Les bandits stochastiques à récompenses d'espérance non-définie

Adam Cohen, Maxime Genest, Vincent Masse

27 novembre 2020

Rappel sur les bandits stochastiques classiques

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

Rappel sur les bandits stochastiques classiques

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

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

- Sélectionne une action $k_t \in \{1, 2, \ldots K\}$
- 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 $t = 1, 2, \ldots, T$, l'agent cumule un regret :

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

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 distributions de probabilité ont une distribution d'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 distributions de probabilité ont une distribution d'espérance non-définie.

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

La loi de Cauchy

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

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

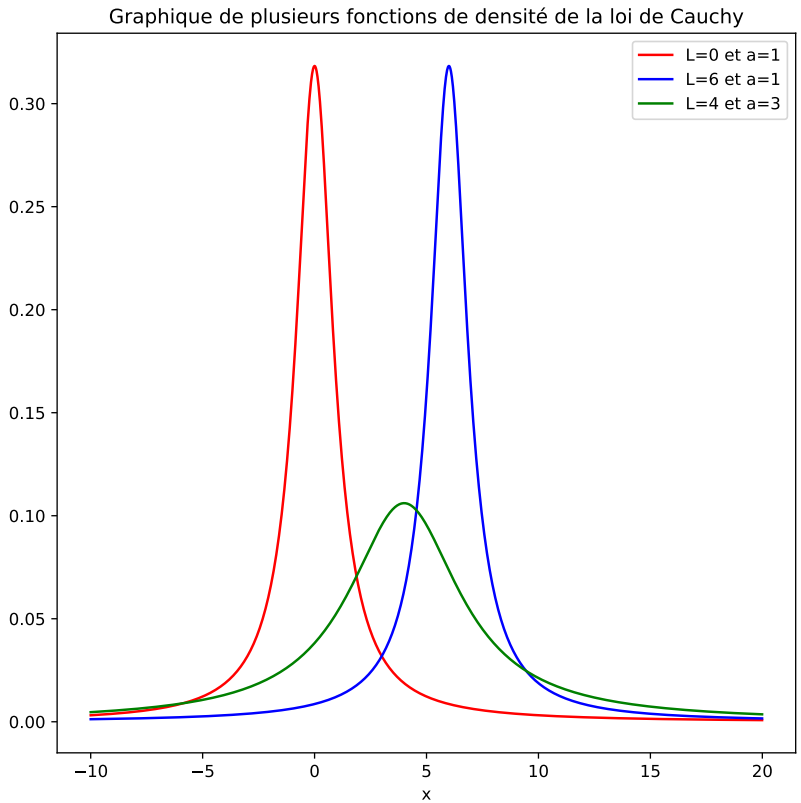
où $L \in \mathbb{R}$ est un paramètre de localisation et $a > 0$ est un paramètre d'échelle.

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.



Les bandits de Cauchy

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

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

Les bandits de Cauchy

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

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

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 $t = 1, 2, \ldots, T$, l'agent :

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

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

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

Les bandits de Cauchy

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

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

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

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

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

Les bandits de Cauchy

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

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

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

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

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

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 $\mathrm{Cauchy}(5, 1)$ et $\mathrm{Cauhy}(6, 1)$.
- Jouer ϵ-greedy et ϵ_t-greedy sur un horizon de $T = 1000$ pas de temps.

Tracer le regret cumulatif empirique moyenné sur les N répétitions.

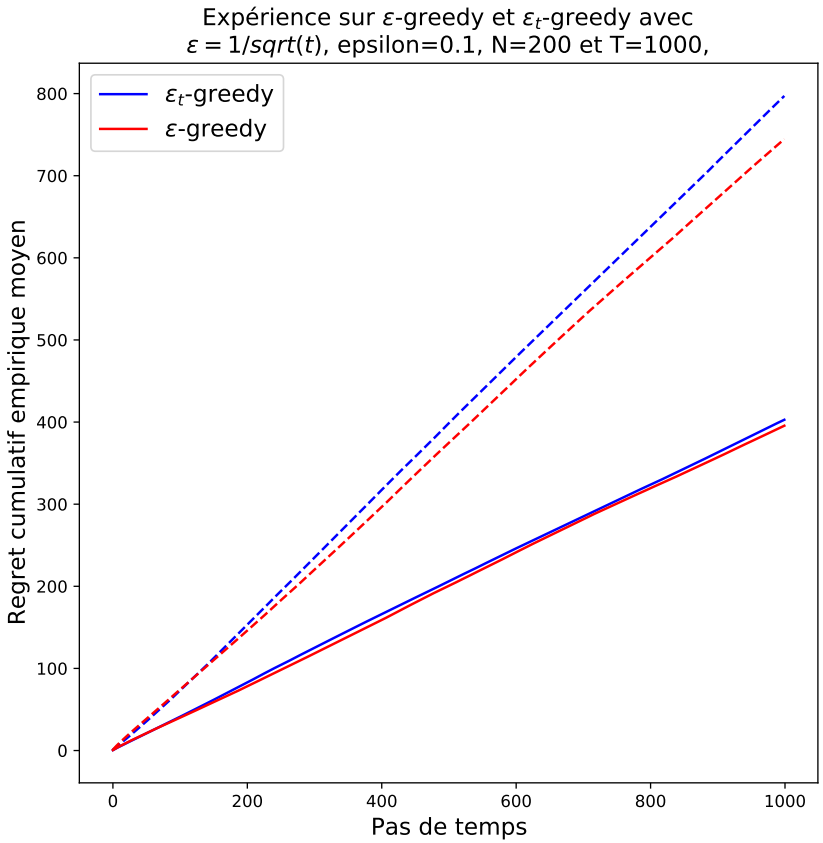
Les algorithmes classiques : Exemple d'échec

Expérience

Pour chacune des $N = 200$ répétitions,

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

Tracer le regret cumulatif empirique moyenné sur les N répétitions.



