TP2 bandits

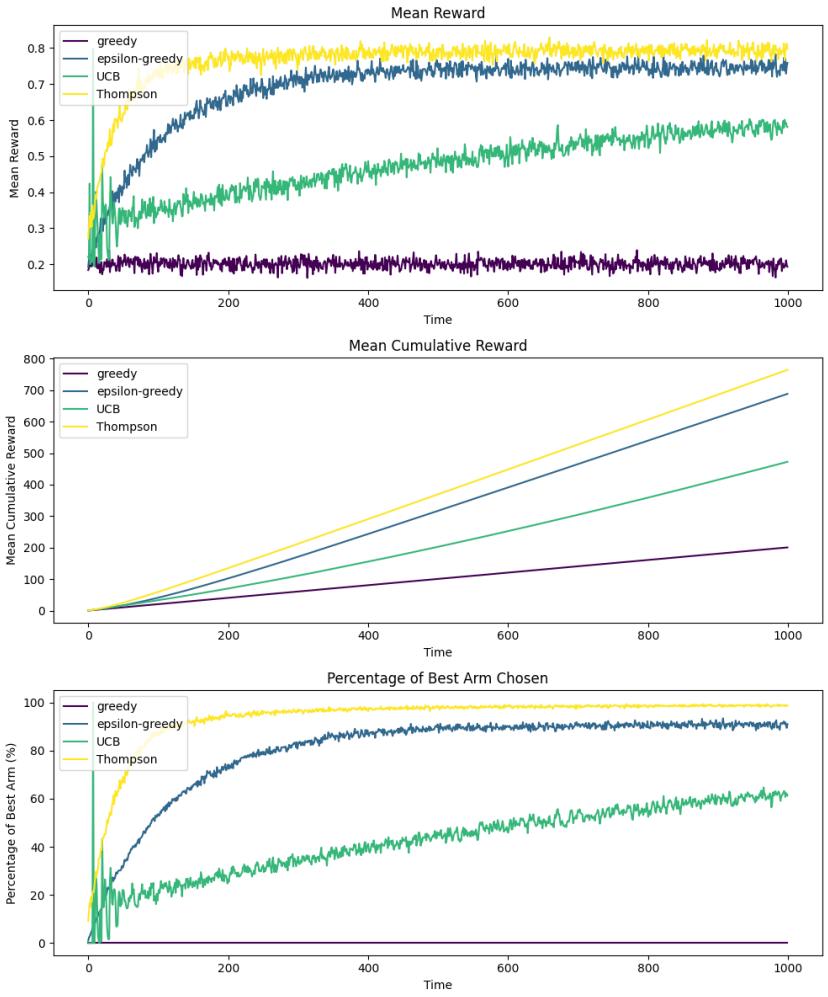
oussama karmaoui Lahmouz Zakaria

November 2023

1 Basic Algorithms

1.1 Question 1

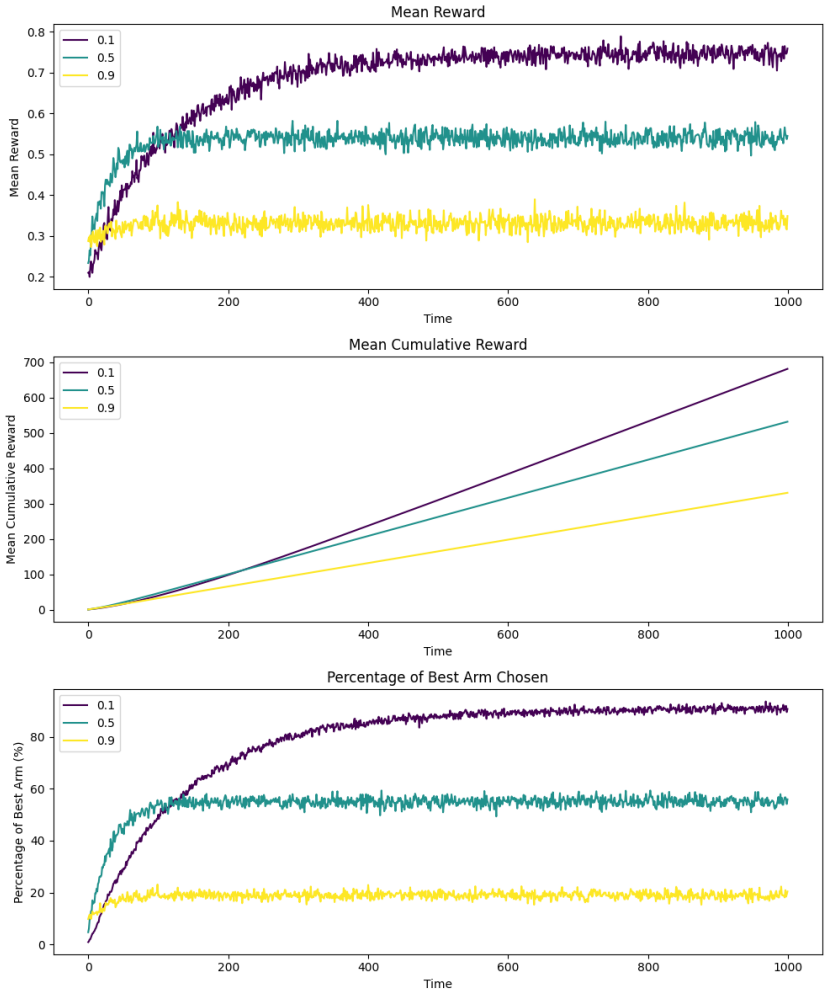
La figure suivante représente ,pour differentes régles(epsilonGreedy , Greedy,UCB,Thompson),la récompense moyenne , le cumul de récompenses et le pourcentage en temps de choisir l'arme optimal



les parametres caractéristiques des differents modèles sont : ϵ vaut 0.1 pour le modèle epsilon greedy, $c = 1$ pour le modèle de thompson et $T = 1000$ le nombre de répetitions.
On remarque qu'au début de la simulation,le modèle de thompson est le plus performant avec un mean reward égal à 0.3 supérieur au mean reward des autres modèle qui est égal à 0.2.Pour $t = 1$ le modele UCB a le plus gros mean reward

