Les bandits stochastiques à récompenses d'espérance non-définie

A Preprint

Maxime Genest
maxime.genest.1@ulaval.ca

Vincent Masse
vincent.masse.4@ulaval.ca

Adam Cohen
adcoh4@ulaval.ca

20 décembre 2020

Abstract

Le problème classique des bandits stochastiques est défini dans la littérature avec des bras de distributions d'espérance définie. Nous souhaitons ici proposer une extension à des distributions de queues très lourdes, d'espérance non-définie. Pour ce faire, nous définissons des mesures de performance pour les agents qui jouent dans ces environnements, en prolongeant de façon la plus naturelle possible le cas classique. Nous montrons également comment on peut adapter les algorithmes reconnues (comme les ETC, epsilon-greedy, Boltzmann/Softmax, Thompson sampling, Bayes-UCB) dans certaines configurations avec récompense d'espérance non-définie. Nous nous intéresserons plus particulièrement aux distributions de Cauchy et de Pareto, qui ont une définition simple et constituent donc un contexte intéressant pour une première approche. Bien que notre contribution est plutôt théorique, notons aussi que ces distributions peuvent avoir des applications intéressantes, notamment on pourrait imaginer que la loi de Pareto peut modéliser des temps d'attente potentiellement catastrophique.

Keywords Sthocastic Bandit · Heavy Queue · Cauchy Bandit · Pareto Bandit

1 Introduction : Le problème des bandits stochastiques

Dans le problème des bandits stochastiques classiques (voir Lattimore and Csaba [2020]), un agent doit choisir à chaque pas de temps t une action $k_t \in \{1, 2, \dots, K\}$. L'agent reçoit alors une récompense $r_t \sim \nu(\mu_{k_t})$, où $\nu(\mu_{k_t})$ est une distribution d'espérance μ_{k_t}. Le but de l'agent est de maximiser la somme des récompenses sur un horizon de temps T, plus formellement on souhaite munir l'agent d'une politique d'action maximisant $\mathbb{E}\left[\sum_{t=1}^{T} r_{k_t}\right]$. Cette espérance sera maximisée si l'action optimale $k_\star = \underset{k}{\text{argmax}}\mu_k$ de moyenne optimale $\mu_\star = \max_k \mu_k$ est sélectionnée le plus souvent possible. L'agent tente donc de minimiser le pseudo-regret cumulatif défini par $\mathcal{R}(T) = T\mu_\star - \mathbb{E}\left[\sum_{t=1}^{T} r_t\right]$. Empiriquement, on peut vérifier la performance d'un agent en calculant son regret cumulatif empirique

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

où $\Delta_{k_t} = \mu_\star - \mu_{k_t}$ est le regret instantané cumulé par l'agent au temps t. La mesure de performance définie par (1) est la base de plusieurs expérimentations comparatives effectuées dans la littérature sur les bandits stochastiques. Le lecteur notera que l'architecture de ce problème suggère que les distributions ν définissant les récompenses ont une espérance. Or, plusieurs distributions de probabilité n'ont pas d'espérance. La loi de Cauchy et la loi de Pareto en sont des exemples et constitueront nos contextes d'études pour aborder

le problème d'extension des bandits stochastiques au cas avec récompense non-définie. Il est possible de s'imaginer ces deux dernières lois comme des analogues respectives à la loi normale et la loi exponentielle, mais avec des queues plus lourdes. Bien que notre contribution est plus de nature théorique, remarquons cependant que la loi de Pareto peut aider à modéliser des contextes d'applications, comme dans l'étude de temps d'attente modéliser par des distributions à queues lourdes. Aussi, nous avons observé que le problème tel que nous le définissons n'est pas couvert dans la littérature, certains couvrent des bandits à récompense à queues lourdes (Bubeck et al. [2013]), mais pas au point de ne pas avoir d'espérance.