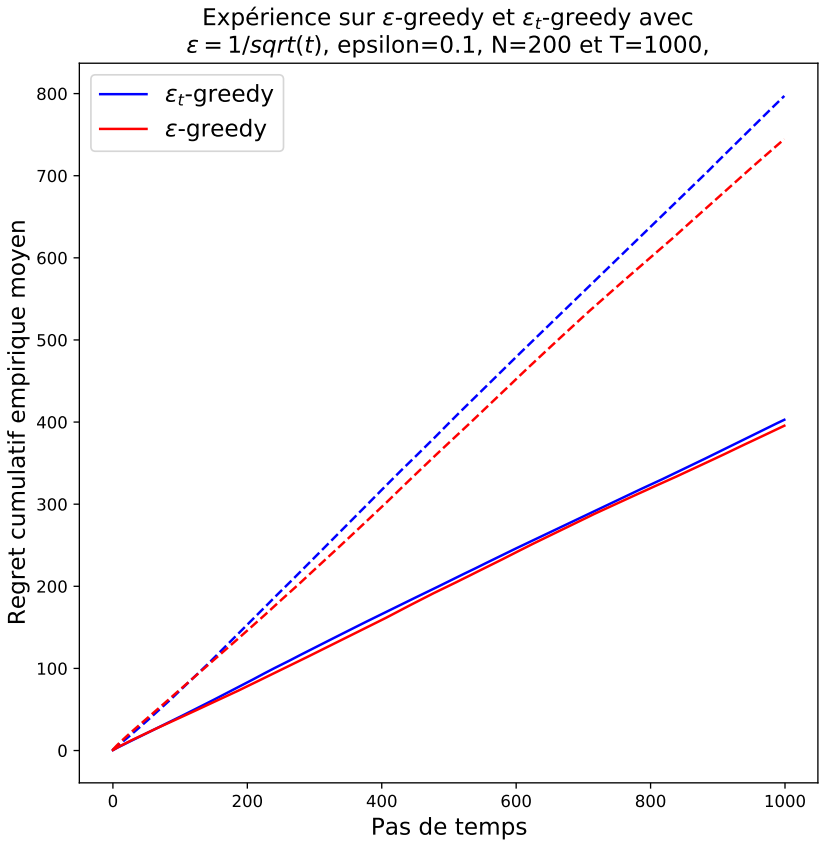
Les algorithmes classiques : Exemple d'échec

Expérience

Pour chacune des $N = 200$ répétitions,

- Créer un bandit Cauchy à deux bras de distributions $\mathrm{Cauchy}(5, 1)$ et $\mathrm{Cauhy}(6, 1)$.
- Jouer ϵ-greedy et ϵ_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 : ϵ-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.

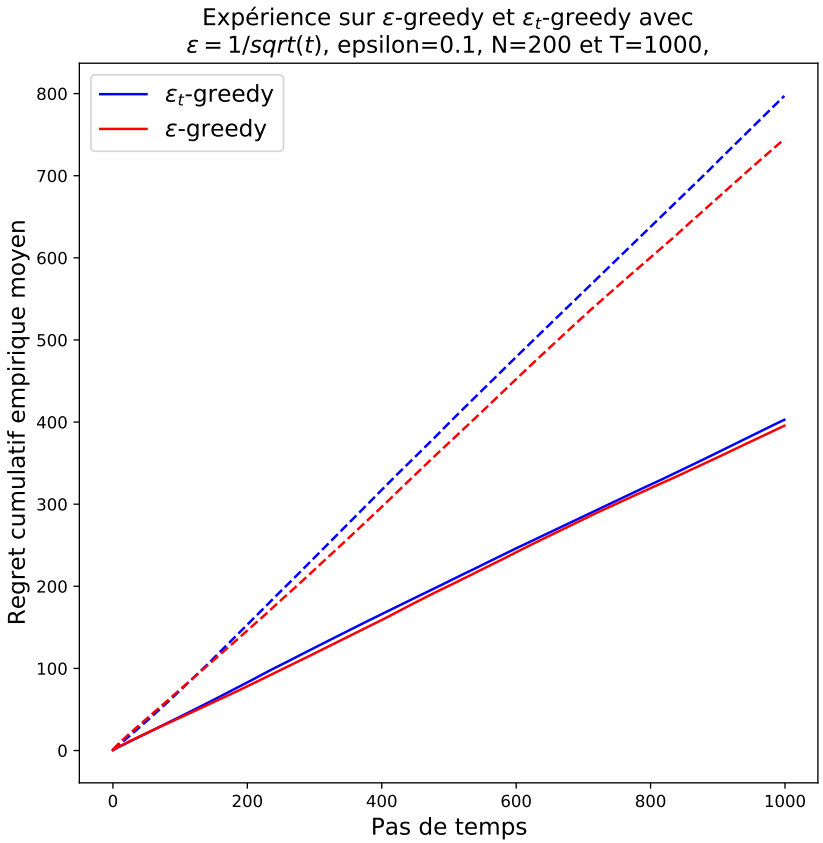
Les algorithmes classiques : Exemple d'échec

Expérience

Pour chacune des $N = 200$ répétitions,

- Créer un bandit Cauchy à deux bras de distributions $\mathrm{Cauchy}(5, 1)$ et $\mathrm{Cauhy}(6, 1)$.
- Jouer ϵ-greedy et ϵ_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 : ϵ-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.

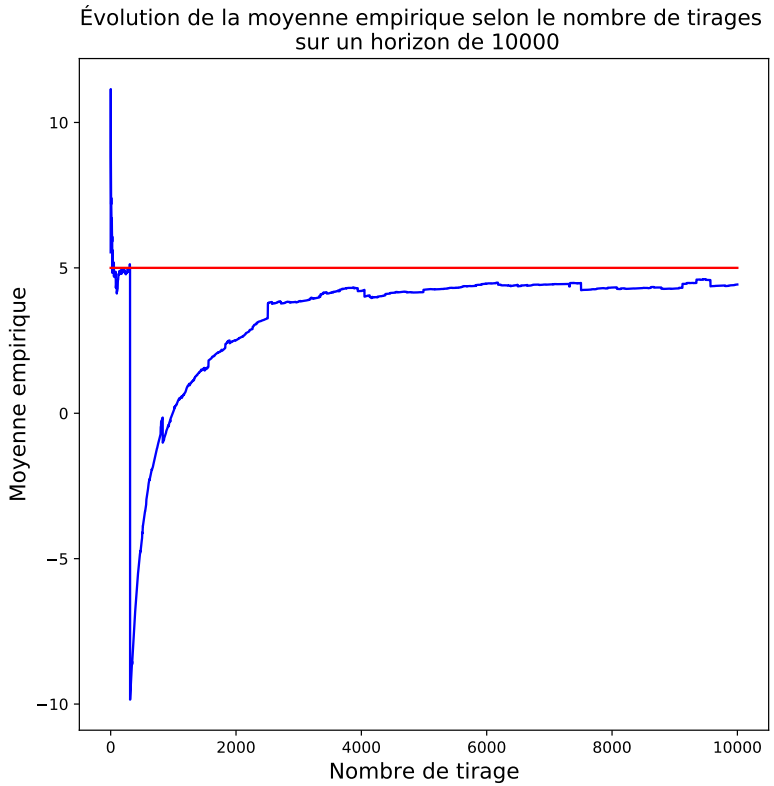
Cet estimateur (la moyenne empirique) n'estime pas bien la localisation L_k du bras numéro k.

