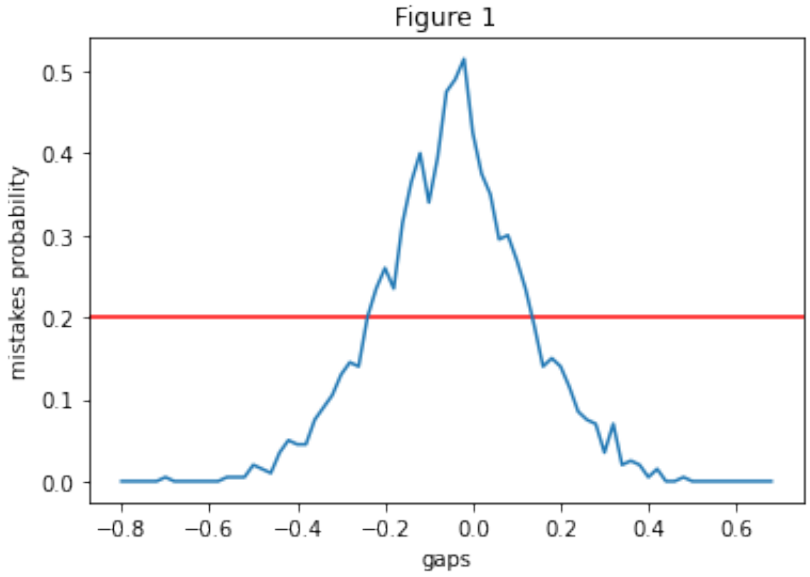
1 Algorithme de base

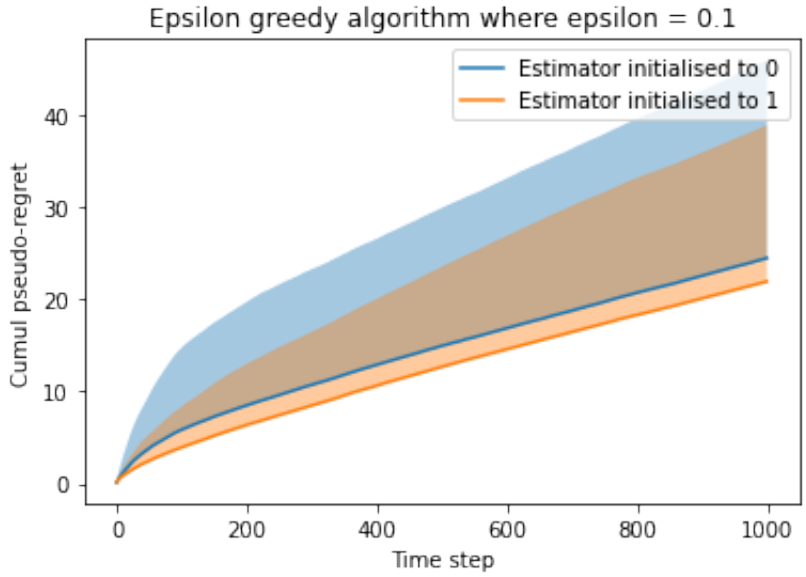
1.1 Experience sur l'algorithme ETC

L'algorithme ETC (explore then commit) n'est pas uniformément efficace, c'est-à-dire qu'il souffrira d'un regret linéaire pour les instances plus difficiles. Il va avoir tendance à converger vers des actions sous-optimal due à leurs exploration limitée. Pour appuyer cela, nous avons réalisé plusieurs experiences avec à chaque fois une instance de two armed bandits. À chaque pas de temps, notre gap variera et nous enregistrerons les performances, c'est-à-dire la probabilité de choisir la mauvaise action.
Nous pouvons voir grâce à la figure 1, que plus le gap va être petit, plus la probabilité d'erreur va être élevée. On peut aussi en déduire que tout gap supérieur à 0.2 positifs ou négatif échouera avec une probabilité d'au moins 20.



