Fouilles de données et Medias sociaux

Master 2 DAC - FDMS

Sylvain Lamprier

UPMC

Décision sur les réseaux

- Décision sur les réseaux
  - Très large volume de données
  - Contexte dynamique
  - Contraintes temps réel
  - Parfois peu d'informations sur chaque entité du réseau
- Exemples de tâches
  - Publicité en ligne
  - Recommendation personnalisée de produits
  - Crawling / Collecte de données
  - Maximisation d'audimat
  - ...
- Modèles traditionnels parfois peu adaptés
  - Apprentissage coûteux
  - Besoin de grandes quantités d'informations pour être performants
  - Peu flexibles aux évolutions sur le réseau
  - Pas de notion de "vitesse d'apprentissage"

⇒ Modèles pour la prise de décision temps réel sur les réseaux

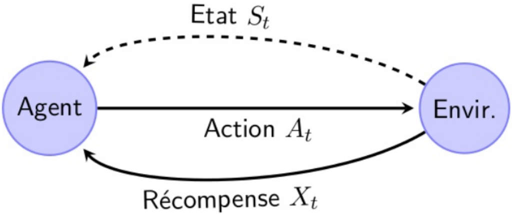
Problèmes de bandits

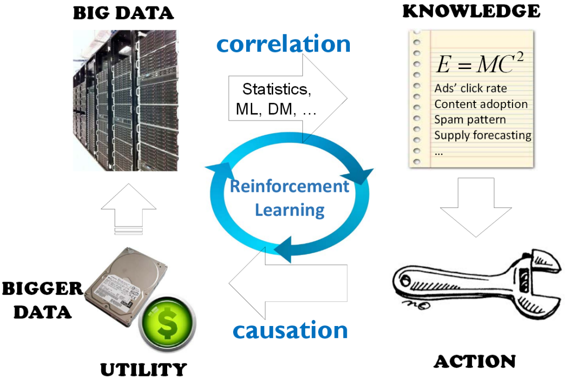
- Prise de décision sur les réseaux
  - Apprentissage en continu
  - Décision Temps réel
  - Pas ou peu d'informations sur les entités manipulées

$\Rightarrow$ Problèmes de bandits-manchots multi-bras

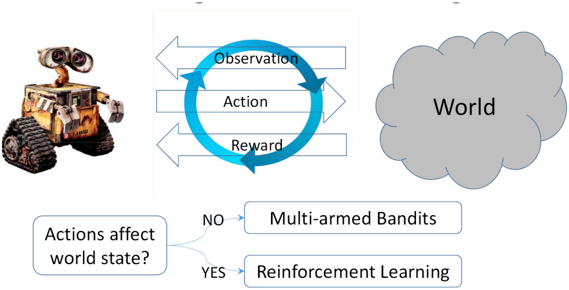
Apprentissage par renforcement

- Apprentissage supervisé
  - On dispose d'une vérité terrain permettant de juger chaque décision
  - $\Rightarrow$ Minimiser les erreurs par rapport à cette vérité terrain
- Apprentissage par renforcement
  - Processus de decision markovien (MDP)
  - Apprentissage faiblement supervisé : on ne dispose que d'indicateurs de l'utilité des décisions prises
  - $\Rightarrow$ Maximiser le reward cumulé




BIG DATA
KNOWLEDGE
correlation
Statistics, ML, DM, …
$E = MC^2$
Ads' click rate
Content adoption
Spam pattern
Supply forecasting
…
Reinforcement Learning
BIGGER DATA
UTILITY
causation
ACTION

Apprentissage par renforcement



Problèmes de bandits

- 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 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 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 $g : \Omega \to [0;1]$ 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 ν_i d'espérance μ_i
  - Une stratégie de décision (ou politique) π 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 ρ_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 $\mathbb{E}(\rho_n)$:

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

Problèmes de bandits

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

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

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:

$$\lim_{x\to\infty} \widehat{\mu}_{i,x} = \mu_i$$

Problèmes de bandits

- Proposition de politique π:
  - $\pi_t = \arg\max_{i \in K} \widehat{\mu}_{i,T_i(t)}$
    avec $T_i(t)$ le nombre de fois que i a été joué au temps t

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

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

$\Rightarrow$ Pas d'exploration

Problèmes de bandits

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

$\Rightarrow$ Pas d'exploration
$\Rightarrow$ Risque de rester "bloqué" sur un bras sous-optimal

Problèmes de bandits

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

$\Rightarrow$ Pas d'exploration
$\Rightarrow$ Risque de rester "bloqué" sur un bras sous-optimal

$\Rightarrow$ 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 $\neq \arg\max_{i \in K} \widehat{\mu}_{i,T_i(t)}$

Problèmes de bandits

Théorème [Lai & Robbins, 1985]

Il est possible de définir des stratégies tel que:

$$\mathbb{E}(\rho_n) \leq cKln(n)$$

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

Problèmes de bandits

- Un premier algo : Epsilon-greedy
- A chaque itération t:
  - Avec une probabilité de $1 - \epsilon_t$, $\pi_t = \arg\max_{i \in K} \widehat{\mu}_{i,T_i(t)}$ (bras de meilleure espérance empirique)
  - Avec une probabilité de ϵ_t, $\pi_t =$ bras choisi au hasard

- Compromis exploitation-exploration défini par ϵ_t
  - Performances très dépendantes de ϵ_t
  - ϵ_t généralement décroissant en fonction de t

  $\Rightarrow$ De nombreuses variantes existent

Problèmes de bandits

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

  $\Rightarrow$ De nombreuses variantes existent
- Est-il possible de spécifier ϵ_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 $O(\frac{K}{dn})$

Problèmes de bandits

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 $O(\frac{K}{dn})$

$\Rightarrow$ Si on connaît Δ^* 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)

Problèmes de bandits

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 $O(\frac{K}{dn})$

$\Rightarrow$ Si on connaît Δ^* 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)

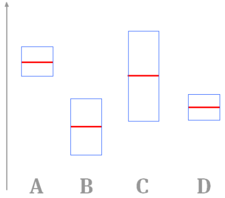
- Pb: Δ^* n'est pas connu a priori $\Rightarrow$ Définition d'un paramètre d efficace difficile

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

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

  - Stratégie optimiste :

    $\Rightarrow$ $B_{t,T_i(t-1)}(i)$ représente une borne supérieure de $\widehat{\mu}_{i,T_i(t-1)}$ à l'iteration t



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

UCB : Borne supérieure du regret

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

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

  - On a alors pour tout bras i :

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

$\Rightarrow$ Cela définit un intervale de confiance de niveau $1 - t^{-4}$:

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

UCB : Borne supérieure du regret

- UCB choisit un bras sous-optimal i, i.e. $B_{t,s}(i) \geq B_{t,s^*}(i^*)$), si:
$$\widehat{\mu}_{i,T_i(t-1)} + \sqrt{\frac{2\log t}{T_i(t-1)}} \geq \widehat{\mu}_{i^*,T_{i^*}(t-1)} + \sqrt{\frac{2\log t}{T_{i^*}(t-1)}}$$
  - Si on est dans l'intervale de confiance, on a alors dans ce cas :
  $\mu_i + 2\sqrt{\frac{2\log t}{T_i(t-1)}} \geq \mu^*$, soit: $T_i(t-1) \leq \frac{8\log t}{\Delta_i^2}$
  - Sinon, c'est que l'une des inégalités (a) ou (b) n'est pas vérifiée