La non-convergence de la moyenne empirique

Comportement de la moyenne empirique sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$

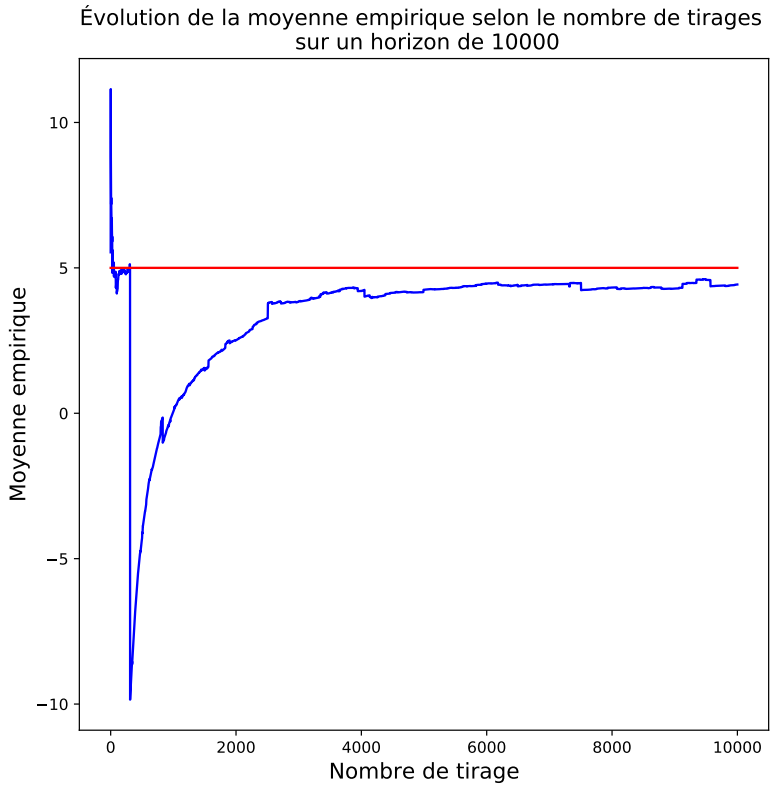
La non-convergence de la moyenne empirique

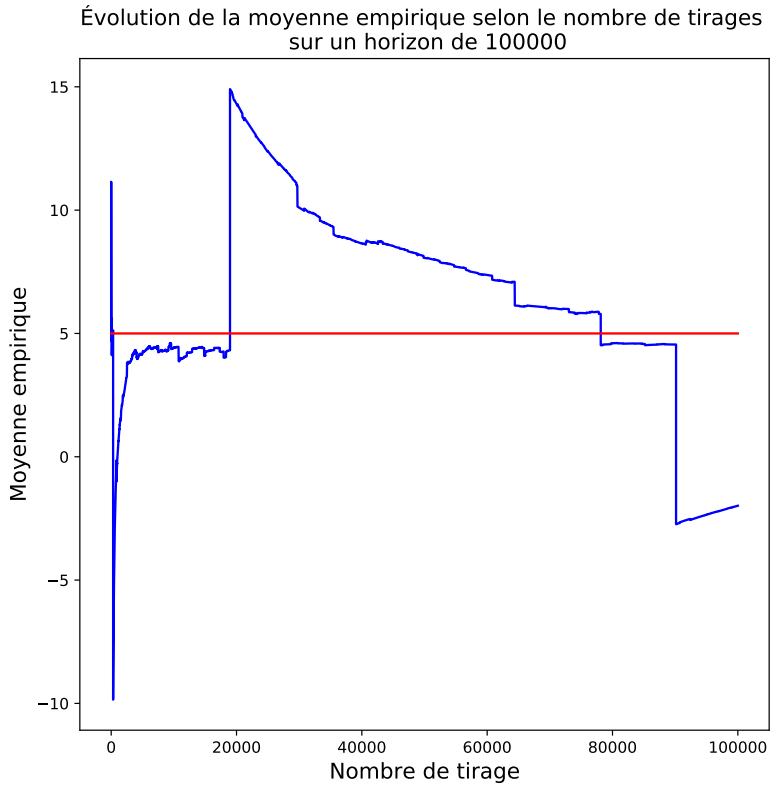
Comportement de la moyenne empirique sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$



La non-convergence de la moyenne empirique

Comportement de la moyenne empirique sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$





Estimateurs de localisation d'une loi de Cauchy

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

Estimateurs de localisation d'une loi de Cauchy

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

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

Estimateurs de localisation d'une loi de Cauchy

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

- La médiane empirique : $\mathrm{MED}(\mathcal{X})$
- La moyenne $\alpha-$tronquée : $TM_\alpha(\mathcal{X}) = \dfrac{1}{T-2r} \displaystyle\sum_{i=r+1}^{T-r} X_{(i)}$

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

Estimateurs de localisation d'une loi de Cauchy

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

- La médiane empirique : $\mathrm{MED}(\mathcal{X})$
- La moyenne $\alpha-$tronquée : $TM_\alpha(\mathcal{X}) = \frac{1}{T-2r} \sum_{i=r+1}^{T-r} X_{(i)}$

  où $r = \lfloor n\alpha \rfloor$ où $0 < \alpha < 0.5$
- L'estimateur de maximum de vraisemblance : $\mathrm{MLE}(\mathcal{X})$

Estimateurs de localisation d'une loi de Cauchy

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

- La médiane empirique : $\mathrm{MED}(\mathcal{X})$
- La moyenne $\alpha-$tronquée : $TM_\alpha(\mathcal{X}) = \dfrac{1}{T-2r}\displaystyle\sum_{i=r+1}^{T-r} X_{(i)}$

  où $r = \lfloor n\alpha \rfloor$ où $0 < \alpha < 0.5$
- L'estimateur de maximum de vraisemblance : $\mathrm{MLE}(\mathcal{X})$
- L-estimator : $\mathrm{LE}(\mathcal{X}) = \dfrac{1}{T}\displaystyle\sum_{i=1}^{T} J\left(\frac{i}{T+1}\right) X_{(i)}$