1.2 Optimisation de l'algorithme epsilon-greedy

Pour rendre l'algorithme epsilon-greedy optimiste, nous avons procédé à une expérience comparative sur des bandits à distribution normale.
Dans un premier temps, nous avons généré N instances de two armed bandits et pour chaque instance nous avons minimisé le regret.
Pour l'expérience 1, nous avons initialisé les estimateurs empiriques à 0 et pour l'expérience 2, nous les avons initialisés à 1. Nous avons ensuite affiché les pseudos regret cumulatif de ses 50 instances dans un graphique. Pour que l'effet soit bien perceptible et non biaisé par l'aléatoire, il est important de prendre un grand nombre d'instances de bandits.
Nous pouvons constater à première vu que le pseudo regret cumulatif moyen de l'expérience 2 est lègèrement plus petit. En initialisant les estimateurs empiriques à 1, nous augmentons les chances que le premier tirage ne biaise pas le résultat final.



2 KL UCB

2.1 Pseudo-code

On consid ?re un bandit ? K bras sur un horizon de temps de T

- Pour chaque $t \leq K$:
  - $N[t] = N[t] + 1$
  - $S[t] = S[t] + Gain(t)$
- Pour chaque $K < t \leq T$:
  - $k_t = amgmax(2)$
  - $N[k_t] = N[k_t] + 1$
  - $S[k_t] = S[k_t] + Gain(k_t)$

$$d(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \tag{1}$$

$$max\left\{q \in \Theta : N[a]d\left(\frac{S[a]}{N[a]}, q\right) \leq \log(t) + c\log(\log(t))\right\} \tag{2}$$

2.2 Implementation bernoulli

Pour implementer cet algorithme, on doit utiliser un optimiseur numerique, car il n'existe pas de forme fermée lorsque les recompenses suivent une distrubution bernoulli. De plus, l'implementation inclu un bris d'egalite. Ainsi si 2 bras ont le meme UCB et qu'ils representent les bras avec le plus grand UCB alors l'algorithme choisira le bras ayant le plus petit nombre de jeu.

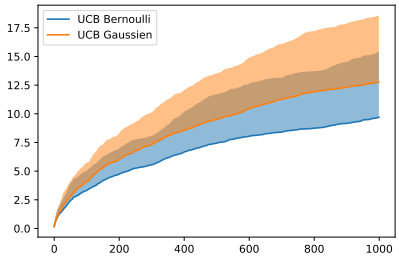
2.3 Implementation UCB Gaussienne

Pour l'implementation de l'algorithme UCB gaussien il existe une forme fermee. Cette forme a ete utilisee dans l'implementation, car elle rend l'exploitation plus rapide, car il n'a pas d'optimisation a faire pour determiner l'UCB. Pour calculer la borne UCB gaussien, l'ecart-type doit etre connu, mais ce n'est pas le cas pour la configuration de gandit utilise. L'ecart-type maximale d'une distribution bernoulli est donc utilisee (0.25).

2.4 Comparaison des modèles

On peut observer que les deux modèles performent bien sur des bandits bernoulli. Les deux modèles ont un regret cumulatif sous linéaire, mais le modèle utilsant la borne KL UCB pour bandit bernoulli performe mieux. Cela peut être dut au fait qu'avec le modèle qui utilise la borne KL UCB pour des bandits gaussien utilise une borne avec une variance de 0.25. Cette variance est utilisée, car elle représente la variance maximale dans un bandit bernoulli ainsi,

elle permet de conserver des garanties. Par contre l'utilisation de cette borne cause une sur exploration. On peut observer l'effet de la sur exploration dans 2.4. On peut aussi observer que la variance semble un peu plus élevée pour le KL UCB optimisé. Cela peut être dut au fait que relativement peu d'instances ont été générées.



3 No 3

3.1 Rappel sur la loi de gamma

X est une variable aléatoire suivant une loi gamma ($X \sim \text{Gamma}(\alpha,\beta)$) si sa fonction de densité est

$$f(x;\alpha,\beta) = \begin{cases} \dfrac{\beta^\alpha x^{\alpha-1} \mathrm{e}^{-\beta x}}{\Gamma(\alpha)} & \text{si } x>0 \\ 0 & \text{sinon} \end{cases} \tag{3}$$

où $\alpha>0$ et $\beta>0$ sont des paramètres.
Avec cette paramétrisation, l'espérance et la variance de cette loi sont

$$\mu = \frac{\alpha}{\beta} \tag{4}$$

$$\sigma^2 = \frac{\alpha}{\beta^2} \tag{5}$$

La loi gamma est asymétrique à droite (asymétrie de $2/\sqrt{\alpha}$).
On observe que le système d'équations (4) et (5) est équivalent à

$$\alpha = \frac{\mu^2}{\sigma^2} \tag{6}$$

$$\beta = \frac{\mu}{\sigma^2} \tag{7}$$

Cela est utile notamment si l'on souhaite obtenir les paramètres d'une loi gamma étant donnée sa moyenne et sa variance.

