Algorithme de bandit et obsolescence : un modèle pour la recommandation

Jonathan Louëdec[1,2], Laurent Rossi[1], Max Chevalier[2], Aurélien Garivier[1], et Josiane Mothe[2]

[1]Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse
[2]Institut de Recherche en Informatique de Toulouse, UMR 5505, CNRS, Université de Toulouse

1er juillet 2016

Résumé

Un nombre croissant de systèmes numériques font appel à des algorithmes de bandits pour combiner efficacement exploration de l'environnement et exploitation de l'information accumulée. Les modèles de bandits classiques sont toutefois assez naïfs : ils se bornent à un nombre fixé de choix disponibles (appelés *bras*), et à des réponses ne variant pas au cours du temps. Pour les moteurs de recommandation, par exemple, il s'agit de limitations sévères : de nouveaux items à recommander apparaissent régulièrement, et les anciens ont une tendance prévisible à perdre de l'attractivité. Pour faire face à ces problèmes, des stratégies capables de gérer l'évolution temporelle du gain moyen associé à chaque bras ont été proposées. Si ces stratégies sont assez générales, elles ne sont pas forcément les plus efficaces dans le cas où la forme de cette évolution temporelle est largement connue a priori.

Dans cet article nous proposons deux nouvelles stratégies capables de prendre en compte d'une part l'obsolescence progressive de chaque bras, et d'autre part l'arrivée de nouveaux bras : *Fading-UCB*, pour laquelle nous fournissons une analyse détaillée de la borne supérieure de regret, et *Trust and abandon*. Nous montrons expérimentalement que les deux stratégies proposées permettent d'obtenir de meilleures performances que celles obtenues par les stratégies de l'état de l'art.

Mots-clef : Systèmes de recommandation, Problèmes de bandit, Stratégies UCB, Flux de données, Bandit non-stationnaire.

1 Introduction

Cadre général Un système de recommandation considère généralement l'ensemble des interactions passées avec les utilisateurs pour estimer la popularité d'un objet [RRS11]. Cela nécessite que les objets soient suffisamment recommandés pour obtenir une estimation précise du gain associé. Cependant recommander tous les objets un grand nombre de fois est une stratégie dangereuse pour deux raisons : le gain moyen n'est pas optimal (car des objets non populaires sont trop souvent recommandés) et l'utilisateur peut choisir de ne plus utiliser le système (si les recommandations sont rarement intéressantes, voir [OT08]). Il faut donc être capable d'apprendre sur la popularité des objets disponibles, tout en maximisant le gain global : un tel problème est connu sous le nom de "dilemme exploration/exploitation". Les *algorithmes de bandits* offrent des solutions à ce dilemme (voir [BCB12]).

Dans la version stochastique des modèles de bandit, un agent choisit à chaque instant $t = 1, 2, \dots$ un *bras* $A_t \in \{1, \dots, K\}$, et reçoit une récompense X_t aléatoire dépendant de ce choix. Le cadre classique est stationnaire : les K bras sont disponibles du début à la fin et ont un gain moyen n'évoluant pas au cours du temps. Si ce modèle capture déjà l'essence du dilemme exploration/exploitation (il faut à la fois essayer les bras mal connus et tirer profit de ceux qui semblent les plus performants), il s'avère insuffisant dans de nombreuses applications. En particulier, de nombreux domaines voient leurs données se renouveler et vieillir au fil du temps. Il faudrait donc être capable de prendre en compte à la fois ce renouvellement et cette obsolescence progressive.

Dans cet article nous proposons un modèle de ban-

dit pour un cadre non-stationnaire, plus réaliste pour la recommandation car capable de prendre en compte l'obsolescence progressive de la popularité des objets. Nous considérons un modèle où l'intérêt moyen d'un bras diminue avec le temps et où de nouveaux bras apparaissent continuellement. Dans un système de recommandation, selon la nature de l'objet recommandé le gain peut être un taux de clics (publicité, site d'information, vidéo, ...), ou bien un nombre de ventes. Si la vitesse d'obsolescence de la popularité d'un objet n'est pas la même selon le type de recommandation, elle apparaît bien dans de nombreux cas.

État de l'art Dans notre modèle, la vitesse d'obsolescence du gain moyen de chaque bras est exponentielle et connue a priori. Les stratégies de bandit généralement connues telles que le *Thompson Sampling* ([Tho33]), les approches *UCB* ([ACBF02], [AMS09], [GC11]) ou encore les approches *Softmax* ([ACBFS02]) ne sont pas conçues pour appréhender un modèle non-stationnaire. Cependant plusieurs travaux de l'état de l'art considèrent des problèmes de bandit dans un cadre non-stationnaire. Garivier et Moulines (2008) [GM08] présentent une analyse d'une version pondérée en fonction du temps de l'approche UCB. Dans ce modèle, les observations les plus récentes sont considérées comme plus importantes. Dans ce même article les auteurs analysent une approche UCB utilisant une fenêtre glissante de taille fixe, où uniquement les informations les plus récentes sont prises en compte dans le calcul de la borne supérieure de confiance : *sliding-window UCB (SW-UCB).* Ces approches sont conçues pour un modèle où aucune information sur la manière dont le gain moyen évolue dans le temps n'est disponible. Chakrabarti et al. (2009) [CKRU09] proposent un modèle de "bandit mortel" dans lequel un bras a une durée de vie prédéfinie au delà de laquelle le bras disparaît. Ce modèle s'est inspiré de ce que l'on peut observer lors de la recommandation de nouvelles (Yahoo ! actualités, Le Monde, ...). Slivkins et Upfal (2007) [SU07] considèrent un cas où le gain moyen de chaque bras suit un mouvement brownien. Besbes et al. (2014) [BGZ14] proposent un modèle où le gain moyen des bras peut changer. Pour ces deux cas, l'exploration doit être suffisamment importante pour être capable d'identifier le possible changement de bras optimal. Tracà et Rubin [TR15] proposent une approche de type UCB prenant en compte les variations périodiques du nombre de visiteurs sur un système de recommandation. Cette approche suggère de réguler la part d'exploration selon l'affluence : plus il y a de visiteurs, plus il faut favoriser l'exploitation à l'exploration. L'exploration se fait donc quand peu de visiteurs sont en ligne. Dans l'article [KQ14], les auteurs proposent d'apprendre une fonction décroissante considérée comme une combinaison linéaire de fonctions connues. Cet article considère que la moyenne des récompenses d'un bras décline en fonction de son âge, comme nous le suggérons dans notre article.