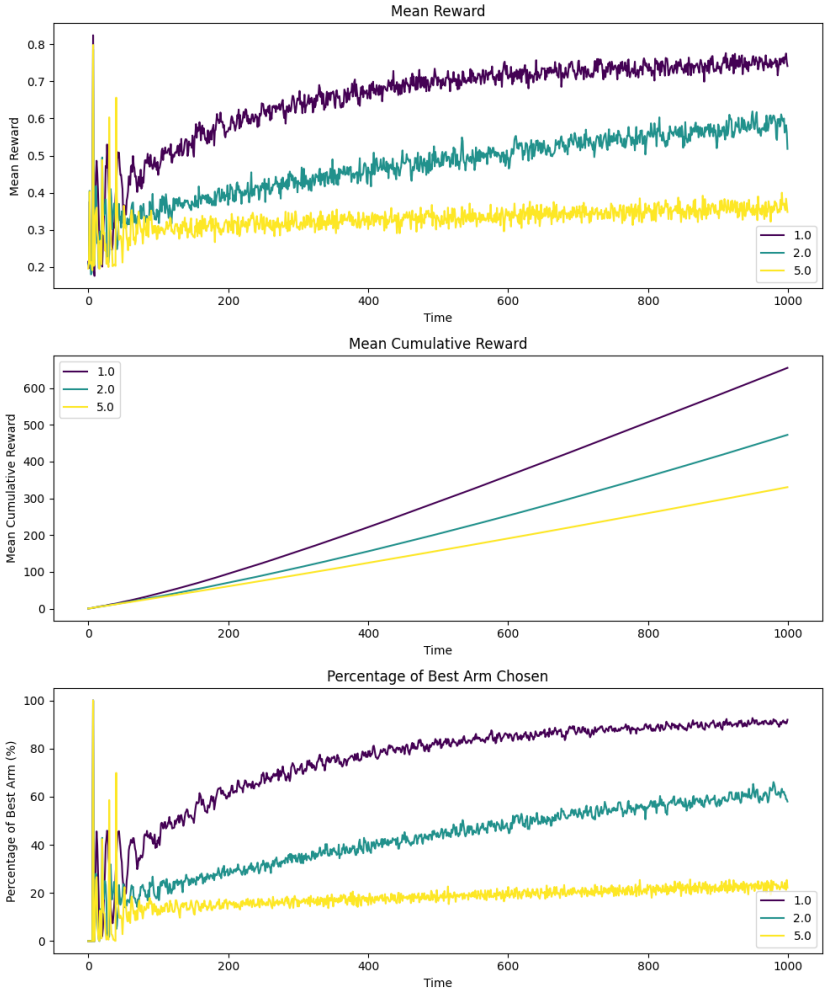
égal à 0.4. Au cours du temps , on remarque que le mean reward du modèle epsilon greedy est croissant,en effet à $t1 = 40$ le mean reward atteint 0.4 , Par contre le modèle greedy est le moins performant avec un mean reward qui oscille autour de 0.25 au fil du temps.En conclusion les modèles thompson et ucb donnent rapidement un reward significatif,le modèle epsilon greedy donne des rewards importants mais ils ne sont pas immediats, contrairement au modèle greedy caractérisé par un reward faible , est cela parait aussi dans le graphe du pourcentage de best arm , qui atteint ses valeurs max (entre 80 et 100) pour modèle Thompson rapidement (t inférieur à 200) .

