Master Recherche IAC
Option 6
Multi-Armed Bandits

Odalric-Ambryn Maillard
LRI

Jan 9th, 2015

Overview

Rappel, notations

Bandits stochastiques

- $\mathcal{A}$ ensemble d'options
- Si on choisit l'option a, on recoit une récompense X_a tirée selon une distribution ν_a ds [0,1]
- Le meilleur bras est a^* (celui tel que $\mathbb{E}[\nu_{a*}] = max_a \mathbb{E}[\nu_a]$)

Regret

$$R_T = \mathbb{E}[\sum_{t=1}^{T}(X_{a^*} - X_{a_t})]$$

où a_t est l'option choisie au temps t.

$$R_T = \sum_a \Delta_a \mathbb{E}[N_T(a)]$$

où $\Delta_a = \mathbb{E}[\nu_{a*}] - \mathbb{E}[\nu_a]$ et $N_T(a)$ est le nombre de fois qu'on a tiré l'option a jusqu'à T.

Algorithmes de base

1. Stratégie gloutonne — regret linéaire
2. Stratégie ϵ-gloutonne — regret linéaire
3. Stratégie UCB, Upper Confidence Bound — regret logarithmique

$$\text{Select } argmax_a\{\hat{\mu}_{a,t} + \delta(t, N_t(a))\}$$

où δ est une fonction de confiance qui dépend de t (le round courant) et de $N_t(a)$
et $\hat{\mu}_{a,t}$ la moyenne empirique des rewards de a.
Ce résultat est distribution ou problem-dependent (il dépend des gaps Δ_a)

Borne inf (problem-dépendant)

$$\frac{Regret_T}{logT} > \sum_a \frac{\Delta_a}{K_{inf}(\nu_a)}$$

où $K_{inf}(\nu_a)$ caractérise la distribution du bras a.

Outils

- Inégalité de Hoeffding
- Inégalité de Bernstein
- Théorême de Sanov.

Une borne independente du probleme (distribution-free)

$$R_T = \sum_a \Delta_a \mathbb{E}[N_T(a)] = \sum_a \Delta_a \sqrt{\mathbb{E}[N_T(a)]} \sqrt{\mathbb{E}[N_T(a)]}$$

Inégalité de Cauchy Schwartz:

$$\langle x, y \rangle \leq ||x||.||y||$$

$$R_T \leq \sqrt{\sum_a \Delta_a^2 . N_T(a)} \sum_a N_T(a)$$

Le deuxième terme est T. Le premier terme est l'espérance du regret, on avait une borne en $6TAlog(T) +$ cste.

ICI VERIFIER

Overview

Bandits Actifs

Un exemple: on veut mettre des capteurs dans une ville pour estimer la pollution (contexte, IBM et smart cities).
On met des capteurs en certains points.
On a un budget donné de capteurs.
a: une position dans la ville;
On mesure, et au bout de T rounds, on veut minimiser le max (sur les endroits de la ville) de l'erreur de prédiction (entre la vraie valeur et la moyenne empirique), en espérance:

$$max_a \mathbb{E}[(\mu_a - \hat{\mu}_{a,T})^2]$$

Bandits Actifs, 2

Etape 1

$$\mathbb{E}\left[\left(\mu_a - \frac{1}{N_T(a)}\sum_t X_t \mathbb{1}a_t = a\right)^2\right] = \frac{Variance_a^2}{N_T(a)}$$

Sachant qu'on a un budget fini, il faut mettre les capteurs là où la variance est la plus grande. On note σ_a la variance du bras a.

On est dans le meme cadre que pour UCB, à ceci près que $N_T(a)$ doit augmenter en fonction de la variance du bras (et non de la moyenne).

Stratégie oracle

Supposons qu'on sache tout, (en particulier on connait les σ_a) on veut avoir le minimum sur toute les répartitions des $N_T(a)$ (telles que leur somme $= T$) de $max_a \frac{\sigma_a^2}{N_T(a)}$.

Solution

$$N_T^*(a) = \frac{\sigma_a^2 T}{\sum_{a'} \sigma_{a'}^2}$$

Stratégie d'apprentissage

On estime $\hat{\sigma}_a$, on fait une borne de confiance, et on veut s'assurer que σ_a soit dans un intervalle de confiance autour de $\hat{\sigma}_a$
Toujours une stratégie optimiste: on prend la borne sup de l'intervalle de confiance autour de l'estimation empirique.

Estimateur empirique
(Attention, il y a un petit biais).

$$\hat{\sigma}^2_{a,t} = \frac{1}{N_t(a)} \sum (X_{t,a} - \hat{\mu}_{a,t})^2$$

Le regret

Ici, plus on a de points, plus l'erreur d'estimation va être petite.
Vis à vis de la stratégie optimale, on est toujours en $log(T)$.