Un autre moyen de prendre en compte l'évolution temporelle du gain moyen peut être d'utiliser des approches contextuelles. Par exemple, l'approche LinUCB proposée par Li et al. [LCLE10] est capable de prendre en compte des variables telles que l'âge du bras et l'heure. Dans notre modèle, nous disposons a priori d'informations sur la façon dont la popularité des bras décroît. Ainsi il est possible d'anticiper la décroissance de la popularité et d'adapter nos estimateurs à chaque instant. Cela nous permet de proposer des algorithmes plus performants.

Cadre envisagé et propositions Dans cet article, nous considérons que de nouveaux bras apparaissent de manière régulière (flux de bras) et que la popularité de l'ensemble des bras décroît de manière exponentielle (obsolescence progressive). Cette décroissance peut être paramétrée selon la nature des bras disponibles. Ce type de décroissance est fondée sur des observations empiriques sur les données du challenge CLEF-NEWSREEL [1]. Ce challenge fournit un jeu de données de deux mois d'interactions entre un système de recommandation d'information et ces utilisateurs. Nous avons pu observer que la popularité d'une information a tendance à décroître de façon exponentielle. Il apparaît que l'information atteint généralement sa popularité optimale dès son apparition, pour décroître très rapidement avec le temps qui passe. Ce phénomène est visible sur la Figure 1, qui représente la moyenne du nombre de clics obtenus sur les 100 informations les plus cliquées de la collection en fonction de leurs âges.

Considérer que la popularité des bras diminue selon un facteur exponentiel constant présente un fort avantage : l'ordre des bras ne change pas. Ainsi nous garantissons qu'un bras non-optimal à un instant donné ne le sera plus jamais, ou encore que si un bras obtient un gain moyen meilleur qu'un autre bras, il le sera toujours.

Le premier algorithme que nous proposons, ***Fading Upper Confidence Bound (F-UCB)***, est une adaptation de l'approche *Upper Confidence Bound (UCB)* pour laquelle une référence populaire est [ACBFS02]. Notre algorithme s'adapte aux flux de bras ainsi qu'à

1. News Recommendation Evaluation Lab de la conférence CLEF, 2015 [HKL+14]

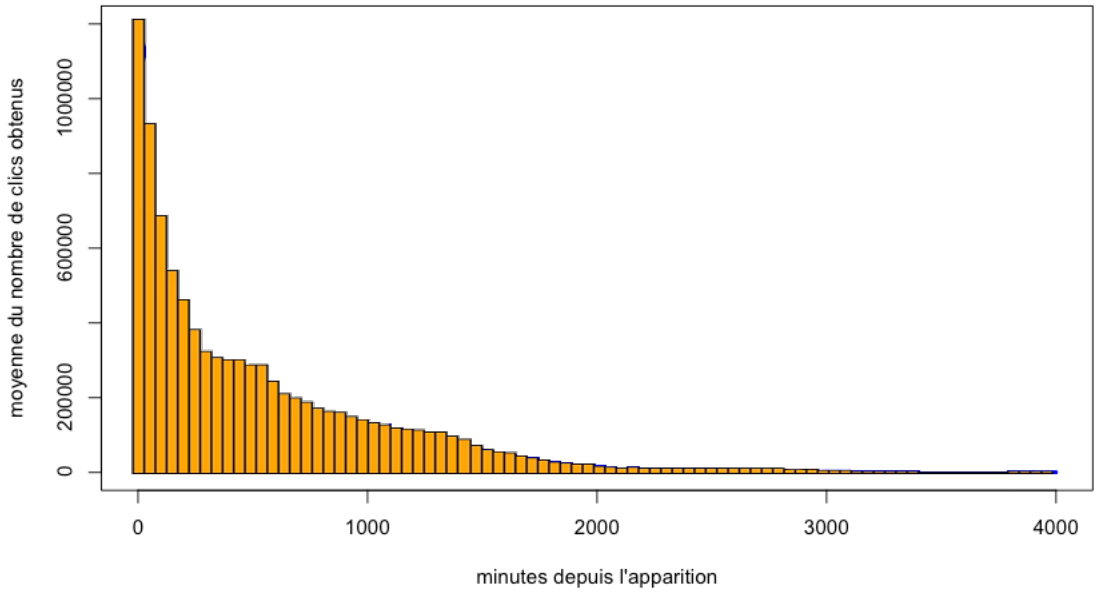

FIGURE 1 – Obsolescence de la popularité des bras. Données issues des 100 documents les plus cliqués du challenge CLEF-NEWSREEL

la décroissance exponentielle de la popularité des bras. L'étude de cet algorithme nous permet d'obtenir des garanties statistiques sur son niveau de performance. Nous proposons également un second algorithme : ***Trust and Abandon (TA)***. Elle s'appuie sur le modèle proposé, ainsi que sur l'hypothèse que lorsqu'un nouveau bras apparaît, les bras déjà existants ont été joués un nombre de fois suffisant pour avoir des estimations de leurs popularités assez précises. L'approche ***TA*** propose de jouer le dernier bras apparu tant qu'il est impossible de certifier qu'il est moins populaire qu'un autre bras. Deux variantes de cet algorithme sont proposées, la différence réside dans le critère qui permet de certifier qu'un bras est plus populaire qu'un autre. Nous montrons expérimentalement que l'algorithme *F-UCB* est légèrement meilleur que l'algorithme TA.