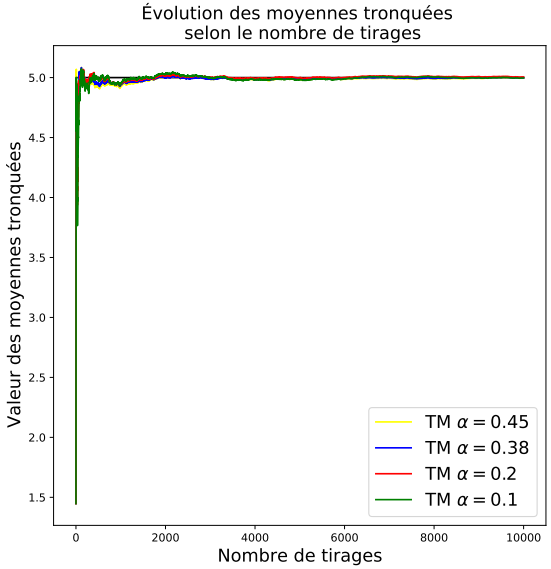
  avec $J(u) = \frac{\sin(4\pi(u-0.5)}{\tan(\pi(u-0.5)}$

Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$

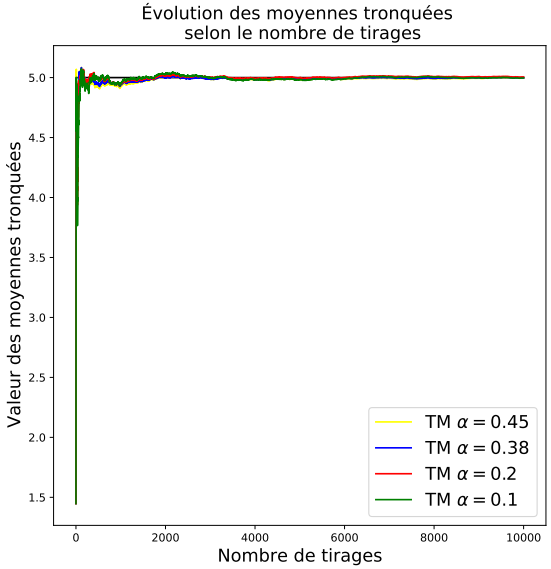
Comportement des estimateurs de localisation de la loi de Cauchy

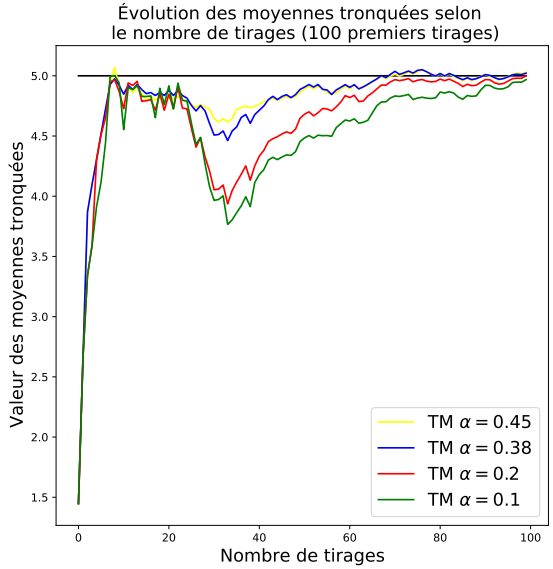
Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$



Comportement des estimateurs de localisation de la loi de Cauchy

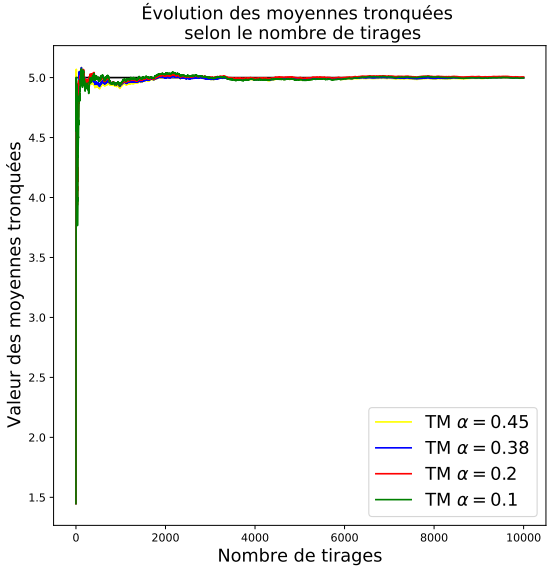
Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$