3.2 Rappel sur la loi de Poisson

La loi de Poisson est une loi de probabilité discrète dont le support est $\mathbb{N}=\{0,1,2,3,...\}$.
On dit que X suit une loi de Poisson ($X \sim \text{Pois}(\lambda)$) si sa fonction de probabilité (fonction de masse) est

$$f(x;\lambda) = \begin{cases} \dfrac{\mathrm{e}^{-\lambda}\lambda^x}{x!} & \text{si } x \in \mathbb{N} \\ 0 & \text{sinon} \end{cases}$$

où λ est un paramètre. L'espérance est $\mu=\lambda$ et la variance $\sigma^2=\lambda$.

3.3 Algorithme Thomas Sampling pour des bandits à distribution Poisson

Pour des bandits à distributions de Poisson, c'est-à-dire dans le cas où chaque bras génère un reward selon une loi de Poisson de paramètre λ inconnu, le conjugué à utilisé est la loi gamma.

Plus précisément, après n observations $x_1,x_2,...,x_n$ tirées d'une loi de Poisson inconnue, la distribution posterior modélisant le paramètre λ inconnu sera

$$\text{Gamma}\left(\alpha_0+\sum_{i=1}^{n} x_i, \beta_0+n\right) \tag{8}$$

où α_0 et β_0 sont les paramètres de la distribution de gamma prior, ils correspondent donc aux paramètres choisis permettant d'initier la distribution.

Cela mène au pseudo-code suivant pour un algorithme Thomas Sampling sur des bandits à K bras à distribution de Poisson.

3.3.1 Pseudo-code

- Initialiser une distribution gamma $\pi_k(0)$ pour $k=1,2,...,K$, c'est-à-dire choisir un couple $(\alpha_{k,0},\beta_{k,0})\forall k$
- Pour chaque $t \geq 1$:
  - Échantillonner un nombre $\theta_{k,t}$ de la distribution $\pi_k(t-1)$, pour $k=1,2,...,K$
  - Sélectionner $k_t = \underset{k}{\text{argmax}}(\theta_{k,t})$
  - jouer l'action k_t et observer le reward r_t
  - Modifier la posterior $\pi_{k_t}(t)$ de l'action jouée : $\alpha_{k_t} \leftarrow \alpha_{k_t} + r_t$ et $\beta_{k_t} \leftarrow \beta_{k_t} + 1$
  - Laisser inchangé les autres posteriors : $\pi_k(t) \leftarrow \pi_k(t-1), \forall k \neq k_t$