2 Extension des bandits stochastiques pour la loi de Cauchy

On rappelle qu'une variable aléatoire X suit une loi de Cauchy ($X \sim \text{Cauchy}(L, a)$) si sa fonction de densité est $f(x; L; a) = \frac{1}{\pi\, a\left[1+\left(\frac{x-L}{a}\right)^2\right]}$ avec $L \in \mathbb{R}$ et $a > 0$. Cette loi de probabilité n'a pas d'espérance, mais est symétrique par rapport L (la localisation de la loi) et possède donc un centre naturel en L. Si on se place dans un problème on l'on doit choisir entre deux lois de Cauchy pour maximiser l'observation tirée de ces lois, il semble naturel de choisir la loi de localisation maximale pour augmenter les chances d'avoir une grande observation. En fait, il est facile de démontrer que si $X_1 \sim \text{Cauchy}(L_1, a_1)$ et $X_2 \sim \text{Cauchy}(L_2, a_2)$, alors

$$\mathbb{P}[X_1 > X_2] > 0.5 \quad \Leftrightarrow \quad L_1 > L_2 \tag{2}$$

$$\text{et} \qquad \mathbb{P}[X_1 > t] > \mathbb{P}[X_2 > t] \quad \forall t \in \mathbb{R} \quad \Leftrightarrow \quad L_1 > L_2 \text{ et } a_1 = a_2 \tag{3}$$

Autrement dit, on peut toujours établir une relation de dominance entre deux lois de Cauchy de paramètres de localisations distincts en utilisant la définition (2). Dans le cas où les paramètres d'échelles sont égaux, on peut établir une relation de dominance plus forte en utilisant (3). Conséquemment, si on définit un bandit stochastique à K bras de distributions Cauchy, il semble naturel de définir l'action optimale comme $k_\star = \underset{k}{\text{argmax}}\, L_k$ à partir de la localisation optimale $L_\star = \max_k L_k$. Le regret associé au choix de l'action k devient donc $\Delta_k = L_\star - L_k$. En appliquant cela dans (1), on obtient une mesure de performance empirique permettant d'évaluer la performance d'un agent dans cette configuration.

Les algorithmes classiques basés sur la moyenne empirique des récompenses pour choisir l'action optimale échoueront à avoir de bons résultats dans cette configuration étant donné que la moyenne empirique n'estime pas bien le paramètre de localisation de la loi. En fait cet estimateur est divergent dans ce cas. Conséquemment, nous devons d'abord trouver de bons estimateurs de la localisation d'une loi de Cauchy.

2.1 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 Cauchy(L, a) où L est inconnu et a connu. Soit $X_{(i)}$ la $i^{\text{ième}}$ statistique d'ordre de $\mathcal{X}$. On définit les estimateurs empiriques suivants pour L.

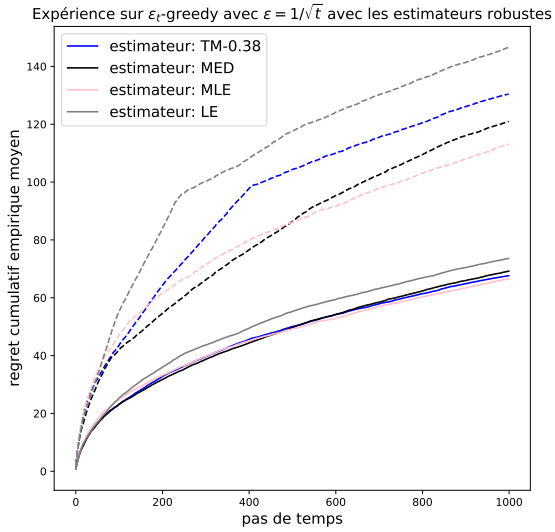
- La médiane : $\text{MED}(\mathcal{X})$
- La moyenne $\alpha-$tronquée : $TM_\alpha(\mathcal{X}) = \dfrac{1}{T-2r}\displaystyle\sum_{i=r+1}^{T-r} X_{(i)}$, où $r = \lfloor T\alpha \rfloor$ et $0 < \alpha < 0.5$
- L'estimateur de maximum de vraisemblance : $\text{MLE}(\mathcal{X})$
- L-estimator : $\text{LE}(\mathcal{X}) = \dfrac{1}{T}\displaystyle\sum_{i=1}^{T} J\left(\frac{i}{T+1}\right) X_{(i)}$, où $J(u) = \frac{\sin(4\pi(u-0.5)}{\tan(\pi(u-0.5)}$,

Remarquons que MLE ne possède pas de formule analytique simple, mais peut être estimé numériquement à l'aide d'un algorithme d'optimisation tel que Newton-Raphson (voir Haas et al. [1970]), ce que nous avons fait. Ce dernier demande donc des ressources computationnelles plus grandes et de faire un choix de critère d'arrêt sur la congergence lors de l'optimsation de la fonction de vraisemblance. L'estimateur LE, tiré de Zhang [2010], a quant à lui l'avantage d'avoir une formule analytique sous forme d'une somme. On a aussi remarqueé que bien que non-noté dans l'article, LE n'était pas défini lorsque T est pair en raison du fait que u pourra alors prendre la valeur 0.5, causant une indétermination de la forme $0/0$ de la fonction $J(u)$.