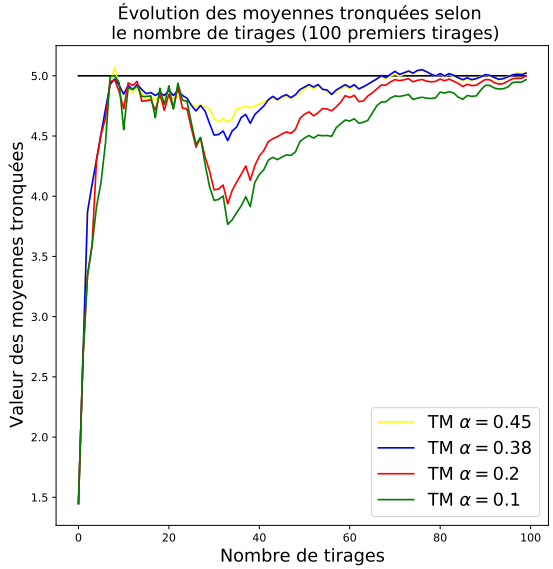


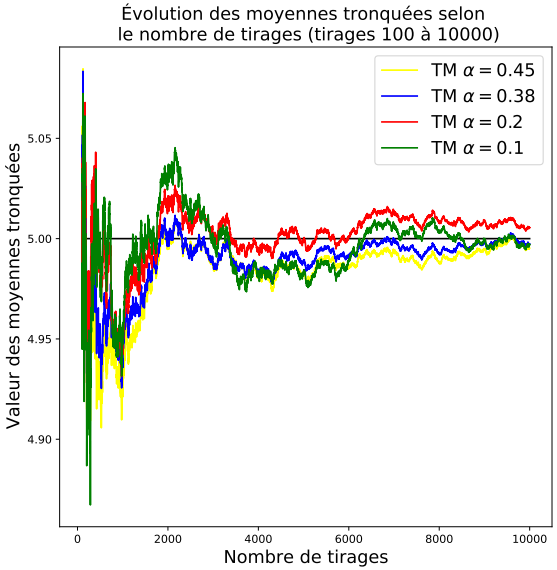


Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$

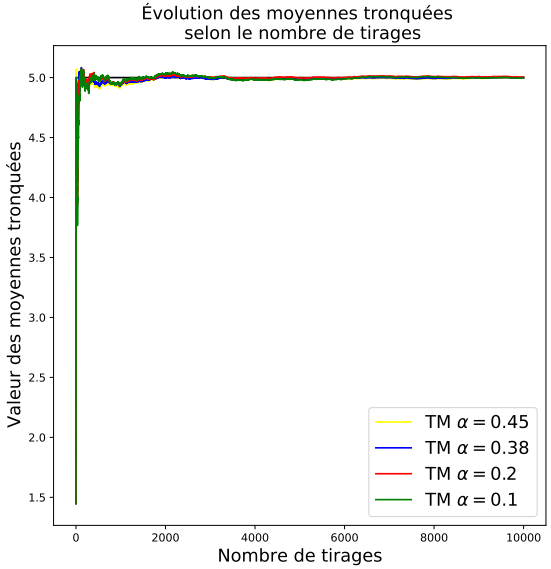


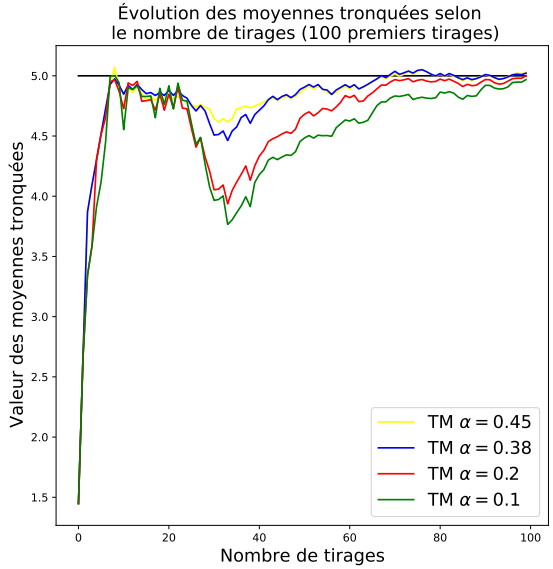




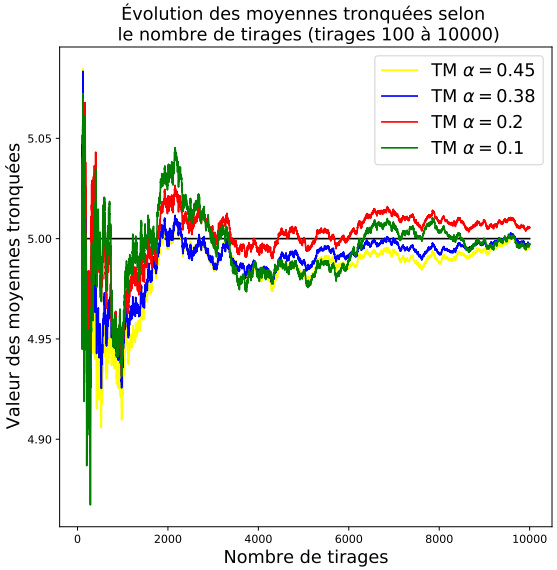
Comportement des estimateurs de localisation de la loi de Cauchy

Comportement des moyennes tronquées sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$





BlaBla

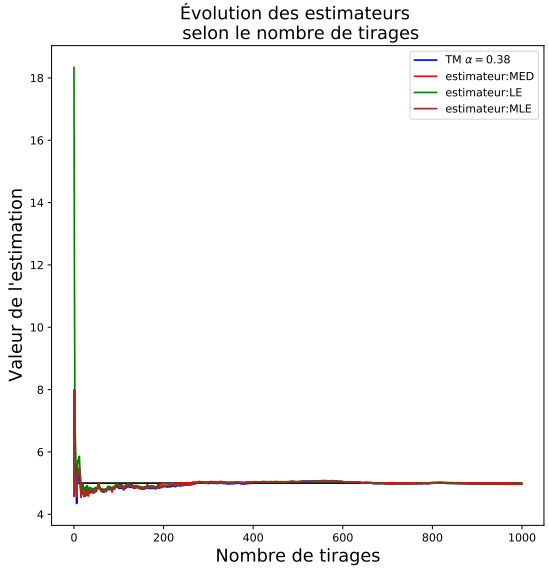


Comportement des estimateurs de localisation de la loi de Cauchy

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

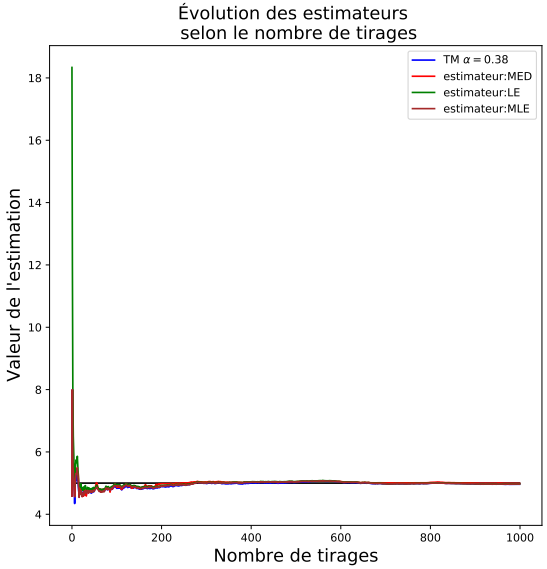
Comportement des estimateurs de localisation de la loi de Cauchy

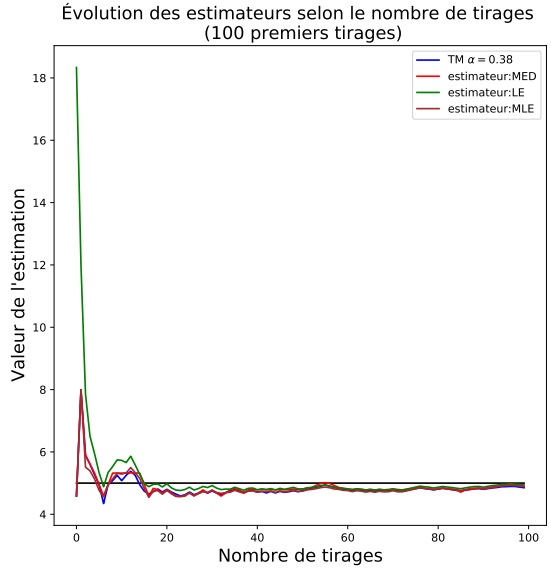
Comportement des estimateurs ME, MLE, LE sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$



