

# Les bandits stochastiques à récompenses d'espérance non-définie

Adam Cohen, Maxime Genest, Vincent Masse

29 novembre 2020

## Rappel sur les bandits stochastiques classiques

- Ensemble de  $K$  actions (bras, machines).
- Chaque action  $k$  est associée à un paramètre inconnu  $\mu_k$  tel que  $X_{k_t} \sim \nu(\mu_k)$  où  $\nu(\mu_k)$  est une distribution d'espérance  $\mu_k$ .

## Rappel sur les bandits stochastiques classiques

- Ensemble de  $K$  actions (bras, machines).
- Chaque action  $k$  est associée à un paramètre inconnu  $\mu_k$  tel que  $X_{k_t} \sim \nu(\mu_k)$  où  $\nu(\mu_k)$  est une distribution d'espérance  $\mu_k$ .

Dans le jeu des bandits stochastiques, à chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- On observe une récompense (reward)  $r_t \sim \nu(\mu_{k_t})$ .

But : Déterminer une politique d'action qui maximisera  $\mathbb{E} \left[ \sum_{t=1}^T r_t \right]$

# Mesure de performance empirique pour les bandits stochastiques

Dans cette situation, à chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent cumule un regret :

$$\Delta_{k_t} = \mu^* - \mu_{k_t}$$

À la fin de l'épisode, on peut calculer le regret cumulatif empirique :

$$R(T) = \sum_{t=1}^T \Delta_{k_t}$$

Cela nous permet de comparer empiriquement la performance de plusieurs politiques, en simulant plusieurs épisodes et en comparant le regret cumulatif moyen sur ces épisodes.

## L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs lois de probabilité ont une **espérance non-définie**.

## L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs lois de probabilité ont une **espérance non-définie**.

Par exemple, la loi de Cauchy ou certaines configuration de la loi de Pareto.

# L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs lois de probabilité ont une **espérance non-définie**.

Par exemple, la loi de Cauchy ou certaines configuration de la loi de Pareto.

## Intérêts du projet

- Intérêt personnel (curiosité intellectuel).

# L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs lois de probabilité ont une **espérance non-définie**.

Par exemple, la loi de Cauchy ou certaines configuration de la loi de Pareto.

## Intérêts du projet

- Intérêt personnel (curiosité intellectuel).
- Modélisation du temps d'attente/temps de service par des distributions à queues lourdes.



# L'hypothèse d'existence de l'espérance

Le jeu des bandits stochastiques ainsi présenté sous-entend que la distribution des rewards associés aux bras du bandit est d'espérance qui existe. Or, plusieurs lois de probabilité ont une **espérance non-définie**.

Par exemple, la loi de Cauchy ou certaines configuration de la loi de Pareto.

## Intérêts du projet

- Intérêt personnel (curiosité intellectuel).
- Modélisation du temps d'attente/temps de service par des distributions à queues lourdes.

*Yu Li. Queuing theory with heavy tails and network traffic modeling. 2018. hal-01891760*

*Whitt, Ward. (2000). The impact of a heavy-tailed service-time distribution upon the  $M/GI/s$  waiting-time distribution. Queueing Syst.. 36. 71-87. 10.1023/A :1019143505968.*

## La loi de Cauchy

La loi de Cauchy est une loi continue de fonction de densité

$$f(x; L; a) = \frac{1}{\pi a \left[ 1 + \left( \frac{x-L}{a} \right)^2 \right]} = \frac{1}{\pi} \left[ \frac{a}{(x-L)^2 + a^2} \right]$$

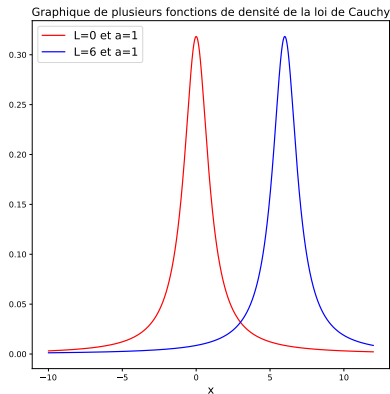
où  $L \in \mathbb{R}$  est un paramètre de localisation et  $a > 0$  est un paramètre d'échelle.

# La loi de Cauchy

La loi de Cauchy est une loi continue de fonction de densité

$$f(x; L; a) = \frac{1}{\pi a \left[ 1 + \left( \frac{x-L}{a} \right)^2 \right]} = \frac{1}{\pi} \left[ \frac{a}{(x-L)^2 + a^2} \right]$$

où  $L \in \mathbb{R}$  est un paramètre de localisation et  $a > 0$  est un paramètre d'échelle.