La suite de l'article est organisée de la manière suivante : dans la section 2 nous précisons notre modèle. La section 3 décrit en détails l'algorithme *F-UCB* et fournit une étude de la borne supérieure de regret de cet algorithme. Nous décrivons dans la section 4 l'algorithme *TA*. Les résultats expérimentaux sont présentés dans la section 5. Nous présenterons enfin nos conclusions ainsi qu'une réflexion sur les perspectives liées à ces travaux.

2 Formalisation du problème

Nous considérons une succession de périodes de taille fixe L. Au début de la période r, commençant à l'instant $1+rL$ le bras a_r entre en jeu et son instant d'apparition est $t_a = 1+rL$. La période en cours à l'instant t est $r(t) = \lfloor \frac{t}{L} \rfloor$. L'expérience se termine après K périodes, la période finale est $r = K-1$. À chaque instant t, l'agent choisit un bras A_t parmi ceux disponibles et reçoit une récompense $Z_t \in \{0,1\}$. Lors de l'apparition des bras, les récompenses sont d'espérances p_a, où a est le bras associé.

Les espérances associées aux bras décroissent à la même vitesse, selon le facteur d'obsolescence progressive τ. Comme nous l'avons souligné plus tôt dans cet article, il est important de préciser qu'un bras qui n'est pas optimal à l'instant t ne le sera jamais par la suite. Le bras optimal durant la période r est noté a_r^*. En outre, un bras a reste optimal un nombre r_a de périodes, pouvant être nul, et ne le sera plus jamais. Lors de son apparition, nous supposons qu'un bras a une espérance supérieure à une certaine valeur η fixée a priori. Cela induit qu'un bras apparu il y a plus de $\tau \log \frac{1}{\eta}$ instants ne peut pas être le bras optimal. L'espérance d'un bras a à un instant $t \geq t_a$ est notée :

$$\mu_a(t) = p_a \exp^{-\frac{t-t_a}{\tau}}$$

On note $\widetilde{N}_a(T)$ le nombre de fois où le bras a a été joué en étant sous-optimal jusqu'à l'instant T :

$$\widetilde{N}_a(T) = \sum_{t=t_a+r_aL}^{\left(t_a+\tau\log(\frac{1}{\eta})\right)\wedge T} \mathbb{1}_{\{I_t=a\}} \tag{1}$$

où r_a est le nombre de périodes pendant lesquels a est optimal.
Une stratégie de bandit vise à minimiser la différence entre la somme des récompenses obtenues en utilisant le bras optimal a_r^* à chaque période et la somme des récompenses obtenues via la stratégie utilisée. Cette différence est appelée le regret cumulé moyen :

$$R(T) = \sum_{r=0}^{K-1}\sum_{t=1+rL}^{(r+1)L} \mathbb{E}[Z^{a_r^*}]e^{-\frac{rL-t_{a_r^*}}{\tau}} - \sum_{t=1}^{T}\mathbb{E}[Z_t]$$

3 La stratégie *Fading-UCB*

L'objectif de cette stratégie est d'adapter la stratégie *Upper Confidence Bound* [ACBFS02] pour la prise en compte de l'obsolescence des bras et d'un flux constant de bras. Cette stratégie utilise non pas l'espérance estimée de chaque bras, mais des bornes de confiance supérieures de cette espérance. C'est une stratégie dites "optimiste dans l'incertain" : c'est à dire qu'elle consiste à jouer le bras ayant potentiellement la plus forte récompense.

L'espérance d'un bras lors de son apparition p_a est estimée par $\hat{p}_a(t)$:

$$\hat{p}_a(t) = \frac{1}{N_a(t)}\sum_{s=1}^{t}\left(Z_s \exp\left(\frac{s-t_a}{\tau}\right)\mathbb{1}_{A_s=a}\right) \tag{2}$$

avec $N_a(t)$ le nombre de fois où le bras a a été joué aux t premiers instants et τ le facteur d'obsolescence, défini a priori.

Pour la suite de l'analyse, nous projetons l'ensemble des estimateurs à l'instant t en cours. Pour cela, nous proposons d'utiliser non pas l'espérance du bras lors de son apparition p_a mais l'espérance à l'instant t $\mu_a(t)$, est estimée par :

$$\hat{\mu}_a(t) = \hat{p}_a(t)\exp\left(-\frac{t-t_a}{\tau}\right) \tag{3}$$

3.1 Calibration de l'intervalle de confiance

Notre objectif est de trouver un intervalle de risque α pour l'espérance en début de période $\mu_a(t)$, variable comprise entre 0 et $M = e^{-\frac{t-t_a}{\tau}}$. En appliquant l'inégalité d'Hoeffding, décrite en annexe, on obtient la valeur

$$M\sqrt{\frac{2}{n}\log\frac{1}{\alpha}} \tag{4}$$

Classiquement la valeur choisie pour α est $1/t$, ainsi au fur et à mesure des interactions, la borne croît lentement si le bras associé n'est pas joué. Le bras sera rejoué à un moment ou un autre sur le long terme. Mais ici, un bras qui est apparu il y a plus de $\tau\log\frac{1}{\eta}$ instants ne peut pas être le bras optimal, il est donc inutile de faire grandir α au delà de cette valeur, cela revient à fixer $\frac{1}{\alpha} = \tau\log\frac{1}{\eta}$.

Au final, en majorant M par 1, on obtient la borne de confiance supérieure $U_a(t)$ suivante :

$$U_a(t) = \hat{\mu}_a(t) + \sqrt{\frac{2\log\left(\tau\log\frac{1}{\eta}\right)}{N_a(t)}} \tag{5}$$

3.2 Algorithme

L'algorithme $F-UCB$ consiste à jouer le bras avec la borne de confiance supérieure $U_a(t)$ la plus grande. Il est décrit dans l'algorithme 1.