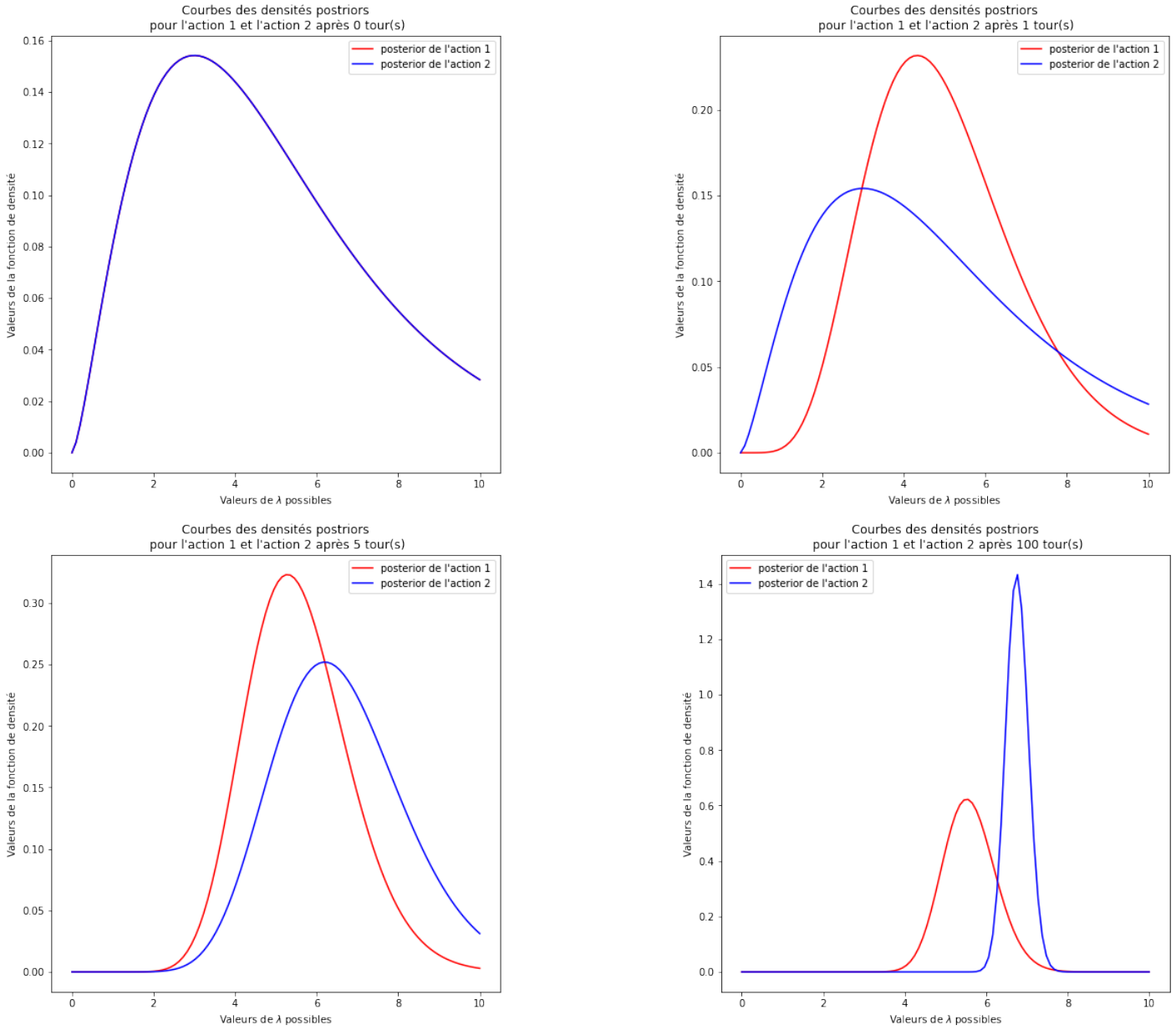
Remarque : même si ce n'est pas toujours mentionné explicitement dans la suite, les différents bras d'un bandit auront toujours la même paramétrisation de leur distribution prior.

3.4 Exploration du comportement de l'alogorihme (validation de la fonctionnalité)

Pour explorer le comportement de l'algorithme et vérifier qu'il fonctionne bien, on propose une expérience. Cette dernière consistera à lancer l'algorithme sur un horizon $T = 100$ sur une instance de bandit à deux bras à distribution de Poisson de moyennes respectives $\lambda_1 = 6$, $\lambda_2 = 7$ et avec une paramétrisation prior $\alpha_0 = 5/2$, $\beta_0 = 1/2$ identiques pour les deux bras. Pour tout $t = 0,1,...,100$, on affiche les 2 distributions posteriors. On pourra ainsi voir l'évolution des posteriors selon les pas de temps. Notons qu'ici, on choisit une loi gamma initiale avec $\mu_0 = 5$ et $\sigma_0^2 = 10$. Nous discuterons plus loin du choix de la paramétrisation initiale. La figure 1 montre le graphiques des densités posteriors après 0 tour (priors), 1 tour, 5 tours, 100 tours. Dans le notebook, l'ensemble des 101 figures sont générées.