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 $u = \dfrac{8 \log n}{\Delta_i^2}$, 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 :

$$\mathbb{E}_{T_i(n)} \leq \frac{8 \log n}{\Delta_i^2} + \sum_{t=u+1}^{n} [\sum_{s=u+1}^{t} t^{-4} + \sum_{s=1}^{t} t^{-4}]$$

$$\leq \frac{8 \log n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

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 $u = \dfrac{8 \log n}{\Delta_i^2}$, 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 :

$$\mathbb{E}_{T_i(n)} \leq \frac{8 \log n}{\Delta_i^2} + \sum_{t=u+1}^{n} [\sum_{s=u+1}^{t} t^{-4} + \sum_{s=1}^{t} t^{-4}]$$

$$\leq \frac{8 \log n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

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

UCB : Borne supérieure du regret

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

UCB : Borne supérieure du regret

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

  $\mathbb{E}(\rho_n)$

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

UCB : Borne supérieure du regret

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

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

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

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

UCB : Borne supérieure du regret

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

$\mathbb{E}(\rho_n)$

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

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

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

UCB : Borne supérieure du regret

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

$\mathbb{E}(\rho_n)$

$$
\begin{aligned}
&= n \times \mu_i^* - \mathbb{E}(\sum_{t=1}^{n} \mu_{\pi_t}) \\
&= n \times \mu_i^* - \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times \mu_i \\
&= \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times (\mu_i^* - \mu_i) \\
&= \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times \Delta_i
\end{aligned}
$$

UCB : Borne supérieure du regret

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

$\mathbb{E}(\rho_n)$

$$
\begin{aligned}
&= n \times \mu_i^* - \mathbb{E}(\sum_{t=1}^{n} \mu_{\pi_t}) \\
&= n \times \mu_i^* - \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times \mu_i \\
&= \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times (\mu_i^* - \mu_i) \\
&= \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times \Delta_i \\
&\leq \sum_{i \in \{1..K\} : \mu_i < \mu_{i*}}^{K} \frac{8 \log n}{\Delta_i} + \Delta_i (1 + \frac{\pi^2}{3})
\end{aligned}
$$

UCB : Borne supérieure du regret

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

$\mathbb{E}(\rho_n)$

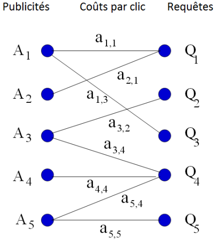
$$
\begin{aligned}
&= n \times \mu_i^* - \mathbb{E}(\sum_{t=1}^{n} \mu_{\pi_t}) \\
&= n \times \mu_i^* - \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times \mu_i \\
&= \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times (\mu_i^* - \mu_i) \\
&= \sum_{i=1}^{K} \mathbb{E}_{T_i}(n) \times \Delta_i \\
&\leq \sum_{i \in \{1..K\}:\mu_i < \mu_{i*}}^{K} \frac{8 \log n}{\Delta_i} + \Delta_i (1 + \frac{\pi^2}{3}) \\
&\leq K \frac{8 \log n}{\Delta^*} + K \Delta^* (1 + \frac{\pi^2}{3})
\end{aligned}
$$

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

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

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



$\Rightarrow$ 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:

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

$\Rightarrow$ 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

    $\Rightarrow$ Meilleure prise en compte de ces bras par considération de la variance empirique des rewards
  - Variance Empirique :

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

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

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

Avec

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

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)
    - $\Rightarrow$ Accélérer la sélection des meilleurs bras en apprenant des ”zones” de l’espace de représentation pertinentes
    - $\Rightarrow$ Cold-start pour nouveaux bras entrant dans le pool
  - Contexte variable : Hypothèse de non-stationnarité des rewards
    - $\Rightarrow$ Prise en compte globale de contextes individuels : rewards des bras suivent une distribution commune définie sur leurs contextes individuels
    - $\Rightarrow$ Prise en compte individuelle d’un contexte global : chaque bras suit une distribution indépendante contionnellement au contexte global de la décision
    - $\Rightarrow$ Prise en compte individuelle d’un contexte individuel : rewards de chaque bras dépendent de son état actuel

Bandits contextuels

- 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 :

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

  - 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

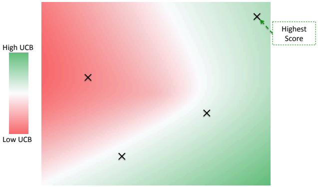
- Lin-UCB *[Li et al., 2010]*
  - Il peut être montré qu'avec une probabilité $1-\delta$:

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

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

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

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



Bandits contextuels

Algorithm 1 LinUCB with disjoint linear models.