Algorithme 1 : Algorithme *F-UCB*

```
A : vecteur contenant les bras disponibles
pour r = 0, ..., K − 1 faire
    A = A ∪ a_r
    pour t = 1 + rL, ..., (r + 1)L faire
        U_a(t) = μ̂_a(t) + sqrt( 2 log(τ log 1/η) / N_a(t) )
        Jouer A_t = argmax_{a∈A} U_a(t)
    fin
fin
```

3.3 Borne supérieure de regret

Pour rappel, $\mu_a(t)$, l'espérance d'un bras a à un instant $t \geq t_a$ vaut $p_a \exp^{-\frac{t-t_a}{\tau}}$ et l'espérance d'un bras optimal a_t^* au même instant est :

$$\mu^*(t) = \mu_{a_t^*}(t) = \max_a \mu_a(t).$$

Pour un bras a sous-optimal lors de la dernière période, le gap minimal Δ_a est défini par :

$$\Delta_a = \min_{t_a + r_a L \leq t \leq t_a + \tau \log(1/\eta)} \mu^*(t) - \mu_a(t)$$

Ce gap minimal est utilisé dans le théorème 1 qui donne une borne pour le regret cumulé moyen :

Théorème 1 *Le regret cumulé moyen $R(T)$ de l'algorithme F-UCB vérifie :*

$$R(T) \leq 8 \log\left(\tau \log \frac{1}{\eta}\right) \sum_a \frac{1}{\Delta_a^2} + (K-1)\left(1 + \frac{2}{\tau \log \frac{1}{\eta}}\right)$$

en sommant sur les bras a sous-optimaux.

La preuve de ce théorème adapte les arguments utilisés pour obtenir la borne du regret dans le cas d'un algorithme *Upper Confidence Bound*. Les éléments importants de la démonstration sont donnés en appendice.

4 La stratégie *Trust and Abandon*

4.1 Motivation

Nous allons maintenant définir une politique où l'on fait l'hypothèse qu'au début d'une période r, les bras a_i avec $i \in 1, \ldots, r-1$ ont été joués un nombre de fois suffisant pour avoir des estimations de leurs popularités assez précises. Les espérances μ_a de ces mêmes bras au début de la période r sont comprises entre $\eta \exp\left(\frac{rL - t_a}{\tau}\right)$ et $\exp\left(\frac{rL - t_a}{\tau}\right)$. Le bras entrant a_r a besoin d'être testés un certain nombre de fois pour pouvoir estimer son espérance, située entre η et 1 durant la période r. Ce bras est le seul à pouvoir atteindre une espérance très proche de 1. L'approche *TA* que nous proposons n'est pas sans rappeler l'algorithme *Successive Elimination* [EDMM06].

Tout comme la stratégie *F-UCB*, nous projetons l'ensemble des estimateurs à l'instant de départ de cette période. Pour cela nous utilisons les estimateurs $\hat{\mu}_a(t)$ et les bornes supérieures de confiance $U_a(t)$ présentées la section 3.1.

4.2 Algorithme

La stratégie *Trust and Abandon* consiste à jouer durant la période r le bras a_r tant qu'il est impossible de certifier que son espérance est plus faible que celle d'un autre bras. Afin de certifier cela, nous utilisons la borne de confiance supérieure $U_a(t)$ de $\mu_a(t)$. Il faut ensuite choisir le moment où $U_{a_r}(t)$ devient trop faible pour certifier qu'un autre bras obtient en moyenne un meilleur gain moyen. Plusieurs choix sont possibles.

Dans une première version, *TA-*μ (algorithme 2), nous cherchons l'instant où cette borne devient inférieure à $\max_{a \in A} \hat{\mu}_a(t)$, l'espérance estimée la plus grande au début de la période en cours, le bras a_r n'est plus joué, au profit du bras $\text{argmax}_{a \in A} \hat{\mu}_a(t)$.

Dans une seconde version, *TA-UCB* (algorithme 3), nous cherchons l'instant où cette borne devient inférieure à $\max_a U_a(t)$, la plus grande borne de confiance au début de la période en cours, le bras a_r n'est plus joué, au profit du bras $\text{argmax}_a \hat{\mu}_a(t)$. Si cette deuxième version ressemble à l'approche *F-UCB*, la principale différence réside dans le choix du bras joué une fois le bras a_r éliminé : *TA-UCB* joue le bras avec l'espérance la plus forte, tandis que *F-UCB* joue le bras avec la borne de confiance supérieure du gain associé la plus forte.

Algorithme 2 : Trust and Abandon μ ($TA - \mu$)

```
A : vecteur contenant les bras disponibles
pour r = 0, ..., K − 1 faire
    A = A ∪ a_r
    pour t = 1 + rL, ..., (r + 1)L faire
        si max_{a∈A} μ̂_a(t) > U_{a_r}(t) alors
            Jouer A_t = argmax_{a∈A} μ̂_a(t)
        sinon
            Jouer A_t = a_r
        fin
    fin
fin
```

Algorithme 3 : Trust and Abandon UCB ($TA - UCB$)

```
A : vecteur contenant les bras disponibles
pour r = 0, ..., K − 1 faire
    A = A ∪ a_r
    pour t = 1 + rL, ..., (r + 1)L faire
        si max_{a∈A} U_a(t) > U_{a_r}(t) alors
            Jouer A_t = argmax_{a∈A} μ̂_a(t)
        sinon
            Jouer A_t = a_r
        fin
    fin
fin
```