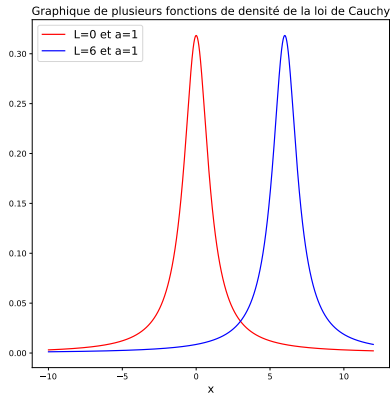


# La loi de Cauchy

La loi de Cauchy est une loi continue de fonction de densité

$$f(x; L; a) = \frac{1}{\pi a \left[ 1 + \left( \frac{x-L}{a} \right)^2 \right]} = \frac{1}{\pi} \left[ \frac{a}{(x-L)^2 + a^2} \right]$$

où  $L \in \mathbb{R}$  est un paramètre de localisation et  $a > 0$  est un paramètre d'échelle.



Soit  $X \sim \text{Cauchy}(0, 1)$  et  $Y \sim \text{Cauchy}(6, 1)$ , alors  
 $\forall t \in \mathbb{R}, \mathbb{P}[Y > t] > \mathbb{P}[X > t]$

## Les bandits de Cauchy

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Cauchy}(L_{k_t}, a)$

# Les bandits de Cauchy

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

## Les bandits de Cauchy

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^* := \max_k L_k \quad \text{et} \quad k^* := \operatorname{argmax}_k L_k$$

## Les bandits de Cauchy

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^* := \max_k L_k \quad \text{et} \quad k^* := \operatorname{argmax}_k L_k$$

le gap (regret) associé à l'action  $k$  devient  $\Delta_k = L^* - L_k$



# Les bandits de Cauchy

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Cauchy}(L_{k_t}, a)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^* := \max_k L_k \quad \text{et} \quad k^* := \operatorname{argmax}_k L_k$$

le gap (regret) associé à l'action  $k$  devient  $\Delta_k = L^* - L_k$

Mesure de performance empirique d'un agent :  $R(T) = \sum_{t=1}^T \Delta_{k_t}$

## Les algorithmes classiques : Exemple d'échec

### Expérience

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon$ -greedy et  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps.

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions.

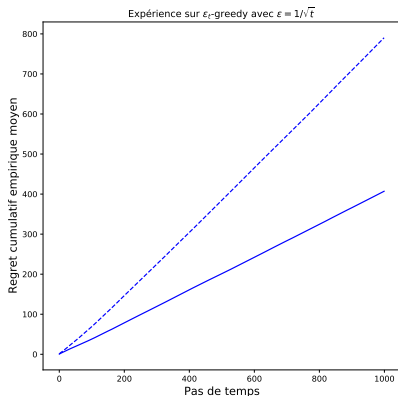
# Les algorithmes classiques : Exemple d'échec

## Expérience

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon$ -greedy et  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps.

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions.



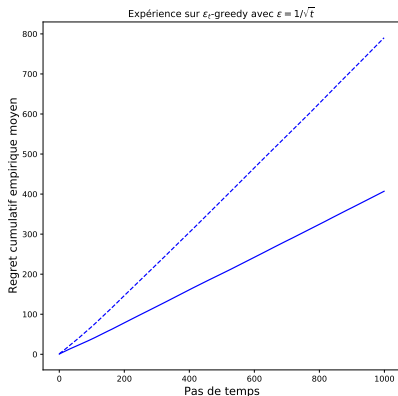
# Les algorithmes classiques : Exemple d'échec

## Expérience

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon$ -greedy et  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps.

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions.



Cause de la mauvaise performance :  $\epsilon$ -greedy base le choix de son action d'exploitation sur la moyenne empirique  $\hat{\mu}_k(t)$  des rewards reçus en jouant l'action  $k$ .

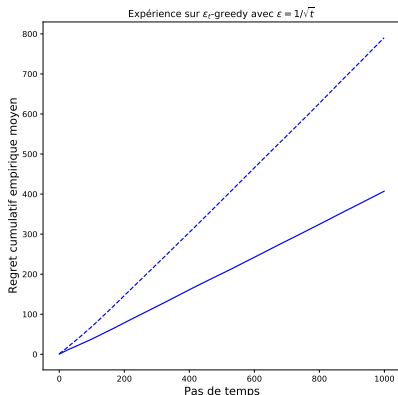
# Les algorithmes classiques : Exemple d'échec

## Expérience

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon$ -greedy et  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps.

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions.



Cause de la mauvaise performance :  $\epsilon$ -greedy base le choix de son action d'exploitation sur la moyenne empirique  $\hat{\mu}_k(t)$  des rewards reçus en jouant l'action  $k$ .