Inputs: $\alpha \in \mathbb{R}_+$
for $t = 1, 2, 3, \ldots, T$ **do**
  Observe features of all arms $a \in \mathcal{A}_t$: $\mathbf{x}_{t,a} \in \mathbb{R}^d$
  for all $a \in \mathcal{A}_t$ **do**
    if a is new **then**
      $\mathbf{A}_a \leftarrow \mathbf{I}_d$ (d-dimensional identity matrix)
      $\mathbf{b}_a \leftarrow \mathbf{0}_{d \times 1}$ (d-dimensional zero vector)
    end if
    $\hat{\boldsymbol{\theta}}_a \leftarrow \mathbf{A}_a^{-1} \mathbf{b}_a$
    $p_{t,a} \leftarrow \hat{\boldsymbol{\theta}}_a^\top \mathbf{x}_{t,a} + \alpha \sqrt{\mathbf{x}_{t,a}^\top \mathbf{A}_a^{-1} \mathbf{x}_{t,a}}$
  end for
  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
  $\mathbf{A}_{a_t} \leftarrow \mathbf{A}_{a_t} + \mathbf{x}_{t,a_t} \mathbf{x}_{t,a_t}^\top$
  $\mathbf{b}_{a_t} \leftarrow \mathbf{b}_{a_t} + r_t \mathbf{x}_{t,a_t}$
end for

Avec $A_i = D_i^T D_i + I$ et $b_i = D_i^T c_i$

Bandits contextuels

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

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)

- Alternative aux stratégies optimistes : Thompson Sampling *[Thompson,1933],[Kaufmann et al., 2012]*
- Maximisation de l'espérance de reward :

$$\pi_t = \arg\max_i \int \mathbb{I}[\mathbb{E}(r_t|i, x_{i,t}, \theta) = \max_{i'} \mathbb{E}(r_t|i', x_{i',t}, \theta)] P(\theta|\mathcal{D})\, d\theta$$

Avec:

- $r_t = g(\omega_{i,\pi_t})$ le reward obtenu au temps t
- $\mathcal{D} = \{(i, t, x_{i,t}, r_t\}$ l'ensemble des observations passées;
- $P(r_t|\theta, i, t, x_{i,t})$ la vraissemblance des rewards en fonctions des observations
- $P(\theta|\mathcal{D}) \propto P(\mathcal{D}|\theta)P(\theta)$ la probabilité postérieure des paramètres conditionnnellement aux paramètres
- $P(\mathcal{D}|\theta)$ la vraissemblance des observations selon les paramètres
- $P(\theta)$ un prior sur l'ensemble de paramètres θ;

Bandits contextuels

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

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

- Thompson Sampling en pratique
- Cas linéaire *[Agrawal & Goyal, 2013]*:
  - $\mathbb{E}(r_t|i, x_{i,t}, \theta) =< \theta, x_{i,t} >$
  - On suppose que que la vraissemblance $P(\theta|\mathcal{D})$ suit une loi normale: $\mathcal{N}(\hat{\theta}(t), v^2 B^{-1}(t))$
  - Avec :
    - $B(t) = I + \sum_{s=1}^{t-1} x_{\pi_i,s} x_{\pi_i,s}^T$
    - $\hat{\theta}(t) = B^{-1}(t) \sum_{s=1}^{t-1} r_s x_{\pi_i,s}$
    - v une constante ≈ 0.25

$\Rightarrow$ Revient à supposer une distribution normale $\mathcal{N}(< \theta, x_{i,t} >, v^2)$ pour les rewards $P(r_t|x_{\pi_i,t})$

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

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.