La figure suivante représente la variation du mean reward , mean cumulative rewards et pourentage de best arm choisi au cours du temps en fonction du parametre ϵ du modèle epsilon greedy. à $t = 0$ le modèle epsilon greedy avec



epsilon égal à 0.9 donne un reward immédiat plus grand que ceux avec epsilon = 0.1 ou 0.5 . Par contre , au fil du temps, le mean reward du modèle avec eps =0.1 ou 0.5 augmente jusqu'à atteindre un mean reward autour de 0.75 (eps =0.1) et 0.70 (eps=0.5).

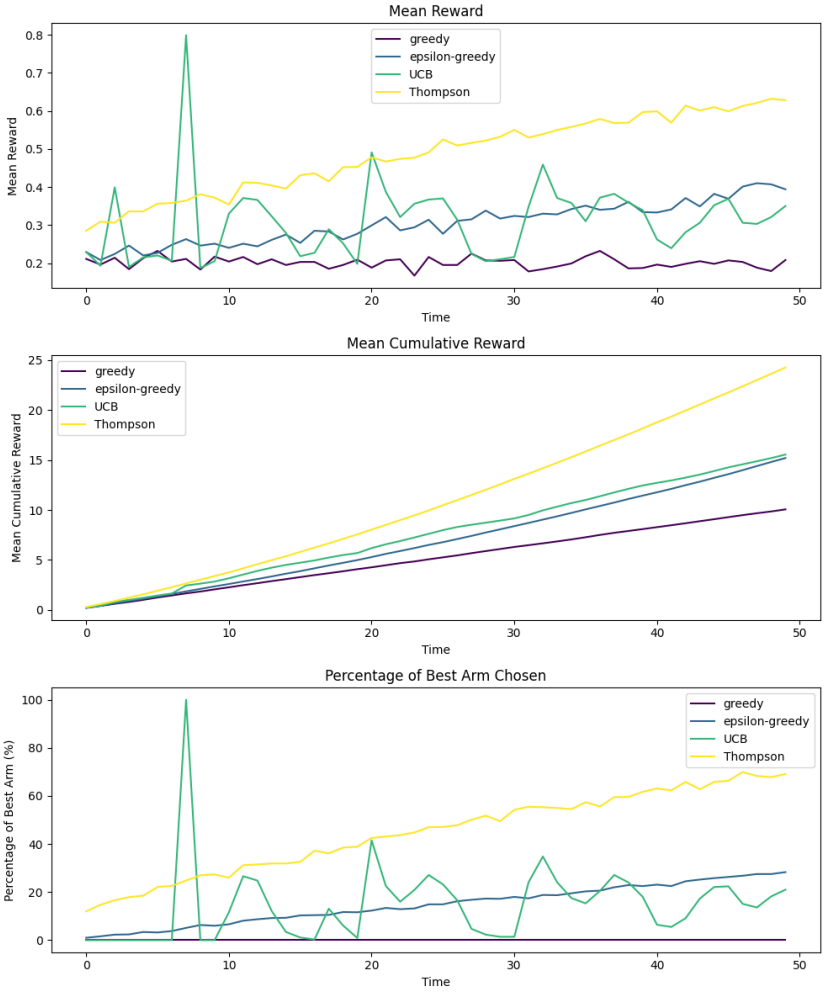
La figure suivante représente la variation du mean reward aucours du temps en fonction du parametre c du modèle ucb. On constate que le modèle ucb le



