \hat{\mu}_{i,s} + \sqrt{\frac{2 \log(t) \hat{\sigma}_{i,s}^2}{s}} + \frac{\log(t)}{2s}$$

# Problèmes de bandits : contexte de décision

- Contexte de décision
  - Contexte global variant à chaque itération
  - Contexte individuel (sur chaque bras) fixe (= profils des bras)
  - Contexte individuel variant à chaque itération
- Prise en compte du contexte
  - Contexte fixe (prise en compte globale)
    - ⇒ Accélérer la sélection des meilleurs bras en apprenant des "zones" de l'espace de représentation pertinentes
    - ⇒ Cold-start pour nouveaux bras entrant dans le pool
  - Contexte variable : Hypothèse de non-stationnarité des rewards
    - ⇒ Prise en compte globale de contextes individuels : rewards des bras suivent une distribution commune définie sur leurs contextes individuels
    - ⇒ Prise en compte individuelle d'un contexte global : chaque bras suit une distribution indépendante conditionnellement au contexte global de la décision
    - ⇒ Prise en compte individuelle d'un contexte individuel : rewards de chaque bras dépendent de son état actuel

- Lin-UCB [*Li et al., 2010*]

- UCB avec prise en compte individuelle du contexte
- Contexte de décision pour un bras  $i$  à l'instant  $t$  :  $x_{i,t}$
- Recherche pour chaque bras des corrélations entre contextes de décision et rewards obtenus :

$$\mathbb{E}_i(g(\omega_{i,t})|x_{i,t}) = \langle x_{i,t}, \theta_i^* \rangle$$

- Mise à jour des paramètres par Ridge Regression au fur et à mesure du processus

$$\hat{\theta}_i = \arg \min_{\theta_i} \|D_i \theta_i - c_i\|^2 + \|\theta_i\|^2$$

Avec  $D_i$  la matrice des contextes observés pour le bras  $i$  et  $c_i$  le vecteur des rewards obtenus correspondants

$$\Rightarrow \hat{\theta}_i = (D_i^T D_i + I)^{-1} D_i^T c_i$$

# Bandits contextuels

- Lin-UCB [Li et al., 2010]

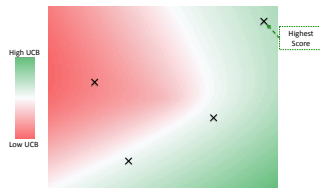
- Il peut être montré qu'avec une probabilité  $1 - \delta$  :

$$| \langle x_{i,t}, \hat{\theta}_i \rangle - \mathbb{E}_i(g(\omega_{i,t}) | x_{i,t}) | \leq \alpha \sqrt{x_{i,t}^T (D_i^T D_i + I)^{-1} x_{i,t}}$$

Avec  $\alpha = 1 + \sqrt{\log(2/\delta)/2}$

- On a donc une borne supérieure de l'intervalle de confiance pour  $\langle x_{i,t}, \hat{\theta}_i \rangle$ , qu'on peut donc utiliser à la manière d'UCB pour définir la politique  $\pi$  :

$$\pi_t = \arg \max_i \langle x_{i,t}, \hat{\theta}_i \rangle + \alpha \sqrt{x_{i,t}^T (D_i^T D_i + I)^{-1} x_{i,t}}$$





---

**Algorithm 1** LinUCB with disjoint linear models.

---

```
0: Inputs:  $\alpha \in \mathbb{R}_+$ 
1: for  $t = 1, 2, 3, \dots, T$  do
2:   Observe features of all arms  $a \in \mathcal{A}_t$ :  $\mathbf{x}_{t,a} \in \mathbb{R}^d$ 
3:   for all  $a \in \mathcal{A}_t$  do
4:     if  $a$  is new then
5:        $\mathbf{A}_a \leftarrow \mathbf{I}_d$  ( $d$ -dimensional identity matrix)
6:        $\mathbf{b}_a \leftarrow \mathbf{0}_{d \times 1}$  ( $d$ -dimensional zero vector)
7:     end if
8:      $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1} \mathbf{b}_a$ 
9:      $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha \sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$ 
10:   end for
11:   Choose arm  $a_t = \arg \max_{a \in \mathcal{A}_t} p_{t,a}$  with ties broken arbitrarily, and observe a real-valued payoff  $r_t$ 
12:    $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t} \mathbf{x}_{t,a_t}^\top$ 
13:    $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$ 
14: end for
```

---

$$\text{Avec } A_i = D_i^T D_i + I \text{ et } b_i = D_i^T c_i$$

# \* Bandits contextuels

Application à la recommandation de news personnalisée [Li et al., 2010]

[www.yahoo.com](http://www.yahoo.com)



$A_t$  : available articles at time  $t$

$\mathbf{x}_t$  : user features (age, gender, interests, ...)