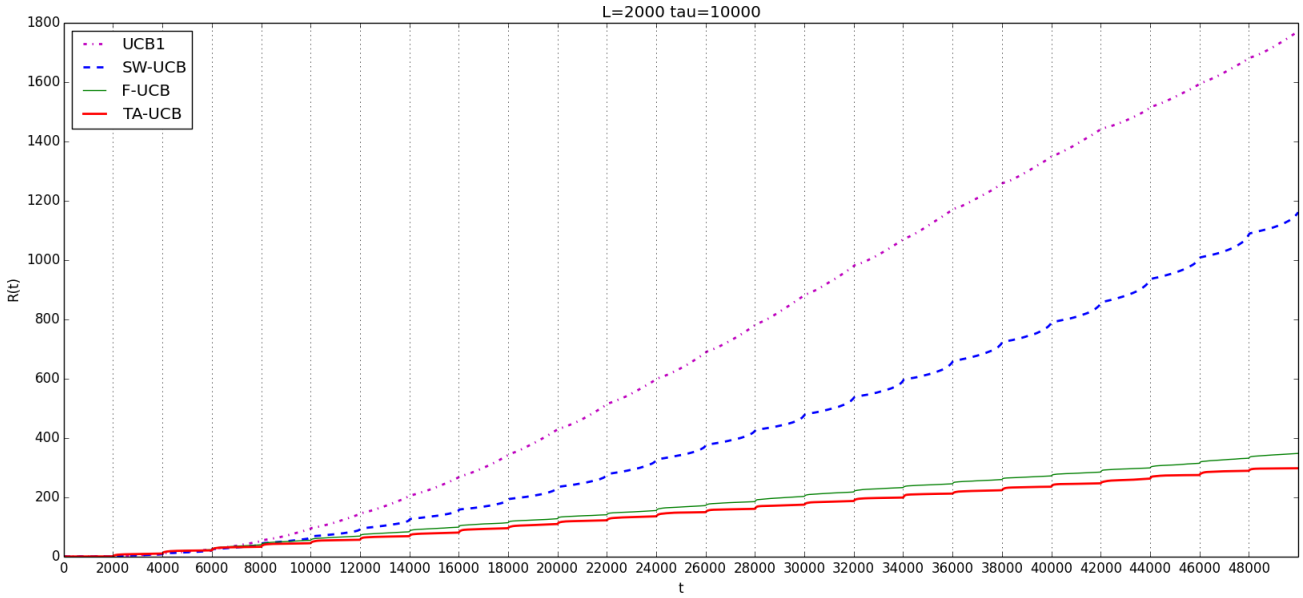

FIGURE 2 – Simulation avec $L = 2000$ et $\tau = 5L$

Algorithme 4 : UCB1

A :vecteur contenant les bras disponibles
pour $t = 1, \ldots, KL$ **faire**
  si $t\%(L+1) = 0$ **alors**
    $A = A \cup a_r$
  fin
  $A_t = \text{argmax}_{a \in A} \frac{1}{N_a} \sum_{s=1}^{t} Z_s \mathbb{1}_{A_s=a} + \sqrt{\frac{2\log(t)}{N_a}}$
  Jouer A_t
fin

Algorithme 5 : Sliding-window UCB (SW-UCB)

A : vecteur contenant les bras disponibles
S : taille de la fenêtre
$N_t(S,a) = \sum_{s=t-S}^{t} \mathbb{1}_{A_s=a}$
$\hat{X}_t(S,a) = \frac{1}{N_t(S,a)_s} \mathbb{1}_{A_s=a}$
pour $t = 1, \ldots, KL$ **faire**
  si $t\%(L+1) = 0$ **alors**
    $A = A \cup a_r$
  fin
  $A_t = \text{argmax}_{a \in A} \hat{X}_t(S,a) + \sqrt{\frac{2\log(t \wedge S)}{N_t(S,a)}}$
  Jouer A_t
fin

5 Résultats expérimentaux

Cadre expérimental Pour évaluer les performances de nos approches et les comparer à celles de l'état de l'art, nous avons mis en place des simulations. Lors de ces simulations, nous varions la longueur d'une période L et le paramètre d'obsolescence τ afin d'évaluer leurs effets sur le regret cumulé. Chaque expérimentation est effectuée sur un intervalle de temps $T = 10\ 000$. Le nombre total de bras k ainsi que le nombre de périodes R sont égaux à $\lfloor \frac{T}{L} \rfloor$. Les résultats présentés sont la moyenne de 200 répétitions de chaque expérimentation. Les performances de l'algorithme *TA-UCB* sont toujours meilleurs que celles de l'algorithme *TA-μ*. Pour alléger les graphiques nous ne présenterons que les résultats de l'algorithme *TA-UCB*.

Les approches *F-UCB* et *TA-UCB* de cet article sont comparées à deux approches de l'état de l'art : *UCB1* (algorithme 4) et *SW-UCB* (algorithme 5). La première approche ne prend pas en compte l'obsolescence du gain moyen des bras. Ce gain va diminuer avec le temps car les récompenses diminues, cela implique que les bras avec un gain moyen faible seront rejouées régulièrement. L'approche *SW-UCB* utilise une fenêtre glissante de taille fixe S, où uniquement les S dernières interactions sont prises en compte dans le calcul de la borne supérieure de confiance. Selon la taille de la fenêtre les résultats varient. Une fenêtre de petite taille implique qu'il faut régulièrement rejouer les bras, car les interactions passées avec ces bras ne sont