Or c'est une singularité effaçable et donc $J(u)$ peut être prolongé analytiquement en $u = 0.5$ en définissant $J(0.5) := \lim_{u \to 0.5} J(u) = 4$. L'efficacité de ces différents estimateurs est étudiée dans Zhang [2010]. Notre but n'étant pas de détailler les différences d'efficacité de ces estimateurs ici, notons tout de même que l'efficacité de ces estimateurs diffèrent selon la taille T du jeu de données, certains étant moins stables lorsque T petit. Conséquemment, il serait possible d'explorer la possibilité d'établir un estimateur basé sur une mixture d'experts. Notons que tous ces estimateurs sauf le MLE exigent d'ordonner les données, ainsi leur valeur sera obtenue en $O(T \log(T))$.

2.2 Adaptation d'algorithmes sur des bandits Cauchy

Les algorithmes ETC (Garivier et al. [2016]), ϵ-greedy et Boltzmann/Softmax (Cesa-Bianchi et al. [2017]) peuvent être adaptés pour jouer sur des bandits Cauchy en remplaçant l'estimation empirique $\hat{\mu}_k(t-1)$ (moyenne des récompenses obtenues aux pas de temps où le bras numéro k a été joué dans les $(t-1)$ premiers pas de temps) permettant d'estimer le bras optimal par un estimateur $\widehat{L}_k(t-1)$ de la localisation L_k de la distribution du bras no k. Chaque estimateurs présentés à la section 2.1 nous donne une adaptation différente de ces algorithmes. Pour illustrer ce fait, nous définissons ϵ_t-greedy avec $\epsilon_t = \frac{1}{\sqrt{t}}$ muni de ces estimateurs. Pour $N = 200$ répétitions, chacun des agents ainsi défini a joué sur un bandit Cauchy à deux bras avec $a_1 = a_2 = 1$ et $L_1, L_2 \sim \mathcal{U}([0,5])$ (loi uniforme sur l'intervalle $[0,5]$) sur un horizon de $T = 1000$ pas de temps.



Le graphique montre le regret empirique cumulatif moyenné sur les répétitions pour chaque version adaptée de l'algorithme. Les courbes en pointillés représentent un écart-type au dessus. Les algorithmes réussissent à réduire la pente du regret cumulatif avec le temps, mission sur laquelle l'algorithme classique basée sur la moyenne empirique échouerait grandement dans cette situation. L'idée ici est de présenter le fait que les versions adaptées de l'algorithme ont un comportement souhaitable au niveau de la performance sans faire une étude exhaustive des performances sous différents contextes. Notons que le problème que nous présentons ici ouvre la porte à comparer la performance des algorithmes sur une nouvelle dimension, celle de l'estimateur de localisation utilisée. Ce qui pourrait être fait dans une analyse plus spécifique. Après avoir choisi un estimateur ou une mixture optimale, il serait possible de comparer plusieurs algorithmes différents tous munis de cet estimateur.

FIGURE 1 – Expérience illustrant la performance de epsilon greedy muni des estimateurs de localisations de la loi de Cauchy

3 Extension des bandits stochastiques pour les lois de Pareto à espérance infinie

Dans cette section, nous considérons un problème où l'on souhaite optimiser des récompenses provenant de lois de Pareto, il serait possible par symétrie transformer ce problème en un problème de minimisation de temps d'attente. Nous ne poursuivrons toutefois pas ce but ici.

Une variable aléatoire X suit une loi de Pareto ($X \sim \text{Pareto}(L, a)$) si sa fonction de densité est

$$f(x; L, a) = \begin{cases} \dfrac{aL^a}{x^{a+1}} & \text{si } x \geq L \\ 0 & \text{sinon} \end{cases} \qquad \text{où } L > 0 \text{ et } a > 0. \tag{4}$$

Les figures 2 et 3 montre l'effet des deux paramètres sur la densité de Pareto.