$a_t$  : the displayed article at time  $t$

$r_{t,a_t}$  : 1 for click, 0 for no-click

Average reward is click-through rate (CTR)

# Bandits contextuels

- Alternative aux stratégies optimistes : Thompson Sampling [*Thompson, 1933*],[*Kaufmann et al., 2012*]
- Maximisation d'espérance par marginalisation selon les paramètres  $\theta$  d'un modèle  $h_\theta$  (tel que  $h_\theta(x_{i,t}) = E[r_{t,i}|x_{i,t}, \theta]$ ) :

$$\pi_t = \arg \max_i \mathbb{E}(r_{t,i}|x_{i,t}, \mathcal{D}_t) \quad (1)$$

$$= \arg \max_i \int \mathbb{E}(r_{t,i}|x_{i,t}, \theta) P(\theta|\mathcal{D}_t) d\theta \quad (2)$$

Avec :

- $r_{t,i} = g(\omega_{i,\pi_t})$  le reward obtenu par  $i$  au temps  $t$
- $\mathcal{D}_t = \{(i, t, x_{i,t}, r_t)\}$  l'ensemble des observations passées ;
- $\mathbb{E}(r_{t,i}|x_{i,t}, \mathcal{D})$  l'espérance du reward obtenu par  $i$  au temps  $t$  en fonction des observations passées
- $P(\theta|\mathcal{D}_t) \propto P(\mathcal{D}_t|\theta)P(\theta)$  la probabilité postérieure des paramètres conditionnellement aux paramètres
- $P(\mathcal{D}_t|\theta)$  la vraisemblance des observations selon les paramètres
- $P(\theta)$  un prior sur l'ensemble de paramètres  $\theta$  ;

- Thompson Sampling en pratique
- A chaque iteration  $t$  :
  - 1 Échantillonnage des paramètres  $\theta^* \sim P(\theta|\mathcal{D}_t)$
  - 2 Choix du bras qui maximise l'espérance du reward en fonction des paramètres et du contexte :

$$\pi_t = \arg \max_i \mathbb{E}(r|x_{i,t}, \theta^*)$$

- Thompson Sampling en pratique
  - Cas linéaire [Agrawal & Goyal, 2013] :
    - $\mathbb{E}(r_t|x_{i,t}, \theta) = \langle \theta, x_{i,t} \rangle$
    - On suppose que les rewards observés suivent une loi normale :  $P(\mathcal{D}_t|\theta) = \mathcal{N}(\theta \mathcal{X}_t, v^2)$ , avec  $\mathcal{X}_t$  la matrice des contextes des bras choisis jusqu'à  $t$
    - On suppose un prior gaussien sur les paramètres  $\theta$  :  $P(\theta) = \mathcal{N}(0, \sigma^2)$
- ⇒ Postérieure normale :  $P(\theta|\mathcal{D}_t) = \mathcal{N}(b_t A_t^{-1}, A_t^{-1})$ , avec :
- $A(t) = \frac{I}{\sigma^2} + \frac{1}{v^2} \sum_{s=1}^{t-1} x_{\pi_i, s} x_{\pi_i, s}^T$
  - $b(t) = \frac{1}{v^2} \sum_{s=1}^{t-1} r_s x_{\pi_i, s}$

## \* Bandits contextuels

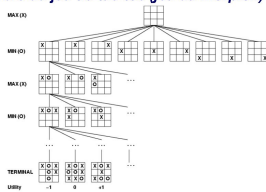
- Thompson Sampling : Application a la selection de messages à publier
  - ⇒ Maximiser le nombre de *retweets* [Lage et al., 2013]
- A chaque ieration  $t$  :
  - 1 Recuperation de la liste des articles candidats au temps  $t$
  - 2 Publication de l'article avec le plus fort potentiel selon ses caractéristiques et les paramètres du modèle
  - 3 Observation de l'impact de la publication pendant une periode de temps donnée
  - 4 Mise à jour du modèle selon le nombre de *retweets* observés
- Caractéristiques considérées :
  - Contenu : tf normalisé des termes
  - Nombre d'Hashtags
  - Nombre de destinataires
  - Taille du message