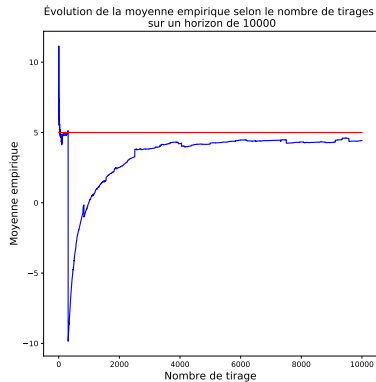
Cet estimateur (la moyenne empirique) n'estime pas bien la localisation  $L_k$  du bras numéro  $k$ .

# La non-convergence de la moyenne empirique

Comportement de la moyenne empirique sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)

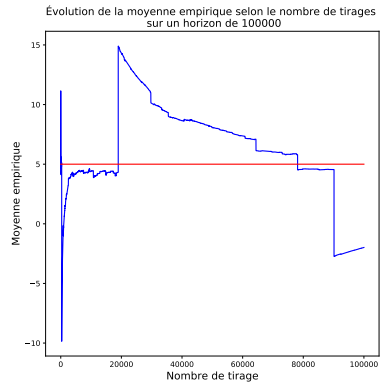
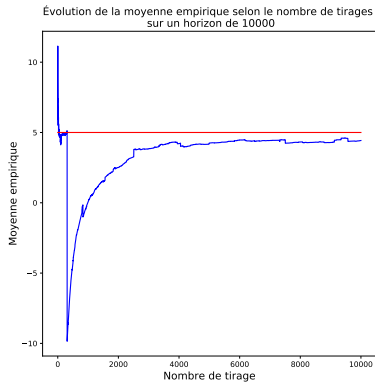
# La non-convergence de la moyenne empirique

Comportement de la moyenne empirique sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)



# La non-convergence de la moyenne empirique

Comportement de la moyenne empirique sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)





## Estimateurs de localisation d'une loi de Cauchy

Soit  $\mathcal{X} = \{X_1, X_2, \dots, X_T\}$  une séquence d'observations provenant d'une loi Cauchy( $L, a$ ) où  $L$  est inconnu et  $a$  connu.

## Estimateurs de localisation d'une loi de Cauchy

Soit  $\mathcal{X} = \{X_1, X_2, \dots, X_T\}$  une séquence d'observations provenant d'une loi Cauchy( $L, a$ ) où  $L$  est inconnu et  $a$  connu.

- La médiane empirique :  $\text{MED}(\mathcal{X})$

## Estimateurs de localisation d'une loi de Cauchy

Soit  $\mathcal{X} = \{X_1, X_2, \dots, X_T\}$  une séquence d'observations provenant d'une loi Cauchy( $L, a$ ) où  $L$  est inconnu et  $a$  connu.

- La médiane empirique :  $\text{MED}(\mathcal{X})$

- La moyenne  $\alpha$ -tronquée :  $TM_\alpha(\mathcal{X}) = \frac{1}{T-2r} \sum_{i=r+1}^{T-r} X_{(i)}$ , où  $X_{(i)}$  est la  $i^{\text{ième}}$  statistique d'ordre de  $\mathcal{X}$ .

où  $r = \lfloor n\alpha \rfloor$  où  $0 < \alpha < 0.5$

## Estimateurs de localisation d'une loi de Cauchy

Soit  $\mathcal{X} = \{X_1, X_2, \dots, X_T\}$  une séquence d'observations provenant d'une loi Cauchy( $L, a$ ) où  $L$  est inconnu et  $a$  connu.

- La médiane empirique :  $\text{MED}(\mathcal{X})$

- La moyenne  $\alpha$ -tronquée :  $TM_\alpha(\mathcal{X}) = \frac{1}{T-2r} \sum_{i=r+1}^{T-r} X_{(i)}$ , où  $X_{(i)}$  est la  $i^{\text{ième}}$  statistique d'ordre de  $\mathcal{X}$ .

où  $r = \lfloor n\alpha \rfloor$  où  $0 < \alpha < 0.5$

- L'estimateur de maximum de vraisemblance :  $\text{MLE}(\mathcal{X})$

## Estimateurs de localisation d'une loi de Cauchy

Soit  $\mathcal{X} = \{X_1, X_2, \dots, X_T\}$  une séquence d'observations provenant d'une loi Cauchy( $L, a$ ) où  $L$  est inconnu et  $a$  connu.