plus performant est celui avec $c = 1$ et un mean reward autour de 0.75 suivi de celui avec $c = 2$ et enfin en comparant de meme les valeurs de mean cumulative rewards et pourcentage best arm, le modèle le moins performant est caractérisé par la plus grande valeur de c = 5.

la figure suivante montre l'évolution des mean reward en fonction du temps

avec le temps final = 50. On remarque que plus T est petit , plus le mean



reward n'atteint pas les valeurs maximales.Mais à partitr d'un temps à definir , par exemple dans le cas de thompson, on ne peut pas dépasser un seuil reward de 0.7, de meme pour T ¡¡ 1000 , le modèle UCB atteind les valeurs max dans le mean reward (0.8) et le pourcentage d'action optimale (100).

1.2 Question 2

la probabilité de choisir l'action optimale dans une seule étape de la politique -greedy :

$P(optimal) = (1 - \epsilon) + (\epsilon/N)$ ou N est le nombre total d'actions

Ici, (1 - ϵ) est la probabilité de choisir l'action optimale, et ϵ / nombre total d'actions est la probabilité de choisir une action au hasard parmi toutes les

actions disponibles.

alors à l'infini , P(optimal) tend vers 1-ϵ , qui représente bien la probabilité asymptotique de prendre l'action optimale

1.3 Question 3

Pour une valeur de T relativement petite , on a pas assez de temps pour explorer et apprendre de diverses actions, on doit donc tirer le meilleur parti de vos connaissances actuelles pour maximiser les récompenses dans le temps limité dont vous disposez. Donc il faut que ϵ **soit proche de 0** . Pour T suffisemment large , on a plus d'occasions d'explorer différentes actions et d'en apprendre davantage sur leurs valeurs ce qui double ls chances d'avoir une convergence vers les vraies valeurs d'actions , donc le choix de ϵ **proche de 1** va etre plus convenable .

1.4 Question 4

Nous remarquons des pics dans le graphique des mean rewards pour plusieurs raisons : Exploration-Exploitation : Lorsque l'on utilise des stratégies basées sur l'exploration, comme Epsilon-Greedy ou UCB, des pics peuvent se produire lorsque l'agent explore un bras sous-optimal. Pendant l'exploration, l'agent peut occasionnellement sélectionner des bras avec des récompenses attendues plus faibles, ce qui entraîne des baisses temporaires de la récompense moyenne.

Aléatoire : Les problèmes de bandits à plusieurs bras impliquent de l'aléatoire dans les récompenses associées à chaque bras. Même si l'agent suit une stratégie optimale, les récompenses réelles qu'il reçoit peuvent fluctuer en raison de cette aléatoire inhérente. Ces fluctuations peuvent entraîner des pics dans le graphique de la récompense moyenne.

Exploration Initiale : Au début d'une expérience, les agents doivent souvent explorer les bras pour en apprendre davantage sur leurs distributions de récompenses. Cette phase d'exploration initiale peut entraîner des variations dans les récompenses et se traduire par des pics dans le graphique.

Conditions Changeantes : Dans certains cas, les distributions de récompenses des bras peuvent changer au fil du temps. Si l'agent n'est pas conscient de ces changements, il peut continuer à exploiter des bras qui étaient auparavant optimaux mais qui sont maintenant devenus sous-optimaux. Cela peut entraîner des pics dans le graphique des récompenses moyennes.

Environnements Stochastiques : Les environnements de bandits avec des récompenses hautement aléatoires, où les distributions de récompenses des bras sont soumises à des fluctuations fréquentes, sont plus susceptibles de présenter des pics dans le graphique des récompenses moyennes.