Initialement, on voit que les deux priors sont superposées. Après un tour, on voit qu'une seule distribution est modifiée (celle de l'action jouée, ici l'action 1). On voit par la suite que les distributions posteriors se déplacent et se concentrent, notamment celle de l'action optimale 2 qui se concentre de plus en plus autour de la moyenne associée à son bras ($\lambda_2 = 7$) et celle de l'action minimale se concentrant plus à gauche. Le comportement des posteriors est attendu et rassurant. À long terme, à chaque tour, il est de plus en plus probable que l'action jouée soit l'action optimale.

FIGURE 1 – Évolution des posteriors



3.5 Expérimentation sur l'algorithme

Dans cette section, on explore le comportement de l'algorithme dans diverses situations tout en s'efforçant de discuter de la sélection des priors. En effet, la loi de gamma n'ayant aucune paramétrisation la faisant correspondre à une loi uniforme, il faut faire un choix de priors qui n'est pas trivial. Pour ce faire, on supposera ici que l'on ne connaît pas les valeurs de λ associée à chaque bras, mais que l'on connaît à tout de moins un intervalle sur lequel ils se trouvent. Dépendant des choix des priors, ces derniers sont des distributions qui peuvent couvrir bien ou non l'intervalle possibles des λ, on suspecte que cela aura un effet sur le pseudo-regret cumulatif.

Pour établir une prior, il faut établir un couple α_0,β_0. Nous établirons plutôt un moyenne (μ_0) et une variance (σ_0^2) pour cette distribution prior et obtiendrons α_0 et β_0 associées avec les équations (4) et (5).

3.5.1 Effet de la variance de la loi prior

Expérience sur des variances trop petites

On génère $N=1000$ instances de bandits de Poisson à deux bras.
Pour chaque instance, les moyennes λ_1,λ_2 des bras sont tirées aléatoirement d'une loi uniforme continue sur l'intervalle [0,10] pour chaque instance.
On considérera le cas ici ou l'agent sait que les moyennes sont dans cet intervalle.

Pour chaque couple (μ_0, σ_0^2,) suivant :

$$(5,10),(5,5),(5,1),(5,1/2),(5,1/4)$$

On trouve α_0 et β_0 associée à l'aide des équations (4) et (5)

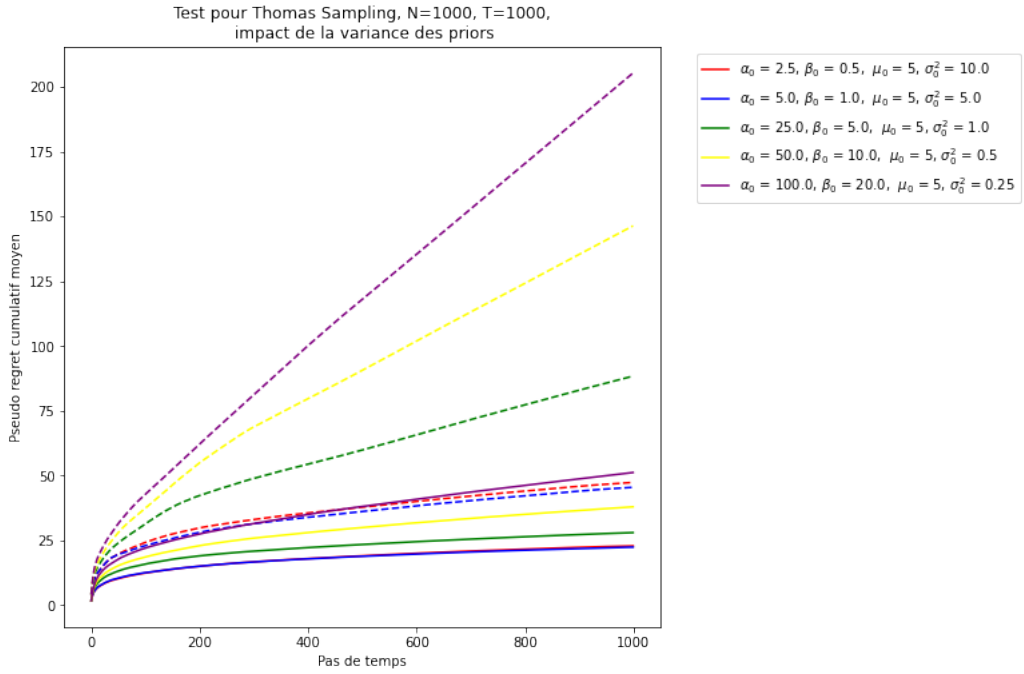
On fait jouer Thomas sampling sur les $N=1000$ instances sur un horizon $T=1000$.