- La médiane empirique :  $\text{MED}(\mathcal{X})$

- La moyenne  $\alpha$ -tronquée :  $TM_\alpha(\mathcal{X}) = \frac{1}{T-2r} \sum_{i=r+1}^{T-r} X_{(i)}$ , où  $X_{(i)}$  est la  $i^{\text{ième}}$  statistique d'ordre de  $\mathcal{X}$ .

où  $r = \lfloor n\alpha \rfloor$  où  $0 < \alpha < 0.5$

- L'estimateur de maximum de vraisemblance :  $\text{MLE}(\mathcal{X})$

- L-estimator :  $\text{LE}(\mathcal{X}) = \frac{1}{T} \sum_{i=1}^T J\left(\frac{i}{T+1}\right) X_{(i)}$ , où  $X_{(i)}$  est la  $i^{\text{ième}}$  statistique d'ordre de  $\mathcal{X}$ ,

et  $J(u) = \frac{\sin(4\pi(u-0.5))}{\tan(\pi(u-0.5))}$ ,

## Estimateurs de localisation d'une loi de Cauchy

Soit  $\mathcal{X} = \{X_1, X_2, \dots, X_T\}$  une séquence d'observations provenant d'une loi Cauchy( $L, a$ ) où  $L$  est inconnu et  $a$  connu.

- La médiane empirique :  $\text{MED}(\mathcal{X})$

- La moyenne  $\alpha$ -tronquée :  $TM_\alpha(\mathcal{X}) = \frac{1}{T-2r} \sum_{i=r+1}^{T-r} X_{(i)}$ , où  $X_{(i)}$  est la  $i^{\text{ième}}$  statistique d'ordre de  $\mathcal{X}$ .

où  $r = \lfloor n\alpha \rfloor$  où  $0 < \alpha < 0.5$

- L'estimateur de maximum de vraisemblance :  $\text{MLE}(\mathcal{X})$

- L-estimator :  $\text{LE}(\mathcal{X}) = \frac{1}{T} \sum_{i=1}^T J\left(\frac{i}{T+1}\right) X_{(i)}$ , où  $X_{(i)}$  est la  $i^{\text{ième}}$  statistique d'ordre de  $\mathcal{X}$ ,

$$\text{et } J(u) = \frac{\sin(4\pi(u-0.5))}{\tan(\pi(u-0.5))},$$

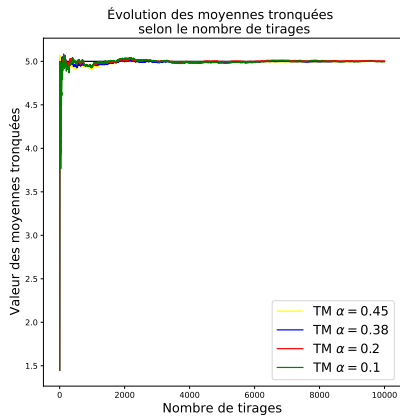
Zhang, J. A highly efficient L-estimator for the location parameter of the Cauchy distribution. *Comput Stat* 25, 97–105 (2010)

# Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)

# Comportement des estimateurs de localisation de la loi de Cauchy

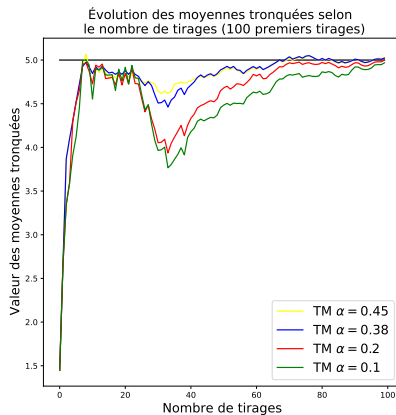
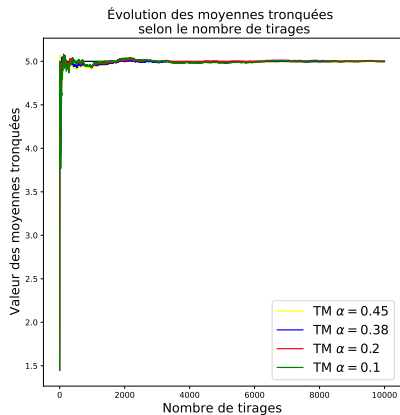
Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)





# Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)

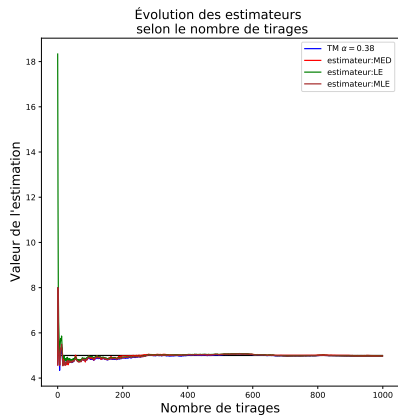


# Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des estimateurs ME, MLE, LE sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)

