Slide (Expérience 1)

Nous sommes donc maintenant prêts à présenter des résultats préliminaires d’algorithmes adaptés avec ces estimateurs. Ici, nous le faisons pour epsilon\_t greedy avec epsilon décroissant selon la fonction 1 sur racine de t.

On mène ici une expérience d’architecture classique basée sur 200 répétitions.

Pour les algorithmes adaptés, le graphique montre notamment de meilleures performances que celle de son analogue standard basée sur la moyenne empirique. Ainsi, comme attendu, les algorithmes basés sur les nouveaux estimateurs apprennent mieux à discerner l’action optimale.

Aussi, on peut comparer la performance de l’algorithme selon l’estimateur utilisé, on remarque que les performances semblent comparables pour la mediane, l’estimateur du maximum de vraisemblance, et la moyenne tronquée -0.38, alors qu’elle est un peu moins bonne pour l’estimateur L, cela est peut-être dû à nos observations précédentes liées à cet estimateur.

Slide (Expérience 2)

Voici une expérience quasi-identique mais où on généralise en utilisant des paramètres L tirées d’une loi uniforme sur l’intervalle [0,5] à chaque répétition. On voit que les résultats restent similaires.

Slide (Loi de Pareto)

On va maintenant transférer notre regard sur la loi de Pareto. Cette distribution de probabilité possède deux paramètres, notées L et a ici. Dans le cas particulier où $a=1,$ on obtient une loi de probabilité avec une espérance qui n’existe pas (infinie).

Voici le graphique de 3 distributions de Pareto dans ce cas particulier. On peut voir que selon la forme des densités, il est évident que la distribution en vert domine les deux autres. Effectivement, dans ces configurations, il est plus probable d’obtenir une grande observation avec une distribution de paramètre L élevé. Conséquemment, ce paramètre, que l’on nommera ici le paramètre de localisation de la loi, nous servira à établir l’action optimale et le regret dans le contexte des bandits. Par exemple, ici, un agent choisissant de jouer l’action en bleu, plutôt que l’action optimale verte, se verra accorder un regret instantané de 1 à ce pas de temps.

Notons que le paramètre L de la loi de Pareto est en lien avec une mesure de tendance centrale, sa médiane, qui est en fait 2L dans le cas étudié où a=1.

Slide (Les bandits de Pareto avec a = 1)

Dans le contexte de bandits Pareto, on peut donc établir une mesure de performance d’un agent basée sur cette définition de regret, de façon similaire que celle utilisée dans le cas des bandits Cauchy.

Slide (Estimateur de la localisation L d’une loi de Pareto et résultats préliminaires)

Si l’on souhaite adaptée les algorithmes simples classiques pour qu’ils puissent affronter des bandits Pareto à distribution d’espérance infinie, ils doivent se baser sur de bons estimateurs de la localisation $L$ d’une loi de Pareto. On a choisi ici l’estimateur <<minimum>> qui est l’estimateur du maximum de vraisemblance et l’estimateur définit comme la médiane divisée par 2. D’autres estimateurs pourraient être éventuellement envisagés.

On est donc prêt à adapter les algorithmes classiques en leur donnant accès à ces deux outils d’estimation.

Voici les résultats obtenus sur 200 répétitions sur lesquels on a joué epsilon\_t greedy. On peut voir que l’agent utilisant le minimum comme estimateur cumul moins de regret en moyenne et que les deux versions adaptées ont une performance grandement meilleure que la version standard basée sur la moyenne empirique.

Slide (La suite)

On pourrait s’imaginer un agent qui doit choisir entre plusieurs façons de réaliser une tâche dont le temps de réalisation est modéliser par une loi de Pareto, l’agent recevant un reward négatif de module égal au temps prit pour résoudre la tâche soumise par l’action jouée. Tenter de coller l’architecture de notre problème à un problème concret de la sorte pourrait être un des buts poursuivis pour la suite.

Un autre but est sans contredit de tenter d’étendre les généralisations déjà faites à des algorithmes plus complexes, comme les UCB par exemple. Dans la littérature est exploré l’idée de rendre les algorithmes plus robustes dans le cas de bandits à distribution à queues lourdes (non sigma-sous gaussienne), mais qui ont quand même une espérance finie. Donc, nous pourrions peut-être nous en inspirer même si notre problème est autre.

Finalement, notre problème est difficile dans sa définition lui-même. Dans cette présentation, nous avons étudier des configurations de lois de probabilités où l’identification d’une action optimale pouvait se faire clairement et naturellement, rendant la définition d’une mesure de performance, et donc du regret, naturellement généralisée des bandits stochastiques standards. Il faudrait tenter de donner un cadre plus générale et rigoureux à l’identification d’une action optimale.