Comportement des estimateurs de localisation de la loi de Cauchy

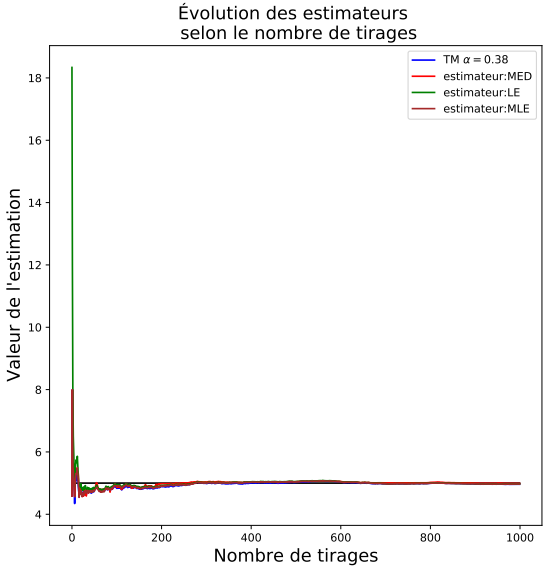
Comportement des estimateurs ME, MLE, LE sur une séquence de réalisations provenant d'une loi $\mathrm{Cauchy}(5, 1)$

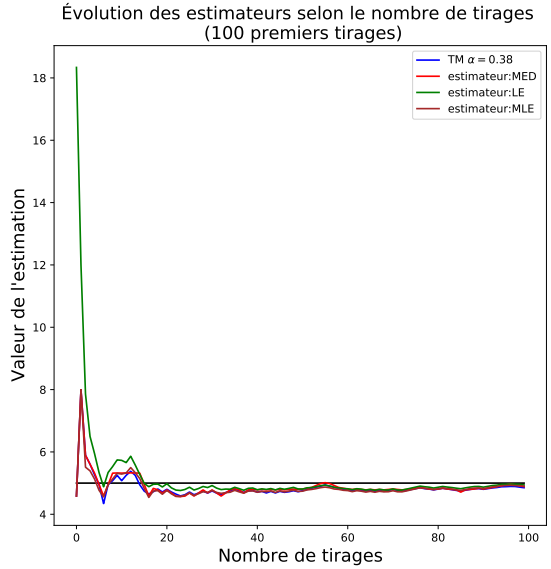


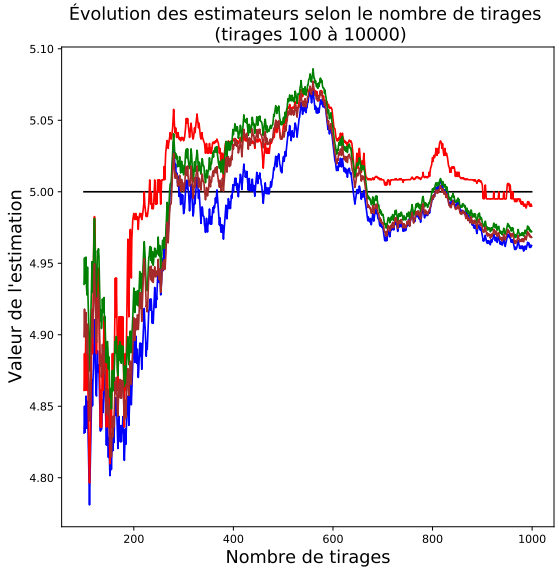


Comportement des estimateurs de localisation de la loi de Cauchy

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

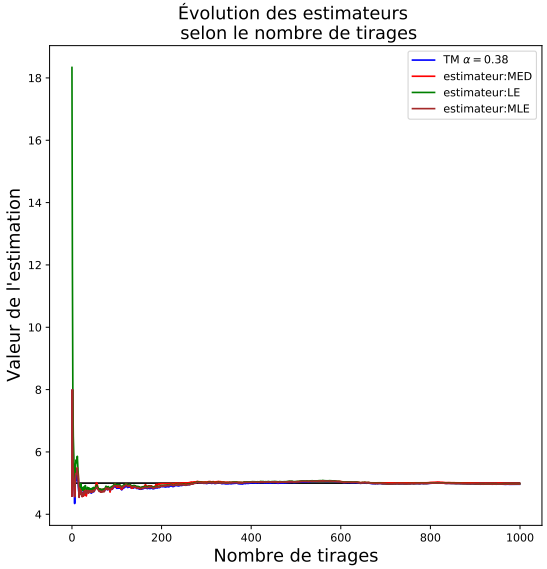


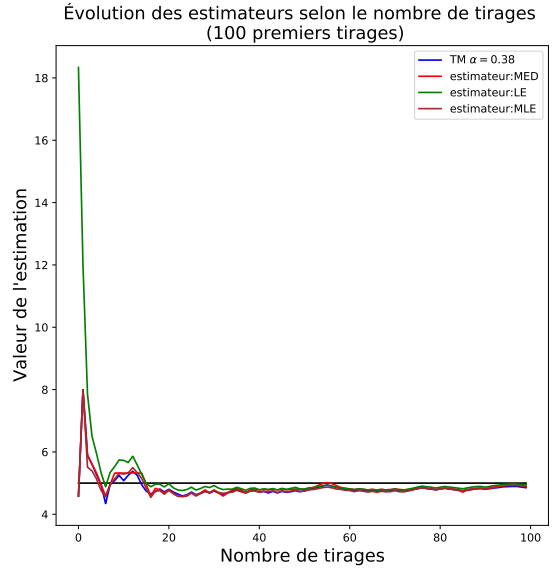




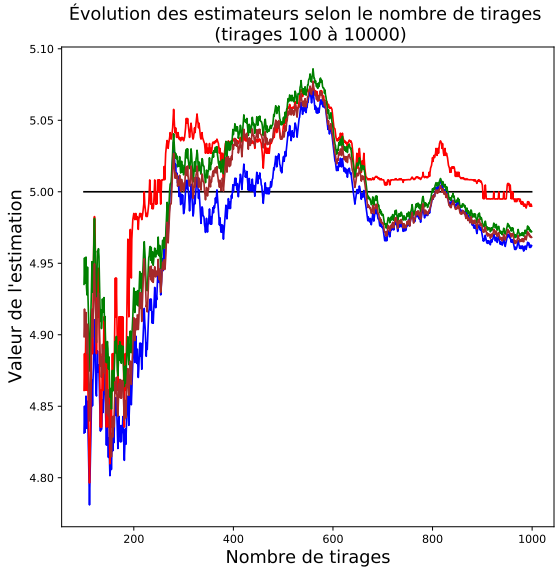
Comportement des estimateurs de localisation de la loi de Cauchy

