Approches softmax pour les problèmes de bandits

Sébastien Gerchinovitz

Institut de Mathématiques de Toulouse

Introduction

Nous allons présenter des techniques de bandits possédant une certaine robustesse vis-à-vis du modèle stochastique sous-jacent : les approches softmax. Ce cadre est lié à la théorie des suites individuelles.

Plan de l'exposé :

1. Cadre : bandits antagonistes
2. Un algorithme et ses garanties théoriques
3. Quelques extensions du problème initial

Bandits antagonistes à K bras

A chaque date $t \in \mathbb{N}^*$,

1. Le statisticien choisit et révèle un vecteur de poids $\boldsymbol{p}_t = (p_{i,t})_{1 \leqslant i \leqslant K} \in \Delta(K)$ à l'aide des données passées $(I_s, \ell_{I_s,s})$, $1 \leqslant s \leqslant t-1$.
2. Simultanément, le statisticien tire $I_t \sim \boldsymbol{p}_t$ et l'environnement choisit $\boldsymbol{\ell}_t = (\ell_{i,t})_{1 \leqslant i \leqslant K} \in [0,1]^K$.
3. Le statisticien encourt la perte $\ell_{I_t,t}$ mais n'observe pas les $\ell_{i,t}$, $i \neq I_t$. L'environnement observe I_t.

But : mimer la meilleure action sur le long terme, i.e., minimiser le regret

$$\sum_{t=1}^{T} \ell_{I_t,t} - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell_{i,t} \ .$$

Hypothèses stochastiques sur les pertes $\ell_{i,t}$: aucune !

La suite $(\boldsymbol{\ell}_t)_{t \geqslant 1}$ peut être déterministe, voire choisie malicieusement en fonction des données $(I_s, \boldsymbol{p}_s)_{s \leqslant t-1}$ et $\boldsymbol{p}_t$ disponibles à l'instant t.

$\rightsquigarrow$ robustesse vis-à-vis d'un éventuel modèle stochastique.

Prélude : sur la nécessité de convexifier

Nous allons montrer qu'une stratégie de la forme "$\boldsymbol{p}_t = \mathrm{dirac}$" (choix de I_t déterministe en fonction du passé) n'est pas adaptée au cadre antagoniste.

Sans convexification : si les poids sont de la forme $\boldsymbol{p}_t = \delta_{i_t}$, alors la suite $\boldsymbol{\ell}_1, \ldots, \boldsymbol{\ell}_T \in [0,1]^K$ définie par $\ell_{i,t} = \mathbb{I}_{\{i=i_t\}}$ engendre un regret p.s. linéaire en T :

$$\sum_{t=1}^{T} \ell_{I_t,t} - \min_{1 \leqslant i \leqslant K} \sum_{t=1}^{T} \ell_{i,t} \geqslant T - \frac{T}{K} = \left(1 - \frac{1}{K}\right) T \ .$$

Ordre de grandeur du regret :
Ce regret croît linéairement en T... Comme on le verra ci-après, il existe des stratégies (choix séq. des $\boldsymbol{p}_t$) qui assurent un regret en $\mathcal{O}_{\mathbb{P}}\left(\sqrt{T}\right)$.

Agrégation par pondération exponentielle (softmax)

Exemple classique d'algorithme séquentiel introduit par Auer et al. (2002) et inspiré d'une méthode en information parfaite des années 1990.

Algorithme (Algorithme Exp3)

Paramètres : suite décroissante $\eta_1, \eta_2, \ldots > 0$

A chaque date $t \geqslant 1$,

- *A l'aide des données passées, calculer le vecteur de poids* $\boldsymbol{p}_t$ *via*

$$p_{i,t} = \frac{\exp\left(-\eta_t \sum_{s=1}^{t-1} \widetilde{\ell}_{i,s}\right)}{\sum_{j=1}^{K} \exp\left(-\eta_t \sum_{s=1}^{t-1} \widetilde{\ell}_{j,s}\right)}, \quad 1 \leqslant i \leqslant K ;$$

  où $\widetilde{\ell}_{i,s} = \frac{\ell_{i,s}}{p_{i,s}} \mathbb{I}_{\{I_s = i\}}$ *est un estimateur dans biais de* $\ell_{i,s}$.