On trace les pseudos-regret cumulatif moyen avec une déviation standard au dessus.

Notons qu'ici, dans les différents couples (μ_0,σ_0^2) explorés, la moyenne est toujours 5 (le centre de l'intervalle [0,10]) et la variance varie (partant de 10 correspondant à la grandeur de l'intervalle et se rapetissant).

La figure suivante montre le résultat obtenu.

FIGURE 2 – Comparaisons de la performance selon des priors de variances plus petites



On remarque que si la distribution prior est choisie avec une variance trop petite, la distribution prior ne couvre pas suffisamment bien l'intervalle [0,10], et cela semble se traduire par une augmentation du pseudo-regret cumulatif moyen. On remarque qu'il semble y avoir un certaine tolérance sur des variations de σ_0^2 qui ne changent pas significativement la courbe du pseudo-regret cumulatif moyen.

Notez que partout dans les graphiques ci-dessous, la ligne continue représente le pseudo-regret cumulatif moyen alors que la ligne pointillée représente la déviation-standard au dessus associée à ce pseudo-regret cumulatif.

Expérience sur des variances trop grandes

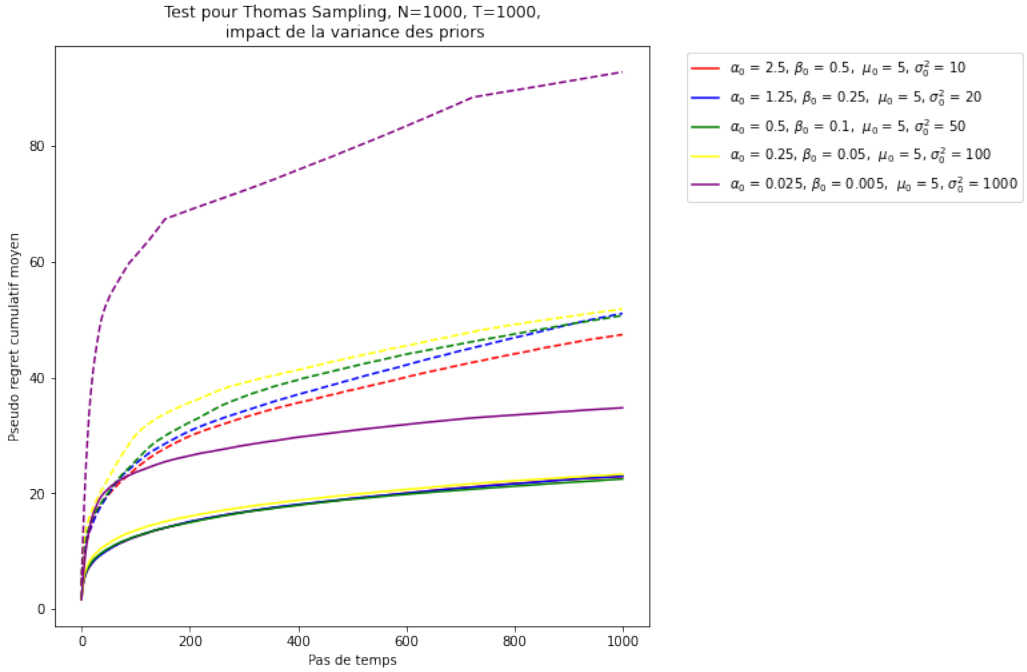
Même méthode que l'expérience 1, mais sur les couples (μ_0,σ_0^2) suivants.

$$(5,10),(5,20),(5,50),(5,100),(5,1000)$$

La moyenne des priors est toujours le centre de l'intervalle [0,10], la variance varie (partant de 10 correspondant à la grandeur de l'intervalle et se grandissant).

La figure suivante montre le résultat obtenu.