Remarque

Cette stratégie n'a pas besoin de connaitre l'horizon T: elle est dite anytime.
Si on connait T, on peut en tenir compte et en tirer parti pour améliorer les bornes de l'algorithme.

Overview

Contexte adversarial

$X_{a,t}$ est tiré par un adversaire.
Seule hypothèse: $X_{a,t}$ est dans $\{0, 1\}$.
On regarde les récompenses cumulées optimales :

$$max_a \sum_t X_{t,a}$$

et les récompenses cumulées de l'algorithme :

$$\sum_t X_{t,a_t}$$

Regret

$$\mathbb{E}\left[max_a \sum_t X_{t,a} - \sum_t X_{t,a_t}\right]$$

Ici, une stratégie déterministe est mauvaise (si elle est déterministe, l'adversaire peut vous planter).

Stratégie EXP3

Principe

On a une distribution $p_t(a)$ sur les bras et on tire le bras a_t selon $p_t(a)$; on recoit $X_{s,a}$.
On peut alors construire

$$\sum_s X_{s,a} \mathbb{1}_{a_s=a}$$

Comment ajuster p_t: Importance sampling

Stratégie EXP3, suite

$$p_{t+1}(a) \propto exp \sum_s X_{s,a} \mathbb{I} a_s = a$$

On voudrait pouvoir dire

$$\mathbb{E}\left[\sum_s X_{s,a} \mathbb{I}_{a_s=a}\right] = \sum_s X_{s,a}$$

Le probleme est que a_s est tirée selon la distribution p_s

$$\mathbb{E}\left[\sum_s X_{s,a} \mathbb{I}_{a_s=a}\right] = \sum_s X_{s,a} p_s(a)$$

Stratégie EXP3, suite

Estimateur de la récompense cumulée du bras a:

$$\mathbb{E} \sum \frac{X_{s,a}}{p_s(a)} \mathbb{1}_{a_s=a} = \sum_s X_{s,a}$$

On veut maximiser, on va minimiser $(1 - X_{s,a})$:

$$p_{t+1}(a) \propto exp\left(-\eta_t \sum_s \frac{1 - X_{s,a}}{p_s(a)} \mathbb{1}_{a_s=a}\right)$$

Notons

$$L_t(a) = exp\left(-\eta_t \sum_s (frac1 - X_{s,a} p_s(a) \mathbb{1}_{a_s=a}\right)$$

Alors $p_{t+1}(a)$ est la normalisation des $L_t(a)$:

$$p_{t+1}(a) = \frac{L_t(a)}{\sum_{a'} L_t(a')}$$

Stratégie EXP3, suite

Remarques

- Ceci est encore un algo anytime qui n'utilise pas l'horizon T.
- Il y a un parametre η_t: toujours un sujet de recherche...
- Pourquoi EXP3: Exploration, Exploitation, avec une Exponentielle.

Algo EXP3

Input: la séquence des η_t

Soit p_1: distribution uniforme sur les actions
Loop pour $t = 1 \dots$

- Tirer $a_t \sim p_t$
- Recevoir X_{t,a_t}
- Calculer $\ell_t(a) = \frac{1-X_{t,a_t}}{p_t(a_t)} \mathbb{I}_{a_t=a}$
- Mettre à jour

$$L_t(a) = \sum_{s=1}^{t} \ell_s(a)$$

- Calculer

$$p_{t+1}(a) \frac{exp\left(-\eta_t L_t(a)\right)}{\sum_{a'} exp\left(-\eta_t L_t(a')\right)}$$

Regret de EXP3

Théorême

Supposons $X_{t,a}$ dans $[0,1]$, alors

$$R_T = \frac{3}{\sqrt{2}}\sqrt{TAlog(A)} \text{ pour } \eta_t = \sqrt{\frac{2logA}{At}}$$

Remarque

- $\sqrt{TA}$, aussi ce qu'on avait dans la borne distribution free.
- Borne inf:

$$R_T \geq \frac{1}{20}\sqrt{TA}$$

- En résumé, le regret est d'ordre $\sqrt{TA}$.

La preuve

Etape 1. A l'instant t:

$$\mathbb{E}_{a_t}\ell_{t,a} = \mathbb{E}_{a'}\ell_{t,a'}p_t(a') = \sum_{a'} \frac{1 - X_{t,a'}}{p_t(a')} \mathbb{I}_{a_t=a'} p_t(a')$$

On simplifie $p_t(a')$.

$$= 1 - X_{t,a}$$

Un estimateur non biaisé.

Etape 2

$$\mathbb{E}_{a\sim p_t}[\ell_{t,a}] = 1 - X_{t,a_t}$$

$$Regret = \sum_{t=1}^{T} X_{t,a} - X_{t,a_t} = \sum (1 - X_{t,a_t}) - (1 - X_{t,a})$$

$$= \sum_t \mathbb{E}_{a\sim p_a}\ell_{t,a} - \mathbb{E}_{a_t\sim p_t}\ell_{t,a_t}$$

Controle avec Log Laplace

aussi appelé génératrice des cumulants.

$$LL(X) = \frac{1}{\lambda}\mathbb{E}[e^{\lambda X}]$$

Pourquoi ?