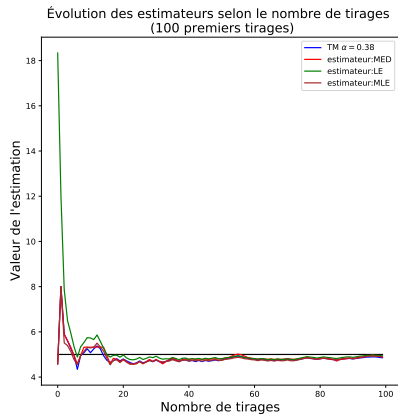
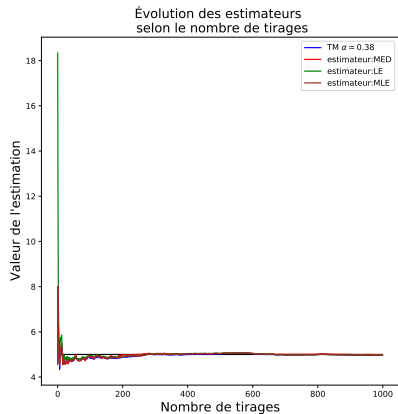
# Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des estimateurs ME, MLE, LE sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)



# Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des estimateurs ME, MLE, LE sur une séquence de réalisations provenant d'une loi Cauchy(5, 1)



# Adaptation des algorithmes existants

# Adaptation des algorithmes existants

**Exemple : adaptation de  $\epsilon$ -greedy**

# Adaptation des algorithmes existants

## Exemple : adaptation de $\epsilon$ -greedy

$\epsilon$ -greedy classique

# Adaptation des algorithmes existants

## Exemple : adaptation de $\epsilon$ -greedy

### $\epsilon$ -greedy classique

Pour tout  $t \geq 1$  :

- Explorer avec probabilité  $\epsilon$  : Sélectionner  $k_t \sim \mathcal{U}(1, 2, \dots, K)$ .
- Exploiter avec probabilité  $1 - \epsilon$  : Sélectionner  $k_t = \underset{k}{\operatorname{argmax}} \hat{\mu}_k(t-1)$



# Adaptation des algorithmes existants

## Exemple : adaptation de $\epsilon$ -greedy

### $\epsilon$ -greedy classique

Pour tout  $t \geq 1$  :

- Explorer avec probabilité  $\epsilon$  : Sélectionner  $k_t \sim \mathcal{U}(1, 2, \dots, K)$ .
- Exploiter avec probabilité  $1 - \epsilon$  : Sélectionner  $k_t = \underset{k}{\operatorname{argmax}} \hat{\mu}_k(t-1)$

### $\epsilon$ -greedy Cauchy

# Adaptation des algorithmes existants

## Exemple : adaptation de $\epsilon$ -greedy

### $\epsilon$ -greedy classique

Pour tout  $t \geq 1$  :

- Explorer avec probabilité  $\epsilon$  : Sélectionner  $k_t \sim \mathcal{U}(1, 2, \dots, K)$ .
- Exploiter avec probabilité  $1 - \epsilon$  : Sélectionner  $k_t = \underset{k}{\operatorname{argmax}} \hat{\mu}_k(t-1)$

### $\epsilon$ -greedy Cauchy

Pour tout  $t \geq 1$  :

- Explorer avec probabilité  $\epsilon$  : Sélectionner  $k_t \sim \mathcal{U}(1, 2, \dots, K)$ .
- Exploiter avec probabilité  $1 - \epsilon$  : Sélectionner  $k_t = \underset{k}{\operatorname{argmax}} \hat{L}_k(t-1)$

où  $\hat{L}_k(t-1)$  est un des estimateurs de la localisation  $L_k$  du bras no  $k$  basée sur les observations obtenus sur ce bras dans les pas de temps passés.

# Adaptation des algorithmes existants

## Exemple : adaptation de $\epsilon$ -greedy

### $\epsilon$ -greedy classique

Pour tout  $t \geq 1$  :

- Explorer avec probabilité  $\epsilon$  : Sélectionner  $k_t \sim \mathcal{U}(1, 2, \dots, K)$ .
- Exploiter avec probabilité  $1 - \epsilon$  : Sélectionner  $k_t = \underset{k}{\operatorname{argmax}} \hat{\mu}_k(t-1)$

### $\epsilon$ -greedy Cauchy

Pour tout  $t \geq 1$  :

- Explorer avec probabilité  $\epsilon$  : Sélectionner  $k_t \sim \mathcal{U}(1, 2, \dots, K)$ .
- Exploiter avec probabilité  $1 - \epsilon$  : Sélectionner  $k_t = \underset{k}{\operatorname{argmax}} \hat{L}_k(t-1)$

où  $\hat{L}_k(t-1)$  est un des estimateurs de la localisation  $L_k$  du bras no  $k$  basée sur les observations obtenus sur ce bras dans les pas de temps passés.