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





BlaBla



Adaptation des algorithmes existants

Adaptation des algorithmes existants

Exemple : adaptation de ϵ-greedy

Adaptation des algorithmes existants

Exemple : adaptation de ϵ-greedy

ϵ-greedy classique

Adaptation des algorithmes existants

Exemple : adaptation de ϵ-greedy

ϵ-greedy classique

Pour tout $t \geq 1$:

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

Adaptation des algorithmes existants

Exemple : adaptation de ϵ-greedy

ϵ-greedy classique

Pour tout $t \geq 1$:

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

ϵ-greedy Cauchy

Adaptation des algorithmes existants

Exemple : adaptation de ϵ-greedy

ϵ-greedy classique

Pour tout $t \geq 1$:

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

ϵ-greedy Cauchy

Pour tout $t \geq 1$:

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

où $\widehat{L_k}(t-1)$ 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 ϵ-greedy

ϵ-greedy classique

Pour tout $t \geq 1$:

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

ϵ-greedy Cauchy

Pour tout $t \geq 1$:

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

où $\widehat{L_k}(t-1)$ 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 ϵ_t-greedy, ETC et Boltzmann/Softmax

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

Pour chacune des $N = 200$ répétitions,

Expérience 1 avec ϵ_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 $\mathrm{Cauchy}(5, 1)$ et $\mathrm{Cauhy}(6, 1)$.

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

Pour chacune des $N = 200$ répétitions,

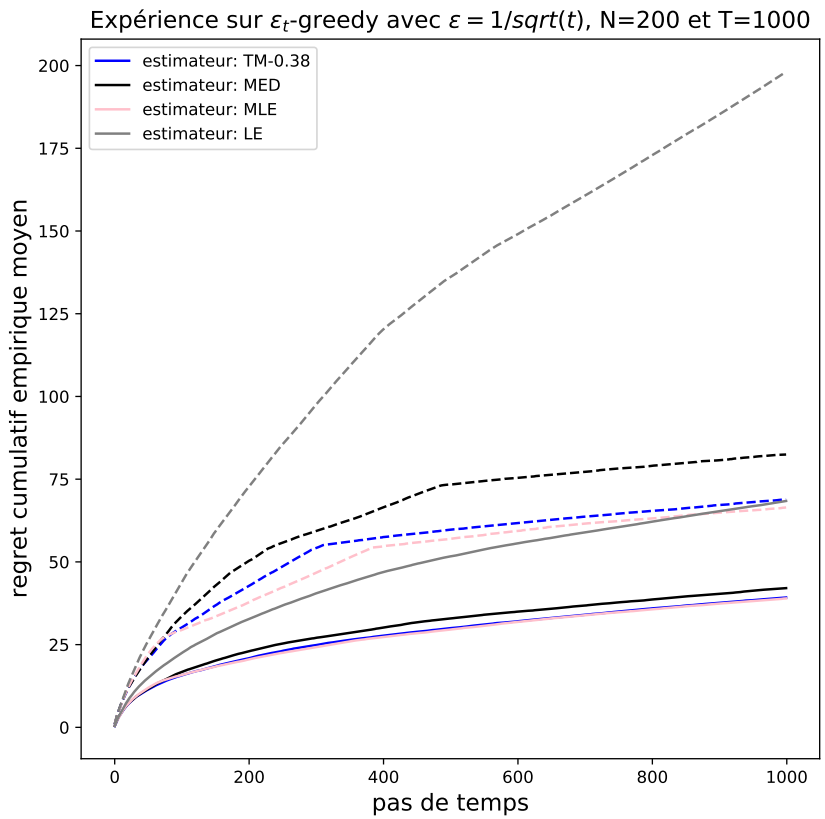
- Créer un bandit Cauchy à deux bras de distributions $\mathrm{Cauchy}(5, 1)$ et $\mathrm{Cauhy}(6, 1)$.
- Jouer ϵ_t-greedy sur un horizon de $T = 1000$ pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter ϵ_t-greedy)

Expérience 1 avec ϵ_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 $\mathrm{Cauchy}(5, 1)$ et $\mathrm{Cauhy}(6, 1)$.
- Jouer ϵ_t-greedy sur un horizon de $T = 1000$ pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter ϵ_t-greedy)

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



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

Pour chacune des $N = 200$ répétitions,

Expérience 2 avec ϵ_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 $\mathrm{Cauchy}(L_1, 1)$ et $\mathrm{Cauhy}(L_2, 1)$ avec $L_1, L_2 \sim \mathcal{U}([0, 5])$