FIGURE 3 – Comparaisons de la performance selon des priors de variances plus grandes



On peut voir qu'une variance légèrement trog grande n'aura pas beaucoup d'effet sur le regret cumulatif moyen, par contre, une variance excessivement trop grande peut en avoir une considérable. Dans ce cas, la distribution prior sur-couvre grandement l'intervalle [0,10] des valeurs possibles de λ. Conséquemment, possiblement que les distributions posteriors ne se concentrent pas suffisamment rapidement pour pouvoir bien cerner les moyennes λ_1 et λ_2 des distributions de Poisson des bras.

3.5.2 Effet de la moyenne de la loi prior avec variance «bonne»

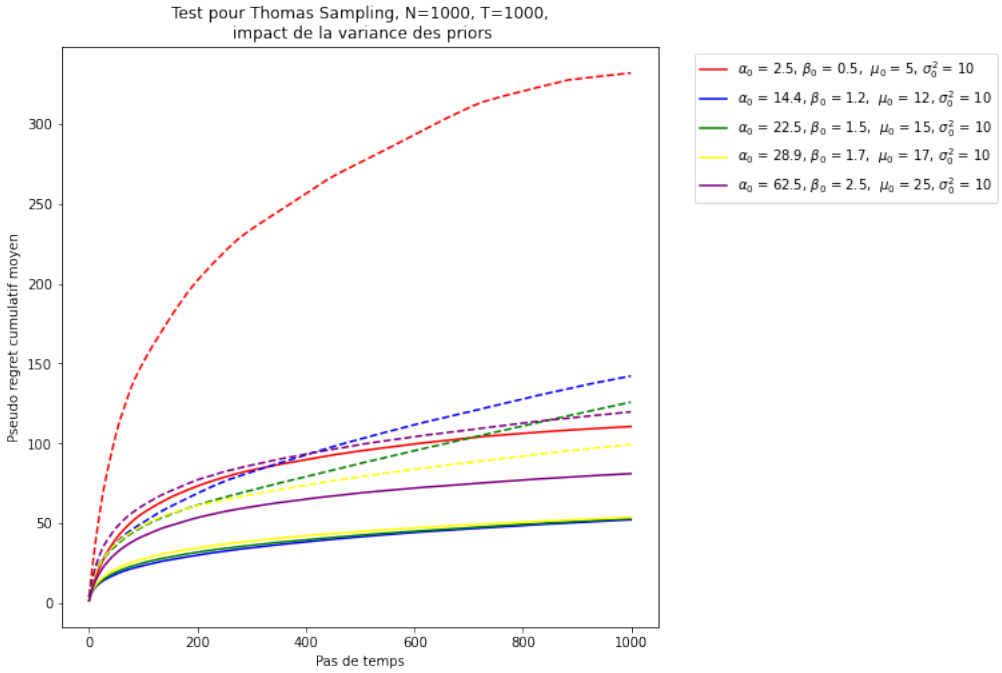
Pour bien pouvoir faire varier les moyennes de la prior choisi, on va ici considérer une situation analogue aux deux premières expériences, mais où l'intervalle sur lequel sont choisis les moyennes λ_1,λ_2 est [10,20] au lieu de [0,10]. On effectue la même expérience que si haut, sur les couples (μ_0,σ_0^2) qui suivent.

$$(5,10),(12,10),(15,10),(17,10),(25,10)$$

La variance de la prior est fixe égale à la longueur de l'intervalle des valeurs possibles de λ.

On obtient la figure suivante

FIGURE 4 – Comparaisons de la performance selon des priors de variances égale à la longueur de l'intervalle et de moyennes variables.



On remarque que si la moyenne du prior est fortement décalée du centre de l'intervalle [10,20], cela aura un effet sur la performance de l'algorithme, notamment le cas où $\mu_0 = 5$. Lorsque la moyenne μ_0 reste dans l'intervalle [10,20], il semble que le fait que la variance soit bonne (distribution prior qui couvre bien l'intervalle) fait en sorte que l'algorithme reste bon même si μ_0 n'est pas centré.