Exploration Inadéquate : Si le paramètre d'exploration de l'agent (par exemple, dans Epsilon-Greedy) est réglé trop bas, il peut ne pas explorer suffisamment, ce qui entraîne des sélections de bras sous-optimales et des pics occasionnels.

1.5 Question 5-6

En se basant sur les observations des simulations des différents algorithmes de bandits, nous pouvons tirer les conclusions suivantes en termes de méthodes et fournir quelques intuitions :

Modele Greedy a tendance à converger rapidement vers une solution sous-optimale en exploitant le bras actuellement perçu comme le meilleur sans beaucoup d'exploration.Ces méthodes ne sont pas adaptées aux problèmes avec des récompenses stochastiques ou changeantes, où l'exploration est cruciale. Le modele Epsilon-Greedy trouve un équilibre entre l'exploration et l'exploitation. Il explore avec une probabilité et exploite avec une probabilité de 1-. Les performances dépendent du choix d'. Un plus petit entraîne plus d'exploration, tandis qu'un plus grand se concentre sur l'exploitation. Bien adaptée aux environnements stationnaires avec un certain degré de stochasticité.

Le modele Thompson utilise l'inférence bayésienne pour prendre des décisions probabilistes. Il maintient une distribution a posteriori des récompenses de chaque bras et échantillonne à partir de celle-ci. Fournit une robustesse dans la gestion de l'incertitude et des changements dans les distributions de récompenses. Efficace pour les environnements non stationnaires et incertains. UCB (Upper Confidence Bound) :

UCB utilise des bornes de confiance pour équilibrer l'exploration et l'exploitation. Il donne la préférence aux bras avec un potentiel d'incertitude plus élevé. Le paramètre c contrôle le niveau d'exploration. Des valeurs de c plus élevées favorisent une exploration plus poussée.Efficace pour atteindre un bon équilibre entre exploration et exploitation. Optimistic Greedy initialise des estimations de récompense élevées pour tous les bras, encourageant l'exploration. Elle devient progressivement plus greedy à mesure qu'elle en apprend plus sur les bras. Bien adaptée aux environnements où l'exploration au début est cruciale pour découvrir le meilleur bras. Considérations de Performance :

Le choix de la meilleure méthode dépend du problème spécifique et de ses caractéristiques. Epsilon-Greedy, UCB et Thompson offrent une polyvalence et une adaptabilité à différents environnements. Pour les problèmes aux récompenses stochastiques ou non stationnaires, des méthodes comme Thompson et UCB ont tendance à surpasser les méthodes Greedy et Epsilon-Greedy. Dans les environnements stationnaires, l'efficacité de ces méthodes dépend des paramètres choisis (par exemple, , c). En résumé, le choix de la méthode de bandits doit être dicté par la nature du problème. Si l'environnement est très incertain ou si les distributions de récompenses changent au fil du temps, les méthodes basées sur le Bayésien, comme Thompson Sampling et UCB, sont des choix robustes. Epsilon-Greedy est un bon point de départ lorsque vous devez équilibrer l'exploration et l'exploitation, et le modele greedy n'est pas généralement recommandé pour les problèmes complexes comportant de l'incertitude.

2 Gradient method

En utilisant l'algorithme du gradient représenté par :