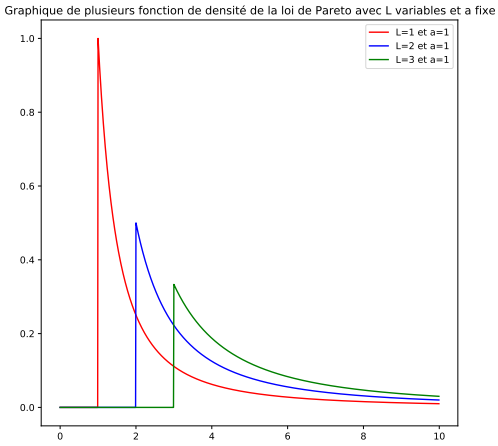


FIGURE 2 – Graphique illustrant l'effet du paramètre L sur la densité de Pareto

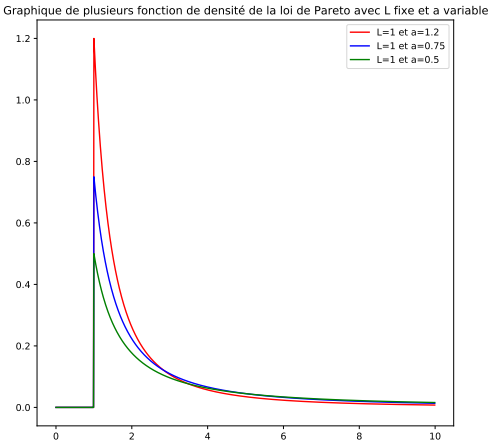


FIGURE 3 – Graphique illustrant l'effet du paramètre a sur la densité de Pareto

La fonction de répartition est $F_X(x) = \mathbb{P}[X \leq x] = \begin{cases} 1-(L/x)^a & \text{si } x \geq L \\ 0 & \text{sinon} \end{cases}$,

et son espérance est $\mu = \begin{cases} \frac{aL}{a-1} & \text{pour } a > 1 \\ \infty & \text{sinon} \end{cases}$

Le cas où $0 < a \leq 1$ est celui qui nous intéresse le plus puisque l'espérance n'existe pas. Dans ce dernier cas, nous souhaiterons définir le regret de façon la plus cohérente possible avec le cas où l'espérance existe. Pour la suite, il importe donc de noter que dans ce dernier cas ($a > 1$), si on pose deux variables $X_1 \sim \text{Pareto}(L_1, a_1)$ et $X_2 \sim \text{Pareto}(L_2, a_2)$, alors la condition $L_1 = L_2 = L$ (localisation fixe) entraîne que

$$\mu_1 > \mu_2 \Leftrightarrow \frac{a_1 L}{a_1 - 1} > \frac{a_2 L}{a_2 - 1} \Leftrightarrow a_1 < a_2 \tag{5}$$

et que la condition $a_1 = a_2 = a$ (paramètre a fixe) entraîne que

$$\mu_1 > \mu_2 \Leftrightarrow \frac{aL_1}{a - 1} > \frac{aL_2}{a - 1} \Leftrightarrow L_1 > L_2 \tag{6}$$

Ainsi, si deux distributions de Pareto sont comparées sur la base de leur espérance, cette dernière grandit lorsque L grandit et grandit lorsque a rapetisse.

Dans le cadre général où l'on peut avoir $0 < a \leq 1$ (espérance infinie), on tentera de comparer deux lois de Pareto à l'aide d'une relation de dominance de première ordre. Soit $\mathcal{D}_1$ et $\mathcal{D}_2$ deux distributions de probabilité, on dira que $\mathcal{D}_1$ domine $\mathcal{D}_2$ et on écrira $\mathcal{D}_1 \succeq \mathcal{D}_2$ si $\mathbb{P}[X_1 > t] \geq \mathbb{P}[X_2 > t] \ \forall t \in \mathbb{R}$, où $X_1 \sim \mathcal{D}_1$ et $X_2 \sim \mathcal{D}_2$. Notons que cela est équivalent à $F_{X_1}(t) \leq F_{X_2}(t) \ \forall t \in \mathbb{R}$. Cette façon de comparer deux distributions correspond en fait à la dominance stochastique de premier ordre (définie dans Schmid and Trede [1996] par exemple). Notons que nous avons déjà utilisé cette relation d'ordre pour les distributions de Cauchy à la relation (3).

