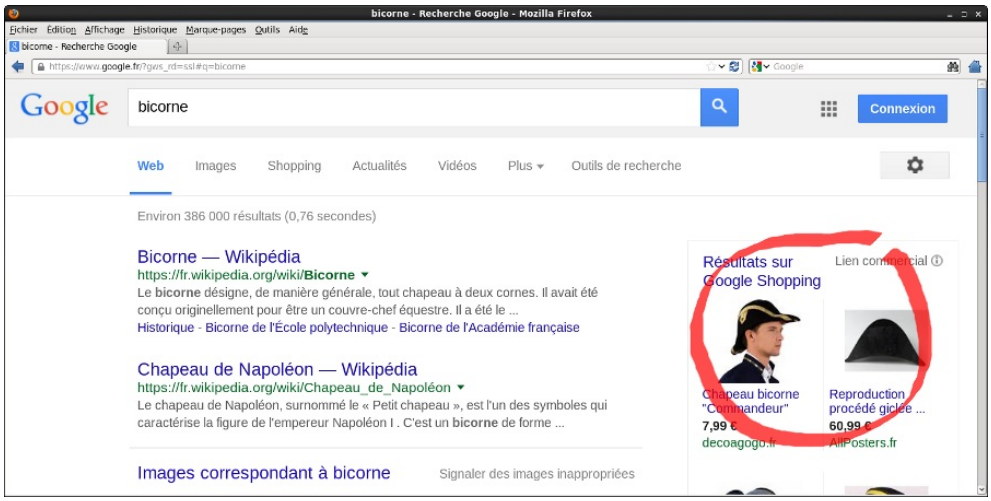
MAP 572 - Mise en oeuvre de méthodes numériques.

François Alouges, Lucas Gerin

Optimisation par apprentissage

1 Apprentissage par renforcement : le bandit à 2 bras



Considérons le problème suivant : un annonceur a le choix d'afficher sur une page web une publicité choisie parmi $\{A, B\}$, l'annonceur est payé au clic et l'objectif est d'afficher la publicité la plus attractive. On modélise le problème de la façon suivante : chaque utilisateur se comporte de façon indépendante des autres et clique sur la publicité A (resp. B) avec probabilité p_A (resp. p_B), on suppose bien sûr que p_A, p_B sont inconnues.

On pose $X_i = 1$ si le i-ème client clique, 0 sinon. On note $E_i \in \{A, B\}$ la publicité affichée sur le site lorsque le i-ème client se connecte, de sorte que

$$X_i \sim \text{Bernoulli}(p_{E_i}).$$

La stratégie E_i à l'instant i est une fonction (éventuellement aléatoire) de $(E_1, X_1), \dots, (E_{i-1}, X_{i-1})$. On cherche à définir une stratégie efficace pour l'annonceur, c'est-à-dire qu'asymptotiquement on propose la meilleure publicité :

$$\frac{X_1 + \cdots + X_n}{n} \overset{n\to+\infty}{\to} \max\{p_A, p_B\}$$

(convergence presque-sûre ou en probabilité). Encore mieux : on souhaite maximiser les gains moyens à horizon fini $\mathbb{E}[X_1 + \cdots + X_n]$.

1.1 Une méthode sous-optimale : la ε-exploration

Considérons la stratégie suivante :
- On choisit $E_1 = A, E_2 = B$.
- Pour $i \geq 3$, on note M_i la publicité qui a eu le meilleur "taux de clic" jusque-là.
  - Avec proba $1 - \varepsilon$, on prend $E_i = M_i$,
  - Avec proba ε on prend $E_i = \text{non}(M_i)$.

(On considère qu'avec probabilité ε on "explore", alors qu'avec probabilité $1 - \varepsilon$ on "exploite".)

Question 1. Intuitivement, quelle est la limite de $(X_1 + \cdots + X_n)/n$? Asymptotiquement, quelle semble être le meilleur choix pour ε ?

Question 2. On fixe $n = 1000$, $p_A = 0.4$, $p_B = 0.6$. Choisir quelques valeurs de ε et tracer par méthode de Monte-Carlo des estimations de courbes

$$i \in \{1, \dots, n\} \mapsto \frac{1}{i}\mathbb{E}\left[X_1 + \cdots + X_i\right].$$

Question 3. Pour le même choix des paramètres, essayer de déterminer la "meilleure" valeur de ε.

1.2 L'algorithme de renforcement de Shapiro-Narendra (*linear reward-inaction*)

On modifie l'algorithme ci-dessus. Soit $c \in (0, 1)$ un paramètre fixé, on note p_i la probabilité d'exploiter A à l'instant i.

- On choisit $p_1 = 1/2$.
- Pour $i \geq 2$,
  - On choisit $E_i = A$ avec probabilité p_i.
  - Si A est choisie et que l'on gagne (*i.e.* $X_i = 1$) alors $p_{i+1} = p_i + \frac{c}{i}(1 - p_i)$.
  - Si B est choisie et que l'on gagne alors $p_{i+1} = (1 - \frac{c}{i})p_i$.
  - Si l'on perd $p_{i+1} = p_i$.

Question 4. On fixe à nouveau $n = 1000$, $p_A = 0.4$, $p_B = 0.6$. On choisit $c = 0.4$, tracer par méthode de Monte-Carlo des estimations de courbes

$$i \in \{1, \dots, n\} \mapsto \frac{1}{i}\mathbb{E}\left[X_1 + \cdots + X_i\right].$$

Question 5. Pour le même choix de paramètres, comparer par simulations les performances de l'algorithme pour différentes valeurs de c.

Question 6. Comment expliquer le terme $\frac{c}{i}$? On pourra par exemple justifier que A, B sont forcément tirés un nombre infini de fois.

Références

[1] V.Rivoirard, G.Stoltz. *Statistiques en Action.* Vuibert (2006).

[2] D.Lamberton, G.Pagès, P.Tarrès. When can the two-armed bandit algorithm be trusted ? *Ann. Appl. Probab.* vol.14 (2004), no. 3, p.1424-1454.