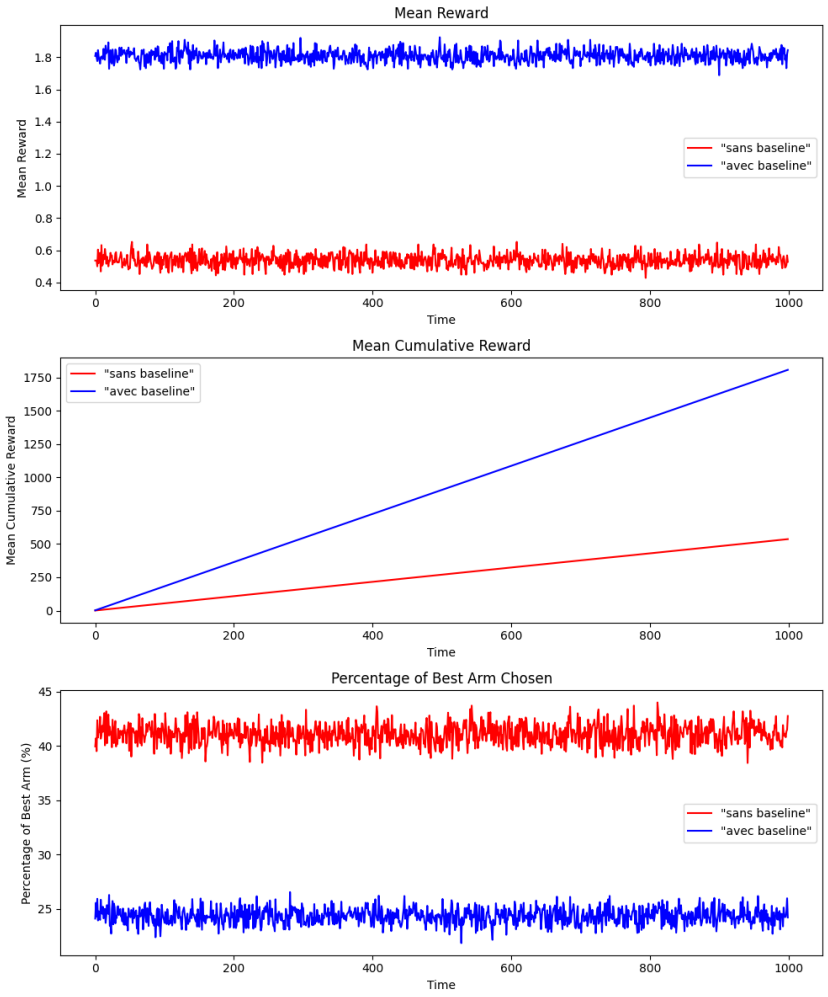


Figure 1: graphe représentant mean rewards, mean cumulative reward pour la méthode du gradient (sans et avec baseline) pour alpha0 = 0.01