3.1 Relation de dominance entre distributions de Pareto

On note que si $\mathcal{D}_1 = \text{Pareto}(L_1, a)$ et $\mathcal{D}_2 = \text{Pareto}(L_2, a)$, alors $\mathcal{D}_1 \succeq \mathcal{D}_2 \Leftrightarrow L_1 \geq L_2$.
Dans le cas où $\mathcal{D}_1 = \text{Pareto}(L, a_1)$ et $\mathcal{D}_2 = \text{Pareto}(L, a_2)$, alors $\mathcal{D}_1 \succeq \mathcal{D}_2 \Leftrightarrow a_1 \leq a_2$.
Les démonstrations sont simples en utilisant l'expression de la fonctions de répartition de la loi de Pareto. Notons que si $\mathcal{D}_1 = \text{Pareto}(L_1, a_1)$ et $\mathcal{D}_2 = \text{Pareto}(L_2, a_2)$ avec $L_1 \neq L_2$ et $a_1 \neq a_2$, il n'est pas toujours

possible d'ordonner $\mathcal{D}_1$ et $\mathcal{D}_2$ selon cette définition de dominance puisque les fonctions de répartition associées à ces distributions pourront alors s'intersecter pour un certain $t \in]\max(L_1, L_2); \infty[$.

On note que les relations de dominance ainsi définies sont donc compatibles avec la «dominance» établie sur la comparaison des espérances, relations (5) et (6), dans le cas d'espérance définie.

Avec ces observations, nous pouvons définir intuitivement et naturellement des notions de regrets pour certaines configurations de bandits de Pareto.

3.2 Bandits de Pareto avec paramètre a constant

On suppose ici un bandit stochastique à K bras dont le numéro k est de distribution Pareto(L_k, a) Ici, $a > 0$ est fixe mais quelconque, le cas où $a \leq 1$ est particulièrement intéressant dans le contexte de notre étude. On définit dans cette situation le bras optimal et la notion de regret en se basant sur les localisations L_k des bras.

$$L_\star = \max_k L_k, \qquad k_\star = \operatorname*{argmax}_k L_k \quad \text{et} \quad \Delta_k = L_\star - L_k$$

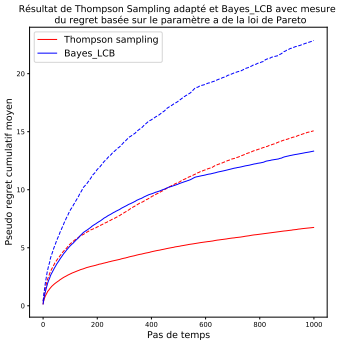
Conséquemment, on peut étendre les algorithmes classiques ETC, $\epsilon-$greedy et Boltzmann/Softmax comme on l'a fait plus tôt (voir section 2.2) à condition de les munir de bons estimateurs de L pour une distribution de Pareto. Un choix naturel est $\widehat{L} := min(\mathcal{X})$ lequel correspond à l'estimateur du maximum de vraisemblance.

3.3 Bandits de Pareto avec paramètre L constant

On suppose ici un bandit stochastiques à K bras dont le numéro k est de distribution Pareto(L, a_k) où L est fixe. On remarque que les a_k étant variables, certains peuvent être inférieure à 1 et d'autres non. Cette configuration est intéressante car elle peut contenir des bras à distribution d'espérance non-définie et d'autres à distribution d'espérance définie. On définit dans cette situation le bras optimal et la notion de regret en se basant sur les paramètres a_k des bras, n'oubliant pas que plus a_k est petit, plus la distribution est dominante.

$$a_\star = \min_k a_k, \qquad k_\star = \operatorname*{argmin}_k a_k \quad \text{et} \quad \Delta_k = a_k - a_\star \tag{7}$$