De la même façon, on peut adapter aisément les  $\epsilon_t$ -greedy, ETC et Boltzmann/Softmax

## Expérience 1 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

## Expérience 1 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

Pour chacune des  $N = 200$  répétitions,

## Expérience 1 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .

## Expérience 1 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter  $\epsilon_t$ -greedy)

## Expérience 1 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter  $\epsilon_t$ -greedy)

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions pour comparer les différentes versions de l'algorithme.

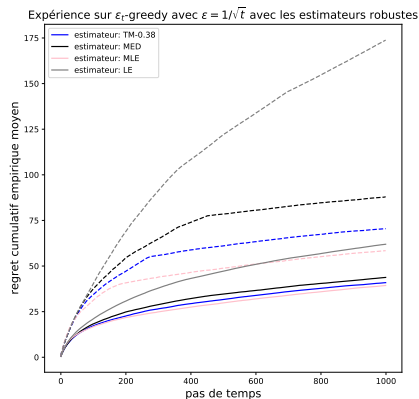


## Expérience 1 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter  $\epsilon_t$ -greedy)

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions pour comparer les différentes versions de l'algorithme.

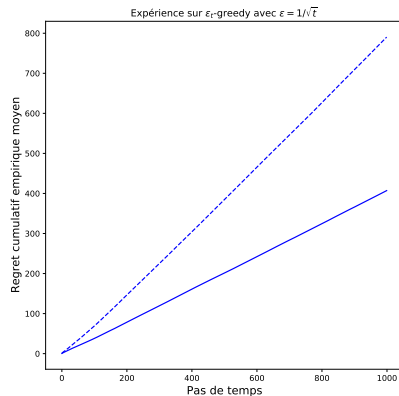
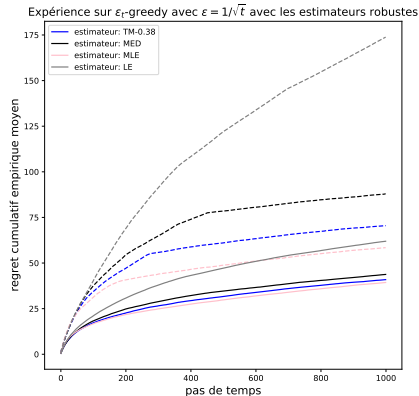


## Expérience 1 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(5, 1)$  et  $\text{Cauchy}(6, 1)$ .
- Jouer  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter  $\epsilon_t$ -greedy)

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions pour comparer les différentes versions de l'algorithme.

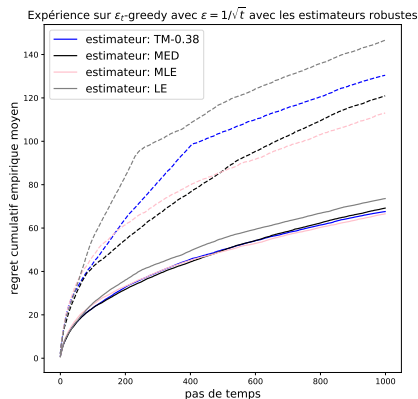


## Expérience 2 avec $\epsilon_t$ -greedy avec $\epsilon_t = 1/\sqrt{t}$ sur des bandits Cauchy

Pour chacune des  $N = 200$  répétitions,

- Créer un bandit Cauchy à deux bras de distributions  $\text{Cauchy}(L_1, 1)$  et  $\text{Cauchy}(L_2, 1)$  avec  $L_1, L_2 \sim \mathcal{U}([0, 5])$
- Jouer  $\epsilon_t$ -greedy sur un horizon de  $T = 1000$  pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter  $\epsilon_t$ -greedy)

Tracer le regret cumulatif empirique moyenné sur les  $N$  répétitions pour comparer les différentes versions de l'algorithme.