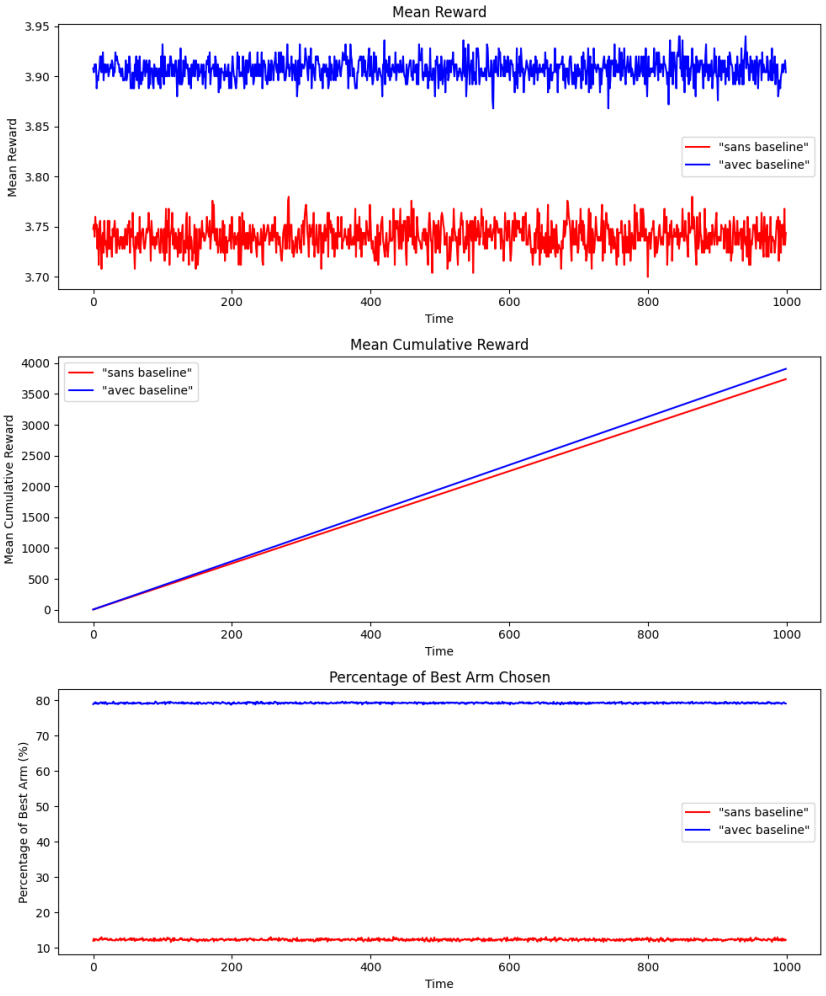


Figure 2: graphe représentant mean rewards, mean cumulative reward pour la méthode du gradient (sans et avec baseline) pour alpha0 = 0.1