3.5.3 Effet de la moyenne de la loi prior avec variance plus petite

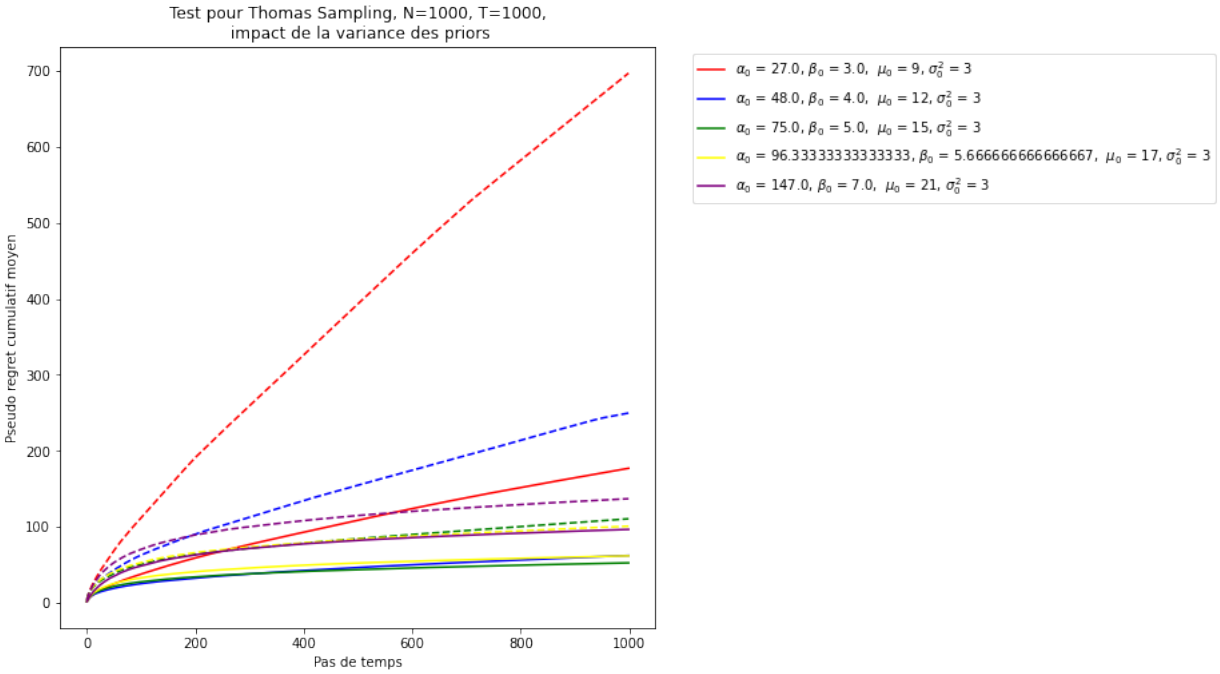
Même expérience que précédemment, mais avec les couples (μ_0,σ_0^2) qui suivent.

$$(9,3),(12,3),(15,3),(17,3),(21,3)$$

Cette fois, la variance est plus petite (3).

On obtient la figure suivante

FIGURE 5 – Comparaisons de la performance selon des priors de variances égalent à 3 et moyennes variables.



On voit qu'ici, avec une variance des priors plus petite, un même décalage de la moyenne μ_0 par rapport au centre de l'intervalle semble être plus coûteux que dans l'expérience précédente. D'ailleurs, le cas $\mu_0 = 9$ semble en être un bon exemple car il atteint une regret cumulatif moyen final plus élevé que le cas $\mu_0 = 5$ dans le cas où la variance était de 10. Conséquemment, si l'information que l'on a sur les λ (intervalle de possibilité) n'est pas bonne pour fixer les priors, lancer un tel algorithme avec une variance trop petite peut avoir un impact sur la performance, notamment si par malchance on choisit un prior avec μ_0 décalé du centre du réel intervalle des valeurs possibles de λ.

En conclusion, à la lumière des simulations effectuées, les priors semblent devoir être choisis pour couvrir raisonnablement bien l'intervalle des valeurs possibles de λ. On pourrait suggérer de prendre la distribution prior telle que μ_0 est le centre de l'intervalle et σ_0^2 la longueur de l'intervalle. Selon nos simulations, cela semble être un choix raisonnable.