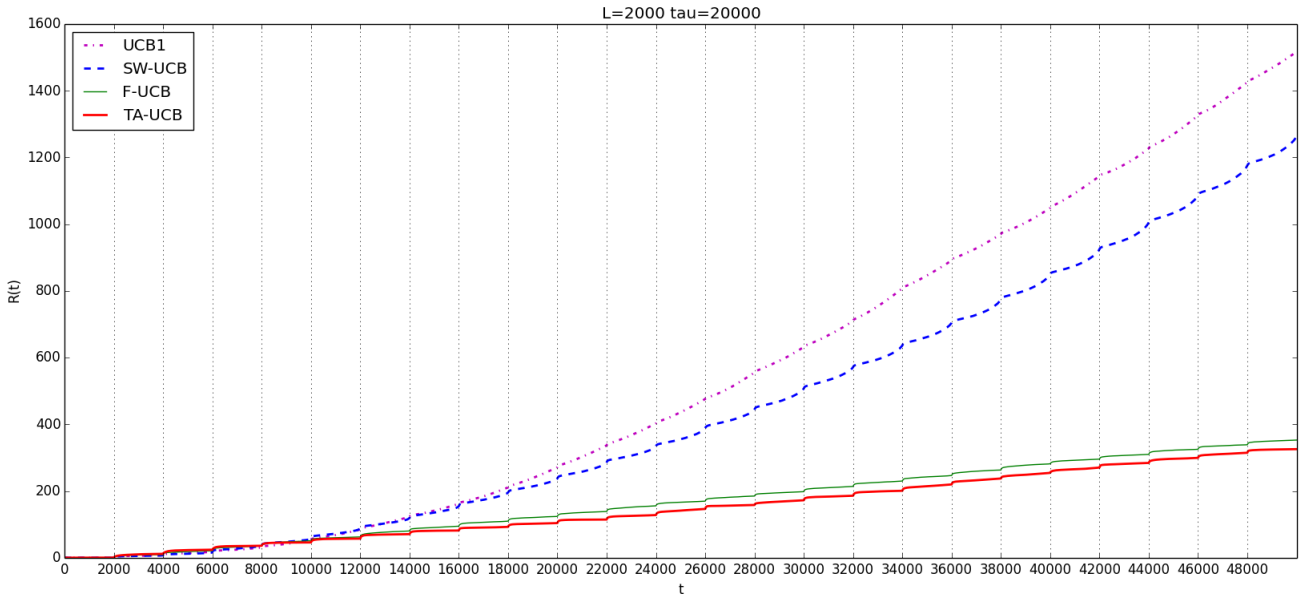

FIGURE 3 – Simulation avec $L = 2000$ et $\tau = 10L$

plus prises en compte. Une fenêtre de très grande taille revient à ne pas prendre en compte l'obsolescence progressive des bras. Nous avons donc testé plusieurs valeurs possibles pour la taille de la fenêtre $S = xL$ avec $x \in [0.5, 1, 1.5, 2, 2.5, 3]$.

Les bornes supérieures de confiance de l'ensemble des algorithmes utilisés dans cette expérimentation sont calculées en utilisant la constante 0.5 au lieu de 2 à l'intérieur de la racine. Si cette valeur habituelle 2 permet d'analyser l'algorithme afin d'obtenir des bornes de regret supérieures, elle donne une part trop forte à l'exploration.

Lors de nos expérimentations, plusieurs valeurs de $\tau = xL$ avec $x \in [3, 4, 5, 6, 7, 8, 9, 10]$ et de $L \in [500, 1000, 2000, 5000]$ ont été testées. La figure 2 est obtenue en utilisant $L = 2000$ et $\tau = 5L$ et la figure 3 en utilisant $L = 2000$ et $\tau = 10L$. Ce choix a été fait pour mettre en évidence comment la vitesse d'obsolescence impacte les résultats. Il est important de rappeler que le choix des valeurs de τ et η induisent la durée maximum de vie d'un bras. Par exemple si $\eta = 0.1$ et $\tau = 5L$, un bras ne peut plus être optimal à partir de $\tau \log \frac{1}{\eta} = 5L$, ce qui signifie qu'un bras peut être optimal au maximum 5 périodes après son apparition, et ensuite mis de côté. Les expérimentations sont réalisées avec $\eta = 0.1$, ainsi un bras peut être optimal au maximum τ périodes. Concernant l'optimisation de la taille S de la fenêtre pour l'algorithme *SW-UCB*. Expérimentalement la valeur qui permet d'obtenir les meilleurs résultats avec $L = 2000$ est $S = 2L$.

Résultats Sur l'ensemble des cas expérimentés selon le cadre précédemment définie, les différents regrets cumulés $R^{\alpha}(t)$ peuvent être classés comme suit :

$$R^{UCB}(t) > R^{SW-UCB}(t) > R^{F-UCB}(t) \geq R^{TA-UCB}(t) \quad (6)$$

L'approche *UCB1* est celle qui obtient les performances les plus faibles en terme de regret cumulé. En ne prenant pas en compte l'obsolescence progressive des bras, cette approche rejoue régulièrement les bras avec des gains cumulés faibles, car le gain moyen du bras optimal diminue et cette décroissance n'est pas répercutée sur l'ensemble des bornes. En utilisant une fenêtre glissante, l'approche *SW-UCB* permet d'obtenir de meilleurs résultats que l'approche *UCB1*, mais ces résultats restent bien plus faibles que ceux obtenus par les approches présentées dans cet article.

L'approche *TA-UCB* est celle qui obtient le regret cumulé le plus faible sur les deux figures. L'approche *F-UCB* obtient des performances très proches de l'approche *TA-UCB*. Lorsque le nouveau bras entrant n'est pas le meilleur, l'utilisation de la borne de confiance calculée par *F-UCB* fait qu'il est assez long d'éliminer les autres bras s'ils n'ont pas été suffisamment joués par le passé, alors que *TA-UCB* va privilégier le bras qui a la plus forte espérance estimée directement. Pour de grandes valeurs de $L \geq 10000$, il se trouve que *TA-UCB* fait aussi bien que *F-UCB* car leurs comporte-

ments est très similaires lorsque les bras sont joués un grand nombre de fois avant l'insertion d'un nouveau bras.

6 Conclusion

Dans cet article nous proposons un modèle dans lequel de nouveaux bras apparaissent régulièrement et où le gain moyen associé à chaque bras décroît de manière exponentielle. Ce modèle est inspiré de plusieurs observations empiriques sur un jeu de données du challenge NEWS-REEL de la conférence CLEF, où l'obsolescence des bras est visible. Une première approche est proposée : *Fading-UCB*. Elle représente une adaptation pour ce modèle de la stratégie UCB, une analyse nous permet de quantifier une borne supérieure de regret pour cette stratégie. La seconde approche proposée, *Trust and Abandon* repose sur une hypothèse stricte : lors de l'apparition d'un bras, nous estimons que l'ensemble des autres bras ont été suffisamment jouer pour avoir une estimation précise des gains moyens associés. Plusieurs simulations nous permettent d'observer que ces deux approches, *Fading-UCB* et *Trust and Abandon*, obtiennent de bien meilleures performances en terme de regret cumulé que l'approche de l'état de l'art *Sliding-Window UCB*, conçue pour prendre en compte la non-stationnarité du gain moyen de chaque bras. L'ensemble de ces approches surpassent largement la stratégie *UCB1*, qui ne prend pas du tout en compte l'obsolescence progressive des bras.