Si on sait controler $LL(X)$, on a le controle des intervalles de confiance. Par ex.,
Si $X \sim \mathcal{N}(0, \sigma^2)$, alors $LL(X) = \lambda\frac{\sigma^2}{2}$

Lemme

$$\mathbb{E}_{a\sim p_t}[\ell_{t,a}] \leq -\frac{1}{\eta_t} log\left(\mathbb{E}_{a\sim p_t} exp\left(-\eta_t \ell_{t,a}\right)\right) + \frac{\eta_t}{2}\mathbb{E}_{a\sim p_t}\mathbb{E}_{a\sim p_t}\left[\ell_{t,a}^2\right]$$

Outils

$$exp(-x) \leq 1 - x - x^2/2$$

$$log(x) \leq x - 1$$

Etape 3

$$\frac{1}{\eta_t} log \mathbb{E}_{a\sim p_t} exp(-\eta_t \ell_{t,a}) = \frac{1}{\eta_t} log \frac{\sum_a exp(-\sum_{s=1}^{t} \eta_t \ell_{s,a})}{\sum_a exp(-\sum_{s=1}^{t-1} \eta_t \ell_{s,a})}$$
$$= \Phi_{t-1}(\eta_t) - \Phi_t(\eta_t)$$

où $\Phi_t(\eta) = -\frac{1}{\eta} log \mathbb{E}_{a\ uniforme} exp(-\eta L_{t,a})$ On voit apparaitre le tirage uniforme: pour nous prémunir contre l'adversaire.

Finalement

$$\sum_{t=1}^{T} \mathbb{E}_{a\sim p_t}[\ell_{t,a}] \leq \sum_t (\Phi_t(\eta_t) - \Phi_{t-1}(\eta_t)) + \sum_t \frac{\eta_t}{2} \mathbb{E}_{a\sim p_t}[\ell_{t,a}^2]$$

Suite

Etape 4

$$\mathbb{E}_{a\sim p_t}[\ell_{t,a}^2] \leq \frac{1}{p_t(a_t)}$$

car les $X_{t,a}$ sont dans $[0,1]$.

Etape 5

On simplifie les $\Phi_t - \Phi_{t-1}$

$$\sum_t \Phi_t(\eta_t) - \Phi_{t-1}(\eta_t) = \Phi_T(\eta_T) + \sum_t \Phi_t(\eta_t) - \Phi_t(\eta_{t+1})$$

Avec:

$$\Phi_t(\eta) = -1/\eta log 1/A \sum_{a'} exp(-\eta L_{t,a'})$$

La somme étant supérieure à chacun de ses éléments (tous positifs), pour tout a,

$$\Phi_t(\eta_t) \leq \frac{1}{\eta_t} log(A) + L_{t,a}$$

En particulier pour l'action optimale.

Suite

Regret

$R_T \leq log(A)/\eta_T + \mathbb{E}[\frac{1}{p_t(a_t)} +$ terme des sommes $\Phi_t(\eta_t) - \Phi_t(\eta_{t+1})$

On admet que le terme droit est négatif.

$$\leq log(A)/\eta + \eta AT/2$$

On optimise en η, ce qui donne $\eta = \sqrt{2log(A)/AT}$ d'où le regret:

$$R_T < \sqrt{ATlogA}$$

Questions

- Peut-on faire mieux que de prendre un η constant ?
- On a optimisé η en supposant qu'on connaissait l'horizon T.

Si on ne connait pas l'horizon, ce qui change est le terme somme des $\Phi(\eta_t) - \Phi(\eta_{t+1})$. On peut montrer que la dérivée de Φ est négative. Si donc η_t décroit avec t (car on a de plus en plus d'information au cours du temps, donc il est moins besoin d'etre agressif), on choisit:

$$\eta_t = \sqrt{\frac{2logA}{At}}$$

Questions

Q. Ne marche pas du tout sur les cours de la Bourse ...
R. Cet algorithme est supposé etre robuste: marche dans le cas le pire. (regret controlé dans le pire cas).
La Bourse: en fait, pas si méchant / adversarial que ca...
D'où un comportement trop conservatif de EXP3.
UCB: va aussi se planter pour la bourse (car le contexte est un peu adversarial quand meme) ...
Les deux algos sont bons dans des régions différentes.

On essaye de faire des algorithmes pour la région intermédiaire.

- Estimer l'environnement: iid ou pas ?
- Faire des algo intermédiaires (poids exponentiels ajustés)

Challenge Yahoo

Input

Un utilisateur arrive, décrit par son contexte, x_t dans $\{0,1\}^{100}$.
On veut lui montrer une annonce qui lui plait.
Qui a gagné: une variante de UCB.