## La loi de Pareto

La loi de Pareto est une loi continue dont la fonction de densité est donnée par

$$f(x; L, a) = \begin{cases} \frac{aL^a}{x^{a+1}} & \text{si } x \geq L \\ 0 & \text{sinon} \end{cases}$$

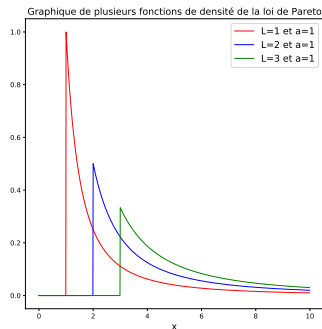
# La loi de Pareto

La loi de Pareto est une loi continue dont la fonction de densité est donnée par

$$f(x; L, a) = \begin{cases} \frac{aL^a}{x^{a+1}} & \text{si } x \geq L \\ 0 & \text{sinon} \end{cases}$$

Dans le cas particulier où  $a = 1$ , l'espérance est non-définie et la fonction de densité est

$$f(x; L) = \begin{cases} \frac{L}{x^2} & \text{si } x \geq L \\ 0 & \text{sinon} \end{cases}$$



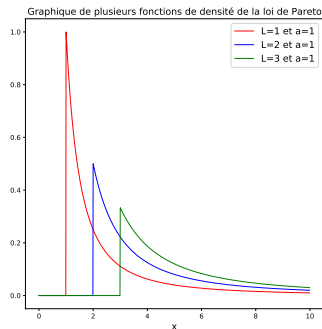
# La loi de Pareto

La loi de Pareto est une loi continue dont la fonction de densité est donnée par

$$f(x; L, a) = \begin{cases} \frac{aL^a}{x^{a+1}} & \text{si } x \geq L \\ 0 & \text{sinon} \end{cases}$$

Dans le cas particulier où  $a = 1$ , l'espérance est non-définie et la fonction de densité est

$$f(x; L) = \begin{cases} \frac{L}{x^2} & \text{si } x \geq L \\ 0 & \text{sinon} \end{cases}$$



Remarque : particularité de la loi de Pareto, pour le cas  $a = 1$ , on a que la médiane de la distribution est  $2L$ .

## Les bandits de Pareto avec $a = 1$

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Pareto}(L_{k_t}, 1)$

## Les bandits de Pareto avec $a = 1$

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Pareto}(L_{k_t}, 1)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :



## Les bandits de Pareto avec $a = 1$

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Pareto}(L_{k_t}, 1)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^* := \max_k L_k \quad \text{et} \quad k^* := \operatorname{argmax}_k L_k$$

## Les bandits de Pareto avec $a = 1$

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Pareto}(L_{k_t}, 1)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^* := \max_k L_k \quad \text{et} \quad k^* := \operatorname{argmax}_k L_k$$

le gap (regret) associé à l'action  $k$  devient  $\Delta_k = L^* - L_k$

## Les bandits de Pareto avec $a = 1$

À chaque pas de temps  $t = 1, 2, \dots, T$ , l'agent :

- Sélectionne une action  $k_t \in \{1, 2, \dots, K\}$
- Observe une reward  $r_t \sim \text{Pareto}(L_{k_t}, 1)$

L'action optimale et la localisation optimale sont définis à partir de la localisation des différents bras :

$$L^* := \max_k L_k \quad \text{et} \quad k^* := \operatorname{argmax}_k L_k$$

le gap (regret) associé à l'action  $k$  devient  $\Delta_k = L^* - L_k$

Mesure de performance empirique d'un agent :  $R(T) = \sum_{t=1}^T \Delta_{k_t}$

## Estimateur de la localisation $L$ d'une loi de Pareto et résultats préliminaires

Un estimateur naturel pour  $L$  à partir d'un jeu de données  $\mathcal{X} = \{X_1, X_2, X_3, \dots, X_T\}$  tirées d'une loi  $Pareto(L, 1)$  est  $\hat{L} = \min(\mathcal{X})$  ou encore  $\hat{L} = \frac{1}{2}MED(\mathcal{X})$

## Estimateur de la localisation $L$ d'une loi de Pareto et résultats préliminaires

Un estimateur naturel pour  $L$  à partir d'un jeu de données  $\mathcal{X} = \{X_1, X_2, X_3, \dots, X_T\}$  tirées d'une loi  $Pareto(L, 1)$  est  $\hat{L} = \min(\mathcal{X})$  ou encore  $\hat{L} = \frac{1}{2}MED(\mathcal{X})$

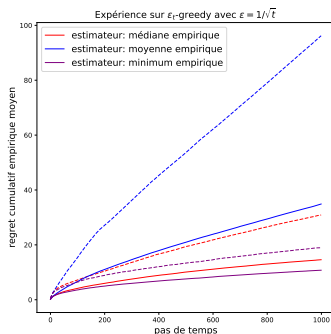
On peut donc généraliser les algorithmes classiques comme  $\epsilon$ -greedy, ETC, Boltzmann/Softmax