Dans nos travaux futurs, nous souhaitons analyser la borne supérieure de regret de l'approche *Trust and Abandon*, qui obtient de bons résultats expérimentalement. Nous souhaitons également proposer un cadre expérimental utilisant des données issues de systèmes de recommandation d'information, tels que le jeu de données *Yahoo News Feed dataset* qui contient plus d'un tera octet d'interactions, ou encore les données proposées par les différents challenges NEWSREEL de la conférence CLEF.

7 Remerciements

Les travaux de Jonathan Louëdec sont financés par le LabEx CIMI. Les auteurs remercient le soutien apporté par l'Agence Nationale de la Recherche (ANR-13-CORD-0020, projet ALICIA, ANR-13-BS01-0005, projet SPADRO)

8 Appendice

Les éléments principaux de la preuve du théorème 1 sont donnés ci-dessous. Un bras optimal lors de la dernière période est optimal sur toute sa durée de vie, on s'intéresse à un bras a sous-optimal à un instant donné donc forcément sous-optimal lors de la dernière période. Le nombre de fois que le bras a est joué en étant sous-optimal jusqu'à l'instant T est noté $\widetilde{N}_a(T)$. Le bras a est sous optimal entre les instants $t_1 = t_a + r_a L$ et $t_2 = (t_a + \tau \log(1/\eta)) \wedge T$. Par conséquent, en posant $t_1 = t_a + (r_a L) \vee 1$,

$$\begin{aligned}
\widetilde{N}_a(T) &= \sum_{t=t_a+r_aL}^{t_2} \mathbb{1}_{\{I_t=a\}} \\
&\leq 1 + \sum_{t=t_1}^{t_2} \mathbb{1}_{\{I_t=a\}} \\
&\leq l + \sum_{t=t_1}^{t_2} \mathbb{1}_{\left\{I_t=a, \widetilde{N}_a(t-1)\geq l\right\}} \\
&\leq l + \sum_{t=t_1}^{t_2} \mathbb{1}_{\left\{U_{a^*}(t-1)\leq U_a(t-1), \widetilde{N}_a(t-1)\geq l\right\}} \\
&\leq l + \sum_{t=t_1}^{t_2} \mathbb{1}_{\{U_{a^*}(t-1)\leq U_a(t-1), N_a(t-1)\geq l\}}
\end{aligned}$$

en notant $a^*(t)$ le bras optimal à l'instant t, $N^*(t) = N_{a^*(t)}(t)$ et $U^*(t) = U_{a^*(t)}(t)$. La moyenne empirique $\hat{\mu}_a(t)$ fait intervenir les t valeurs de la suite

$$\left(Z_s \mathbb{1}_{A_s=a} \exp^{\frac{s}{\tau}}\right)_{1\leq s\leq t}.$$

Cette suite est constituée des $N_a(t)$ valeurs $Z_s \exp^{\frac{s}{\tau}}$ lorsque le bras a est tiré et de $t - N_a(t)$ valeurs nulles lorsque le bras a n'est pas tiré. Soit $(X_{a,i})_{1\leq i\leq N_a(t)}$ la suite constituée des valeurs (comprises entre 0 et 1) obtenues lorsque le bras a est tiré. En utilisant cette suite, $\hat{\mu}_a(t)$ s'écrit

$$\hat{\mu}_a(t) = \frac{1}{N_a(t)} \sum_{i=1}^{N_a(t)} X_{a,i} \exp\left(-\frac{t}{\tau}\right)$$

et on introduit l'expression :

$$\hat{\mu}_{a,s}(t) = \frac{1}{s} \sum_{i=1}^{s} X_{a,i} \exp\left(-\frac{t}{\tau}\right).$$

Cette notation et la définition de la borne supérieure de confiance permet d'écrire les inclusions suivantes :

$$\left\{U_{a^*(t-1)}(t-1) \leq U_a(t-1), N_a(t-1) \geq l\right\} \subset$$

$$\left\{ \min_{1\leq s^*\leq N^*(t-1)} \hat{\mu}_{a^*_{t-1},s^*}(t-1) + \sqrt{\frac{2m}{s^*}} \leq \max_{l\leq s\leq N_a(t-1)} \hat{\mu}_{a,s}(t-1) + \sqrt{\frac{2m}{s}} \right\} \subset \bigcup_{s^*=1}^{N^*(t-1)} \bigcup_{s=l}^{N_a(t-1)} A_{s,s^*}(t-1)$$

en notant $m = \log\left(\eta \log \frac{1}{\tau}\right)$ et $A_{s,s^*}(t)$ l'événement

$$\left\{ \hat{\mu}_{a,s^*}(t) + \sqrt{\frac{2m}{s^*}} \leq \hat{\mu}_{a,s}(t) + \sqrt{\frac{2m}{s}} \right\}.$$

Observer l'événement $A_{s,s^*}(t-1)$ implique au moins l'un des 3 cas suivants :

$$A^1_s(t-1) = \left\{ \hat{\mu}_{a,s}(t-1) - \sqrt{\frac{2m}{s}} \geq \mu_a(t-1) \right\}$$

$$A^2_{s^*}(t-1) = \left\{ \hat{\mu}_{a^*_{t-1},s^*}(t-1) + \sqrt{\frac{2m}{s^*}} \leq \mu^*(t-1) \right\}$$

$$A^3_s(t-1) = \left\{ \mu^*(t-1) \leq \mu_a(t-1) + 2\sqrt{\frac{2m}{s}} \right\}$$