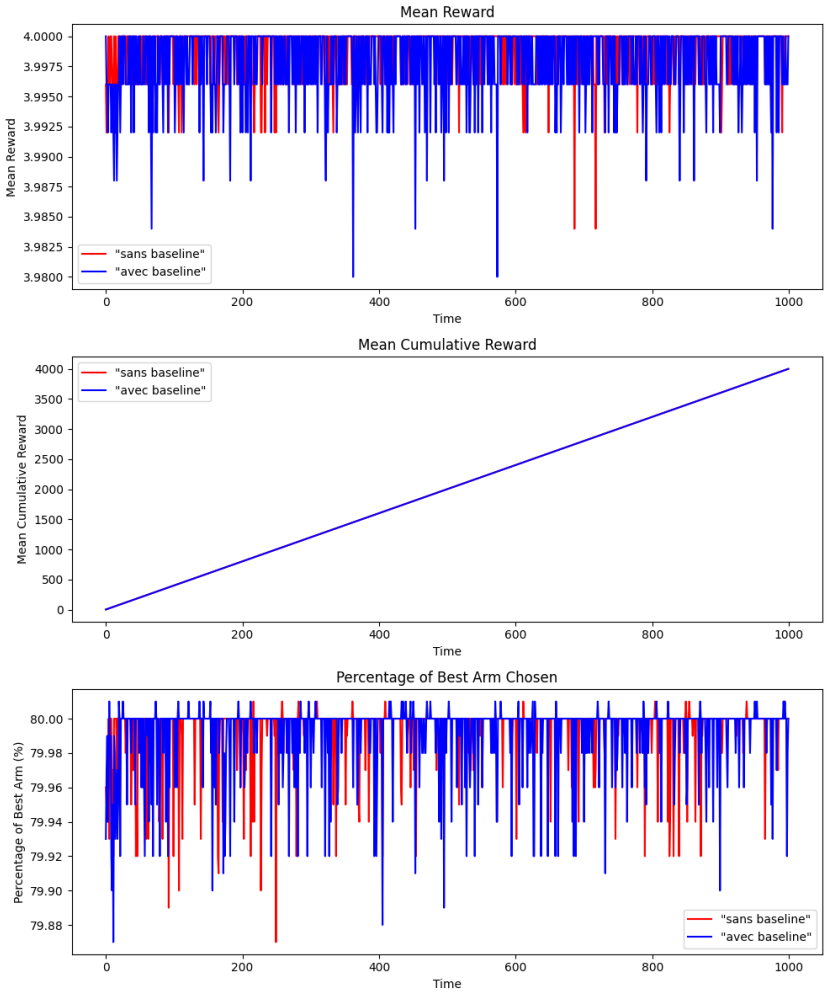


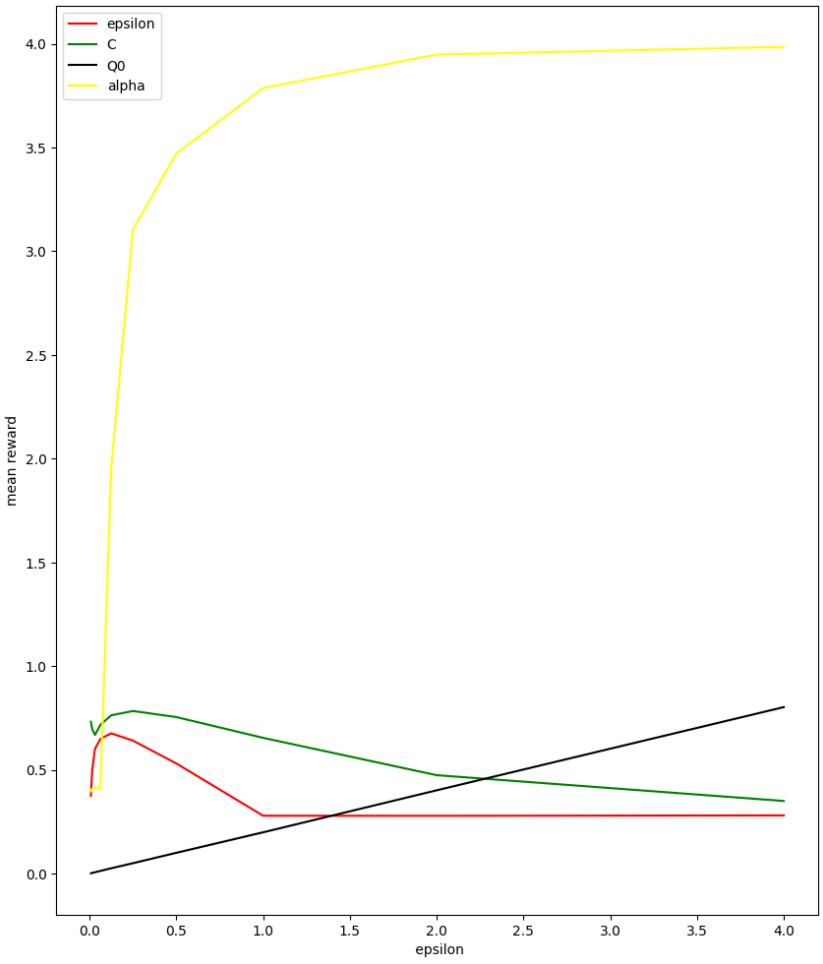
Figure 3: graphe représentant mean rewards, mean cumulative reward pour la méthode du gradient (sans et avec baseline) pour alpha0 = 2

Question 1 : Les 6 figures montrent en grande partie l'évolution de la performance des résultats en choisissant 4*maxk pk comme baseline (pour alpha0 le mean reward est environ 3.9 pour le cas avec baseline qui est plus grand que 3.75 de celle sans baseline), et en augmentant la valeur de alpha0 aussi (valeurs inférieures de mean rewards (entre 0 et 1) pour alpha0 = 0.01, tandis que le mean rewards atteint sa valeur max (environ 4) pour alpha0 = 2), avec l'augmentation de alpha0, les résultats de simulation des cas "sans baseline" et "avec baseline" deviennent très proches entre elles (cas alpha0 0.1 et alpha0 = 2).

Question 2 : La méthode de gradient présente une bonne performance en comparant avec les autres méthodes surtout pour les valeurs moyennes ou petites de steps (T) car en comparant avec les méthodes de base à l'exo 1 , leur mean rewrads commencent de presque 0 jusqu'à une valeur maximale après 200 steps ou 400 , tandis qu'avec la méthode gradient on recoit le mean rewards maximale dès les premières steps , ce qui améliore aussi le mean cumulative rewards et le pourcentage du meilleur bras choisi.

3 Parameter study by Learning curve

On remarque que les deux fonctions des modèles Ucb et thompson sont croissantes sur l'intervale [0,0.2],puis sur [0.3,4] le mean reward diminue en fonction du parametre caracteristique epsilon ou c.
.D'autre part, le modèle gradient donne le meilleur mean reward à peu près égal à 4 avec alpha autour de 2 et la courbe du reward en fonction du parametre caractéristique alpha0 est croissante .De meme pour le model optimistic-greedy mais avec un reward plus faible (il vaut en moyenne 0.75 pour Q0 égal à 4). Pour le modèle epsilon-greedy le meilleur mean reward est de 0.75 avec epsilon égal à 0.25.


epsilon
C
Q0
alpha
mean reward
epsilon
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0