# Bandits : Application dans les arbres de décision

Un processus de décision peut être représenté par un arbre :

- Noeud = situation dans un problème de décision
- Fils d'un noeud  $s_i$  = situation atteignable à partir de  $s_i$  selon une action
- Racine = situation de départ
- Feuilles = situations terminales
- Possiblement infini (si boucles dans le MDP correspondant ou si actions/etats continus)

Arbre du jeu tic-tac-toe (jeu du morpion)



IFT615

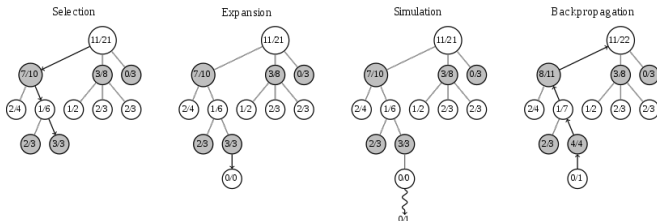
© Frédéric Kuhn

Monde stochastique et/ou avec adversaire : comment évaluer les bonnes situations ?

- Dans le cadre des jeux à deux joueurs : algorithmes minmax, alpha-beta, etc.
- ... Mais souvent arbre trop grand  $\Rightarrow$  impossible à évaluer complètement
- Possibilité : s'arrêter à une profondeur donnée et retourner une estimation de la qualité de la situation à cette profondeur
- Problème : comment estimer la qualité d'une situation ?

# Bandits : Application dans les arbres de décision

## Monte-Carlo Tree Search



- Sélection : A chaque niveau de l'arbre connu, sélection d'une action selon une stratégie exploitation/exploration
- Expansion : Arrivé sur une feuille, création des situation filles du noeud
- Simulation : Jeu aléatoire jusqu'à situation de victoire ou de défaite (Rollout)
- Backpropagation : Mise à jour des scores des noeuds de l'arbre selon victoire ou défaite

Stratégie de Sélection classique : UCT (upper-confidence Tree)

$$\frac{w_i}{n_i} + c\sqrt{\frac{\ln N_i}{n_i}}$$



# References

- [Agrawal & Goyal, 2013] S. Agrawal and N. Goyal. Thompson sampling for contextual bandits with linear payoffs. In ICML (3), pages 127–135, 2013
- [Audibert et al., 2007] J.-Y. Audibert, R. Munos, and C. Szepesvari. Tuning bandit algorithms in stochastic environments. In ALT'07, pages 150–165. 2007.
- [Auer et al., 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem. Mach. Learn. 47, 2-3 (May 2002), 235-256.
- [Gisselbrecht et al., 2015] Thibault Gisselbrecht, Ludovic Denoyer, Patrick Gallinari and Sylvain Lamprier. WhichStreams : A Dynamic Approach for Focused Data Capture from Large Social Media. ICWSM 2015 : 130-139
- [Kaufmann et al., 2012] E.Kaufmann, N.Korda, and R.Munos. Thompson Sampling : an asymptotically optimal finite-time analysis. In ALT'12.
- [Lage et al., 2013] Ricardo Lage, Ludovic Denoyer, Patrick Gallinari et al. (2013) Choosing which message to publish on social networks : A Contextual bandit approach. In IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.
- [Lai & Robbins, 1985] Lai, T. and Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. Advances in Applied Mathematics ,6,4–22.
- [Li et al., 2010] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. 2010. A contextual-bandit approach to personalized news article recommendation. In Proceedings of the 19th international conference on World wide web (WWW '10). ACM, New York, NY, USA, 661-670.
- [Pandey & Olston, 2007] Sandeep Pandey and Christopher Olston. Handling advertisements of unknown quality in search advertising. Advances in Neural Information Processing Systems , 19 :1065, 2007
- [Thompson, 1933] Thompson, William R. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples". Biometrika, 25(3-4) :285–294, 1933.