La probabilité de $A^1_s(t-1)$ peut s'écrire

$$\mathbb{P}\left(\sum_{i=1}^{s^*} X_{a^*_{t-1},i} e^{-\frac{t-1}{\tau}} \leq s^* p_{a^*_{t-1}} e^{-\frac{t-1-T_a}{\tau}} - \sqrt{2s^*m} \right)$$

et l'inégalité de Hoeffding, dont un énoncé est rappelé en fin d'appendice, permet de majorer cette probabilité par :

$$\exp\left(-2\frac{\left(\sqrt{2s^*m}\right)^2}{s^*} \right) = \left(\tau \log \frac{1}{\eta}\right)^{-4}$$

De manière analogue, $\mathbb{P}\left(A^2_{s^*}(t-1)\right)$ se majore par la même expression. Lorsque s devient supérieur à la valeur

$$\left\lceil \frac{8\log\left(\eta \log \frac{1}{\tau}\right)}{\Delta_a^2} \right\rceil$$

la probabilité $\mathbb{P}\left(A^3_s(t-1)\right)$ devient nulle. Finalement, en choisissant comme valeur de l l'expression ci-dessus, $\mathbb{E}\left(\widetilde{N}_a(T)\right)$ est majoré par :

$$\left\lceil \frac{8\log\left(\tau \log \frac{1}{\eta}\right)}{\Delta_a^2} \right\rceil + \sum_{t=t_1}^{t_2} \sum_{s^*=1}^{N^*(t-1)} \sum_{s=1}^{N_a(t-1)} 2\left(\tau \log \frac{1}{\eta}\right)^{-4}$$

$$\leq \frac{8\log\left(\tau \log \frac{1}{\eta}\right)}{\Delta_a^2} + 1 + 2\left(\tau \log \frac{1}{\eta}\right)^3 \left(\tau \log \frac{1}{\eta}\right)^{-4}$$

en utilisant que $t_2 - t_1 + 1$, $N_a(t)$ et $N^*(t)$ sont majorés par $\tau \log \frac{1}{\eta}$. Finalement,

$$\mathbb{E}\left(\widetilde{N}_a(T)\right) \leq \frac{8\log\left(\tau \log \frac{1}{\eta}\right)}{\Delta_a^2} + 1 + \frac{2}{\tau \log \frac{1}{\eta}}$$

ce qui permet d'obtenir une borne pour le regret $R(T)$ en sommant sur les bras a sous-optimaux :

$$\begin{aligned} R(T) &\leq \sum_a \mathbb{E}\left(\widetilde{N}_a(T)\right) \\ &= \sum_a \left(\frac{8\log\left(\tau \log \frac{1}{\eta}\right)}{\Delta_a^2} + 1 + \frac{2}{\tau \log \frac{1}{\eta}} \right) \\ &\leq 8\log\left(\tau \log \frac{1}{\eta}\right) \sum_a \frac{1}{\Delta_a^2} \\ &\quad + (K-1)\left(1 + \frac{2}{\tau \log \frac{1}{\eta}}\right) \end{aligned}$$

où K est le nombre total de périodes.

Inégalité d'Hoeffding : Soit (X_i) avec $1 \leq i \leq n$ une suite de variables aléatoires indépendantes de même espérance p avec $X_i \in [a_i, b_i]$. Pour tout $s > 0$,

$$\mathbb{P}(|\bar{X} - p| \geq s) \leq 2\exp\left(\frac{-2n^2s^2}{\sum_{i=1}^n (b_i - a_i)^2} \right)$$

Références

[ACBF02] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3) :235–256, 2002.

[ACBFS02] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1) :48–77, 2002.

[AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19) :1876–1902, 2009.

[BCB12] Sebastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1) :1–122, 2012.

[BGZ14] Omar Besbes, Yonatan Gur, and Assaf Zeevi. Optimal exploration-exploitation in a multi-armed-bandit problem with non-stationary rewards. *arXiv preprint arXiv :1405.3316*, 2014.

[CKRU09] Deepayan Chakrabarti, Ravi Kumar, Filip Radlinski, and Eli Upfal. Mortal multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 273–280, 2009.

[EDMM06] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *The Journal of Machine Learning Research*, 7 :1079–1105, 2006.

[GC11] Aurélien Garivier and Olivier Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. In *Proceedings of the 24th Conference on Learning Theory (COLT)*, 2011.

[GM08] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *arXiv preprint arXiv :0805.3415*, 2008.

[HKL+14] Frank Hopfgartner, Benjamin Kille, Andreas Lommatzsch, Till Plumbaum, Torben Brodt, and Tobias Heintz. Benchmarking news recommendations in a living lab. In *CLEF'14 : Proceedings of the 5th International Conference of the CLEF Initiative*, LNCS, pages 250–267. Springer Verlag, 09 2014.

[KQ14] Junpei Komiyama and Tao Qin. Time-decaying bandits for non-stationary systems. In *Web and Internet Economics*, pages 460–466. Springer, 2014.

[LCLE10] Lihong Li, Wei Chu, John Langford, and Robert E.Schapire. A contextual-bandit approach to personalized news article recommandation. In *Proceedings of 19th International World Wide Web Conference*, pages 661–670, 2010.

[OT08] Heather L O'Brien and Elaine G Toms. What is user engagement ? a conceptual framework for defining user engagement with technology. *Journal of the American Society for Information Science and Technology*, 59(6) :938–955, 2008.

[RRS11] Francesco Ricci, Lior Rokach, and Bracha Shapira. Introduction to recommender systems handbook. In *Recommender Systems Handbook*, pages 1–35. Springer, 2011.

[SU07] Aleksandrs Slivkins and Eli Upfal. Adapting to a stochastically changing environment : The dynamic multi-armed bandits problem. Technical report, Technical Report CS-07-05, Brown University, 2007.

[Tho33] William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, pages 285–294, 1933.

[TR15] Stefano Tracà and Cynthia Rudin. Regulating greed over time. *arXiv preprint arXiv :1505.05629*, 2015.