- *Choisir aléatoirement* $I_t \sim \boldsymbol{p}_t$.

Borne supérieure sur le regret

Théorème (Garanties robustes, Auer et al. 2002)

L'algorithme Exp3 encourt un petit regret en environnement antagoniste : quelle que soit la stratégie de l'adversaire $\ell_1, \ell_2, \ldots \in [0,1]^K$, $\forall T \geqslant 1$,

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{I_t,t}\right] - \min_{1 \leqslant i \leqslant K} \mathbb{E}\left[\sum_{t=1}^{T} \ell_{i,t}\right] \leqslant 2\sqrt{T\,K \ln K}$$

où la borne est obtenue pour le choix de $\eta_t = \sqrt{\ln(K)/(tK)}$.

- En biaisant les estimateurs $\widetilde{\ell}_{i,t}$ et en mélangeant avec l'uniforme, on peut majorer le regret *stricto sensu*, et avec grande probabilité.
- La vitesse $\sqrt{TK \ln K}$ est optimale au sens minimax à un facteur logarithmique près, cf. borne inférieure en $\sqrt{TK}$ de Auer et al. (2002).

Quelques extensions du problème de bandits à K bras

Les algorithmes de type softmax ont originellement été étudiés dans le cadre des suites individuelles, en information parfaite, lorsque toutes les pertes $\ell_{i,t}$, $1 \leqslant i \leqslant K$, sont révélées à la fin de chaque tour t.

La théorie des suites individuelles fournit, en plus de la robustesse, des extensions rapides aux premiers algorithmes de bandits :

1. Bornes adaptatives

   Bornes de regret plus fines que $\sqrt{TK \ln K}$ pour des suites (ℓ_t) plus simples. Cf. par ex la borne en variance de Hazan et Kale (2011).

Quelques extensions du problème de bandits à K bras

Les algorithmes de type softmax ont originellement été étudiés dans le cadre des suites individuelles, en information parfaite, lorsque toutes les pertes $\ell_{i,t}$, $1 \leqslant i \leqslant K$, sont révélées à la fin de chaque tour t.

La théorie des suites individuelles fournit, en plus de la robustesse, des extensions rapides aux premiers algorithmes de bandits :

1. Bornes adaptatives

   Bornes de regret plus fines que $\sqrt{TK \ln K}$ pour des suites (ℓ_t) plus simples. Cf. par ex la borne en variance de Hazan et Kale (2011).

2. Tracking the best expert

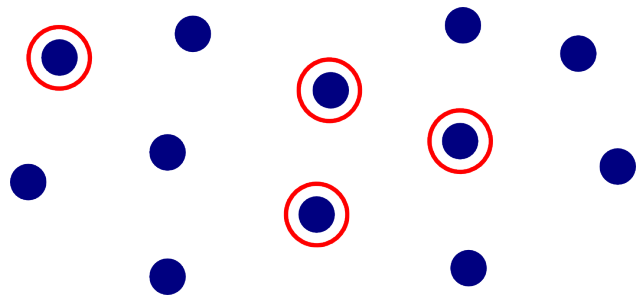
   On se compare maintenant à la meilleure suites d'actions $i_1, i_2, \ldots$ avec au plus s changements :

$$\mathbb{E}\left[\sum_{t=1}^{T} \ell_{I_t,t} - \min_{\substack{i_1,\ldots,i_T \text{ avec} \\ \text{au plus } s \text{ chgts}}} \sum_{t=1}^{T} \ell_{i_t,t}\right] \leqslant C\sqrt{TKs \ln(TK/s)}$$

   Obtenu par Audibert et Bubeck (2010) avec l'algo EXP3.P (efficace).

Un exemple de bandits combinatoires : "multiple plays"

Tâche séquentielle : choisir un ensemble d'actions $\mathcal{S}_t \subset \{1, \dots, K\}$ avec certaines contraintes, et encourir la perte $\sum_{i \in \mathcal{S}_t} \ell_{i,t}$.