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

Pour chacune des $N = 200$ répétitions,

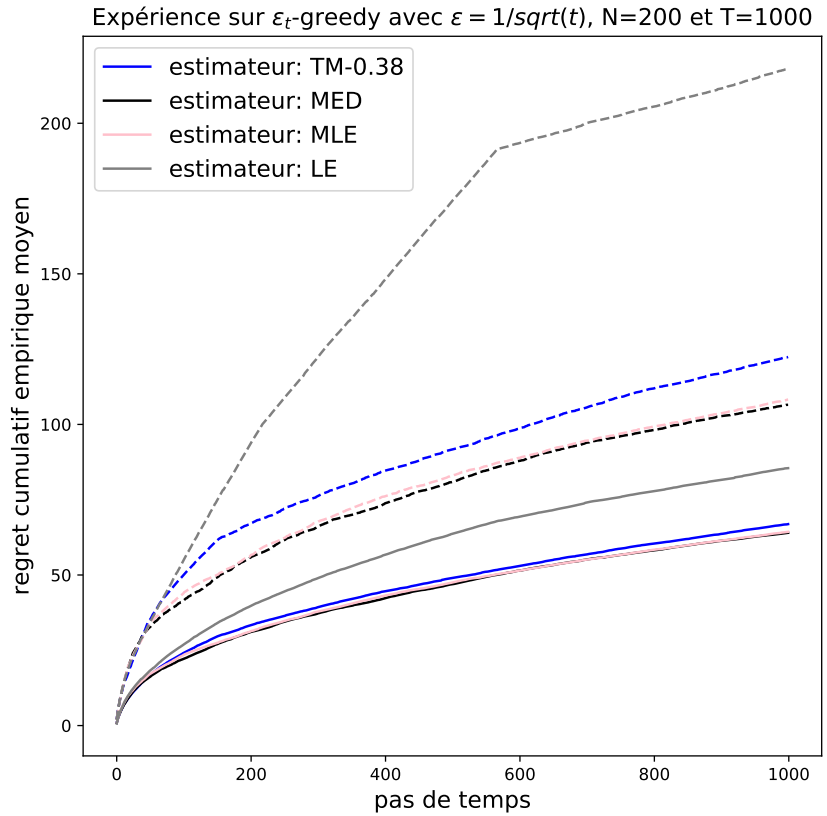
- Créer un bandit Cauchy à deux bras de distributions $\mathrm{Cauchy}(L_1, 1)$ et $\mathrm{Cauhy}(L_2, 1)$ avec $L_1, L_2 \sim \mathcal{U}([0, 5])$
- Jouer ϵ_t-greedy sur un horizon de $T = 1000$ pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter ϵ_t-greedy)

Expérience 2 avec ϵ_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 $\mathrm{Cauchy}(L_1, 1)$ et $\mathrm{Cauhy}(L_2, 1)$ avec $L_1, L_2 \sim \mathcal{U}([0, 5])$
- Jouer ϵ_t-greedy sur un horizon de $T = 1000$ pas de temps (refaire avec plusieurs estimateurs de localisation pour adapter ϵ_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

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

La loi de Pareto

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

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

Dans le cas particulier où $a = 1$, on obtient que

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

La loi de Pareto

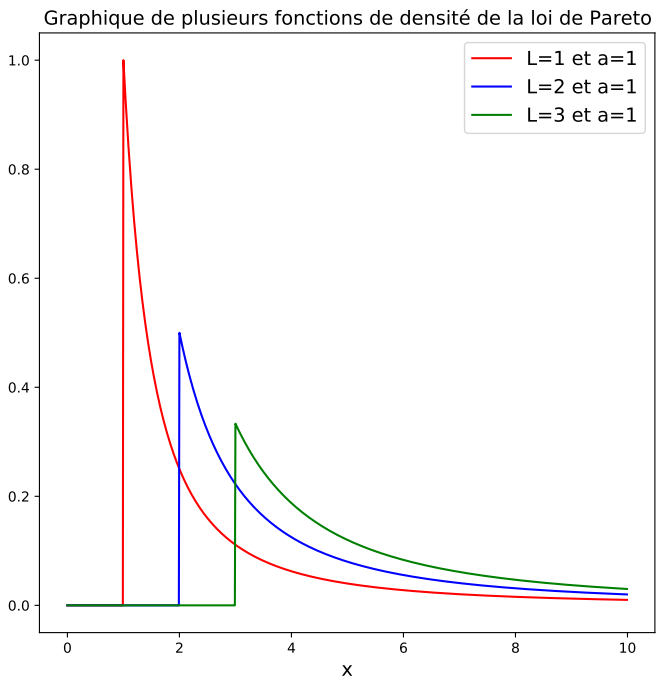
La loi de Pareto est une loi continue dont la fonction de densité est donnée par

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

Dans le cas particulier où $a = 1$, on obtient que

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

Dans ce cas particulier, la loi de Pareto possède une espérance non-définie (infinie).



La loi de Pareto

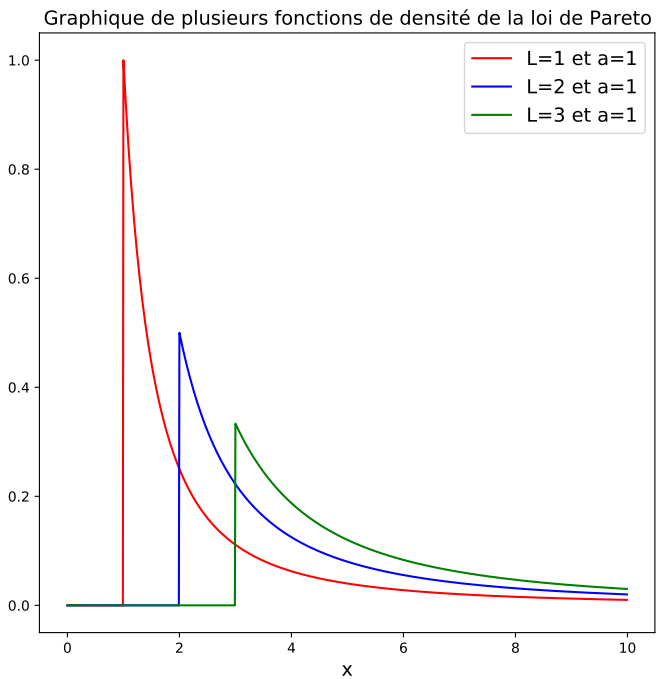
La loi de Pareto est une loi continue dont la fonction de densité est donnée par

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

Dans le cas particulier où $a = 1$, on obtient que

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

Dans ce cas particulier, la loi de Pareto possède une espérance non-définie (infinie).



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

Les bandits de Pareto avec $a = 1$

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

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

Les bandits de Pareto avec $a = 1$

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

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

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 $t = 1, 2, \ldots, T$, l'agent :

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

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

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

Les bandits de Pareto avec $a = 1$

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

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

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

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

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

Les bandits de Pareto avec $a = 1$

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

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

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

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

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

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

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

Estimateur de la localisation L d'une loi de Pareto

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

On peut donc généraliser les algorithmes classiques comme ϵ-greedy, ETC, Boltzmann/Softmax

Estimateur de la localisation L d'une loi de Pareto

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

On peut donc généraliser les algorithmes classiques comme ϵ-greedy, ETC, Boltzmann/Softmax

Voici le résultat d'une expérience sur $N = 200$ instances de bandits Pareto avec $L \sim \mathcal{U}([0, 1])$ et $a = 1$, où l'on a joué ϵ_t-greedy avec $\epsilon_t = 1/\sqrt{t}$