## Estimateur de la localisation $L$ d'une loi de Pareto et résultats préliminaires

Un estimateur naturel pour  $L$  à partir d'un jeu de données  $\mathcal{X} = \{X_1, X_2, X_3, \dots, X_T\}$  tirées d'une loi  $Pareto(L, 1)$  est  $\hat{L} = \min(\mathcal{X})$  ou encore  $\hat{L} = \frac{1}{2}MED(\mathcal{X})$

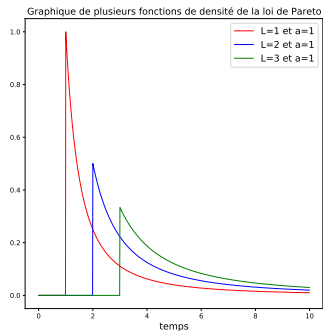
On peut donc généraliser les algorithmes classiques comme  $\epsilon$ -greedy, ETC, Boltzmann/Softmax

Voici le résultat d'une expérience sur  $N = 200$  instances de bandits Pareto à deux bras avec  $L_1, L_2 \sim \mathcal{U}([0, 1])$  et  $a = 1$ , où l'on a joué  $\epsilon_t$ -greedy avec  $\epsilon_t = 1/\sqrt{t}$  sur un horizon de  $T = 1000$  pas de temps.



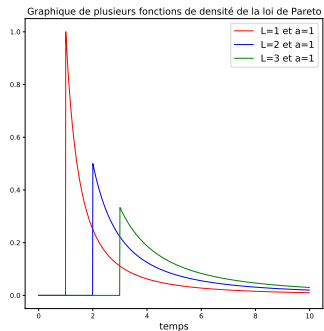


- Modélisation d'un problème concret.

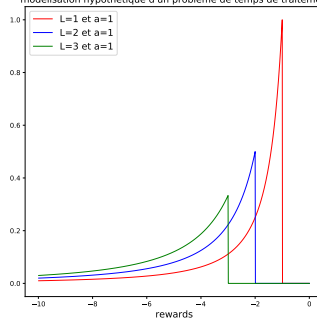




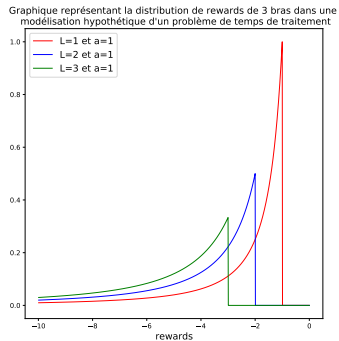
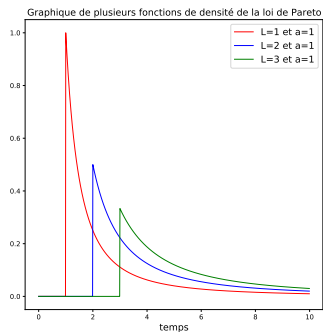
- Modélisation d'un problème concret.



Graphique représentant la distribution de rewards de 3 bras dans une modélisation hypothétique d'un problème de temps de traitement

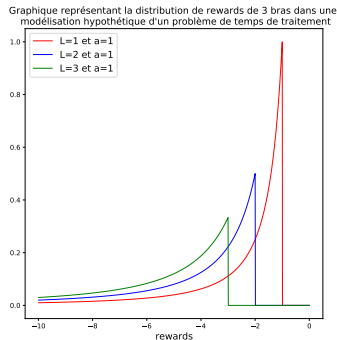
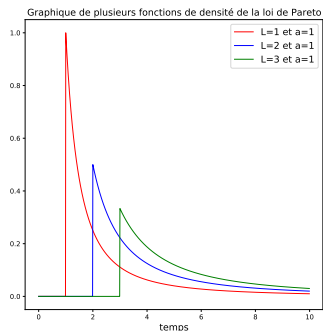


- Modélisation d'un problème concret.



- Extension à des algorithmes plus complexes, par exemple les UCB, Thompson sampling, etc...  
*S. Bubeck, N. Cesa-Bianchi and G. Lugosi, "Bandits With Heavy Tail," in IEEE Transactions on Information Theory, vol. 59, no. 11, pp. 7711-7717, Nov. 2013, doi : 10.1109/TIT.2013.2277869.*

- Modélisation d'un problème concret.



- Extension à des algorithmes plus complexes, par exemple les UCB, Thompson sampling, etc...  
*S. Bubeck, N. Cesa-Bianchi and G. Lugosi, "Bandits With Heavy Tail," in IEEE Transactions on Information Theory, vol. 59, no. 11, pp. 7711-7717, Nov. 2013, doi : 10.1109/TIT.2013.2277869.*
- Généralisation du problème (définition de la mesure d'une performance générale qui ne dépend pas de la loi), possiblement basée sur des notions de dominance entre des variables aléatoires.