Objectif : minimiser le regret

$$\sum_{t=1}^{T} \sum_{i \in \mathcal{S}_t} \ell_{i,t} - \min_{\mathcal{S} \subset \{1,\dots,K\}} \sum_{t=1}^{T} \sum_{i \in \mathcal{S}} \ell_{i,t} \ .$$

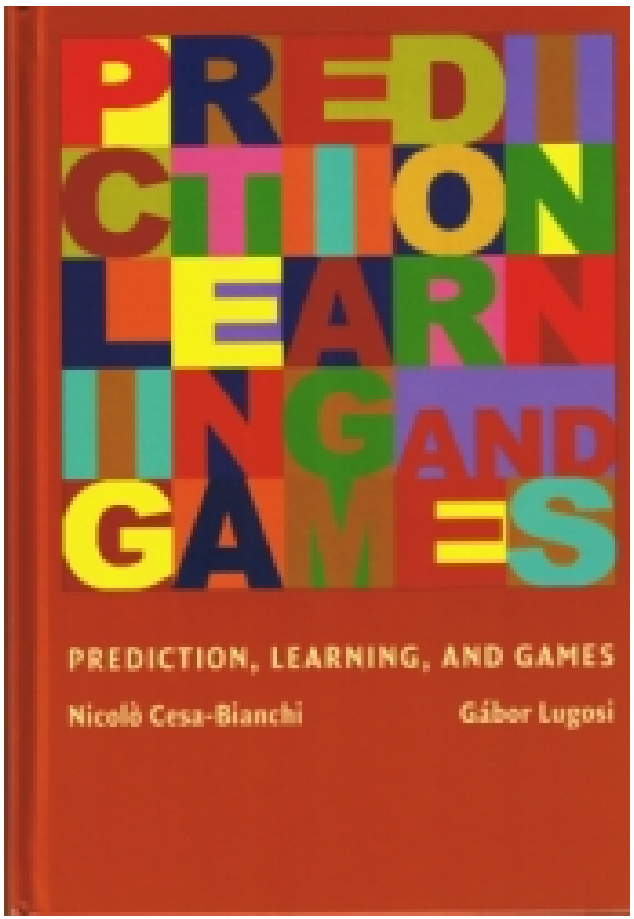
Ex de résultat : avec un algo de descente de gradient généralisée qui exploite statistiquement et algorithmiquement la structure combinatoire,

$$\mathbb{E}[\text{regret}] \leqslant C \sqrt{TKs \ln(K/s)}$$

si on contraint $|S_t| = s$ (cf. Uchiya et al. 2010).

Conseils de lecture

Cet exposé est inspiré de l'article de survol de Bubeck et Cesa-Bianchi (2012) et de l'ouvrage **Prediction, learning, and games** (Cesa-Bianchi et Lugosi, 2006), dont je recommande vivement la lecture :



Bibliographie I

J.-Y. Audibert et S. Bubeck. Regret bounds and minimax policies under partial monitoring. *J. Mach. Learn. Res.*, 11 :2785–2836, 2010.

P. Auer, N. Cesa-Bianchi, Y. Freund, et R.E. Schapire. The nonstochastic multi-armed bandit problem. *SIAM J. Comput.*, 32(1) :48–77, 2002.

S. Bubeck et N. Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1) : 1–122, 2012.

N. Cesa-Bianchi et G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

E. Hazan et S. Kale. Better algorithms for benign bandits. *J. Mach. Learn. Res.*, 12 (Apr) :1287–1311, 2011.

T. Uchiya, A. Nakamura, et M. Kudo. Algorithms for adversarial bandit problems with multiple plays. In Marcus Hutter, Frank Stephan, Vladimir Vovk, et Thomas Zeugmann, editors, *Algorithmic Learning Theory*, volume 6331 of *Lecture Notes in Computer Science*, pages 375–389. Springer Berlin Heidelberg, 2010.