Puisque le conjugué de la loi de Pareto pour l'estimation du paramètre a existe (il s'agit d'une loi gamma), nous pouvons donc songer définir les algorithmes de perspectives bayésiennes sur ces configurations de bandits (voir Agrawal and Goyal [2013] et Kaufmann et al. [2012]). On peut facilement programmer Thompson Sampling en utilisant le fait que la posterior d'une prior Gamma(α, β) après n tirages est une Gamma $(\alpha + n, \beta + \sum_{i=1}^{n} \ln(x_i/L))$ dans cette situation. Par contre on devra prendre le minimum des θ_k échantillonnés à chaque tour et non le maximum. Aussi, on peut programmer un Bayes_LCB (Bayes lower confidence bound) en s'inspirant de Bayes_UCB mais en prenant le quantile de probabilité cumulé de $\frac{1}{t(\ln(T))^c}$ au lieu de $1 - \frac{1}{t(\ln(T))^c}$ où T est l'horizon joué et c un paramètre. Par exemple, le graphique suivant montre les résultats obtenus en jouant ces deux algorithmes sur $N = 200$ répétions sur un horizon de $T = 1000$ sur des bandits Pareto avec $L_1 = L_2 = 1$ et $a_1, a_2 \sim \mathcal{U}([0, 1])$ pour chaque répétition. On a pris $c = 5$ pour Bayes_LCB. Aussi, On a pris les priors $\alpha = 1$, $\beta = 1$ pour que la loi Gamma initiale couvre bien l'espace $[0, 1]$ des valeurs possibles de a dans cette expérience



On voit que ces algorithmes peuvent performer de façon correcte sur des configurations à queues lourdes au point de n'avoir pas d'espérance. Notons que nous n'avons pas tenter d'optimiser la performance de Bayes_LCB en travaillant sur le paramètre c, le but n'étant pas nécessairement de comparer les deux algorithmes mais de montrer comment les définir dans le contexte actuel. Le lecteur intéressé pourrait comparer ces algorithmes avec des algorithmes classiques plus simples basées sur des estimateurs du paramètre a, lesquels peuvent s'obtenir facilement par maximum de vraisemblance pour la loi de Pareto.

FIGURE 4 – Thompson Sampling adapté et Bayes LCB

3.3.1 Rendre cohérente la mesure du regret basé sur le paramètre a

Puisque la fonction $\mu(L, a) = aL/(a-1)$ donnant l'espérance en fonction de a et L dans le cas d'espérance finie n'est pas de décroissance linéaire en a, il serait souhaitable d'ajuster la valeur du regret donnée par (7) pour que son expression soit plus en phase avec la fonction $\mu(L, a)$ pour tenter que les deux mondes se connectent mieux.

4 Conclusion

Nous avons montré qu'il était possible d'étendre la notion de regrets sur des bandits stochastiques de récompense d'espérance non-définie, en se basant sur certaines propriétés des distributions dans les cas spécifiques étudiées ici. Cela entraîne que nos définitions de regrets dépendent beaucoup des distributions définissant l'environnement du bantit. Il serait intéressant de tenter définir les éléments dans un cadre plus général. On peut aussi voir un certain parallèle avec le problème du reward shaping qu'on observe dans les algorithmes de bandits séquentiel, car si on désire minimiser un temps d'attente, mais qu'on désire ne pas avoir de temps d'attente catastrophiques, les algorithmes doivent aussi prendre en compte les queues des distributions et non seulement un de leur paramètre de tendance centrale (moyenne, médiane,...).

Références

Tor Lattimore and Szepesvari Csaba. *Bandit algorithms*. Cambridge University Press, 2020.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11) :7711–7717, 2013. doi:10.1109/TIT.2013.2277869.

Gerald Haas, Lee Bain, and Charles Antle. Inferences for the cauchy distribution based on maximum likelihood estimators. *Biometrika*, 57(2) :403–408, 1970.

Jin Zhang. A highly efficient l-estimator for the location parameter of the cauchy distribution. *Computational Statistics*, 25(1) :97–105, 2010.

Aurélien Garivier, Emilie Kaufmann, and Tor Lattimore. On explore-then-commit strategies, 2016.

Nicolò Cesa-Bianchi, Claudio Gentile, Gabor Lugosi, and Gergely Neu. Boltzmann exploration done right. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6284–6293. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/b299ad862b6f12cb57679f0538eca514-Paper.pdf`.

Friedrich Schmid and Mark Trede. Testing for first-order stochastic dominance : A new distribution-free test. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 45(3) :371–380, 1996. ISSN 00390526, 14679884. URL `http://www.jstor.org/stable/2988473`.

Shipra Agrawal and Navin Goyal. Further optimal regret bounds for thompson sampling. In Carlos M. Carvalho and Pradeep Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 99–107, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL `http://proceedings.mlr.press/v31/agrawal13a.html`.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On bayesian upper confidence bounds for bandit problems. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 592–600, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL `http://proceedings.mlr.press/v22/kaufmann12.html`.