Bandit Problems

The Accounts algorithm and thoughts on exploration

Tolga Yenisey

Multi-armed Bandit problem

- Each round i of T the player chooses arm j of K and receives a payout of $p_{i,j}$ (alternatively, pays cost $c_{i,j}$)

- Adaptive (adversarial): payouts (or costs) of each round are dependent on the outcomes of the previous rounds
- Non-adaptive (stochastic): payouts (or costs) of each round are independent of any previous rounds

- Full information: The player receives the payout from the arm he chose, but also learns the payouts of all other arms each round
- Bandit: The player only receives the payout of the arm chosen, and learns nothing of the payouts of the other arms

EXP3

(Auer et al. 2002b)

EXP3(K)

$\mathsf{p}_1 \leftarrow (\frac{1}{K}, \ldots, \frac{1}{K})$

for $t \leftarrow 1$ **to** T **do**

  $\textsc{Sample}(I_t \sim \mathsf{p}_t)$

  $\textsc{Receive}(\ell_{I_t,t})$

  for $i \leftarrow 1$ **to** K **do**

    $\widetilde{\ell}_{i,t} \leftarrow \frac{\ell_{i,t}}{p_{i,t}} 1_{I_t = i}$

    $\widetilde{L}_{i,t} \leftarrow \sum_{s=1}^{t} \widetilde{\ell}_{i,s}$

    $p_{i,t+1} \leftarrow \frac{e^{-\eta \widetilde{L}_{i,t}}}{\sum_{j=1}^{K} e^{-\eta \widetilde{L}_{j,t}}}$

return p_{T+1}

EXP3 (Exponential weights for Exploration and Exploitation)

EXP3 algorithm

- High probability (at least 1- ε) regret of the form $O(\sqrt{TKlog(TK/\varepsilon)})$
- Accounts: with probability at least 1-ε, $O(\sqrt{TKlogK} * log\frac{1}{\varepsilon})$
- Improvement achieved with better balance of exploration vs exploitation
- The Accounts algorithm is a refinement of EXP3

Accounts algorithm

- Motivation:
  Explore any given arm until enough confident enough that it’s a poor choice to overcome regret due to variance of exploration at low probability

  “Absorb” the cost of exploring poor choices to increase likelihood of better payouts in the remaining rounds
- For each arm have an account, or allowance, for exploration. If the exponential weighting would reduce the probability that an arm is chosen below a certain amount, and it still has allowance for exploration, keep the probability and take from the account

Let $S \subset \mathbb{R}^K$ denote the simplex of probability distributions over $\{1, \ldots, K\}$. Our algorithm is defined in terms of two functions $f : \mathbb{R}^K \to S$ and $g : \mathbb{R}_{\geq 0} \to [0, 1]$. The boldface variables are vectors in $\mathbb{R}^K$.

Algorithm 3.1: ACCOUNTS(f, g)

$\widehat{\mathbf{C}} := \mathbf{A} := \mathbf{0}$.
for $i := 1$ **to** T
  Set $\mathbf{p} = (p_1, \ldots, p_K) = f(\widehat{\mathbf{C}})$.
  Sample $M = M^i$ from $1, \ldots, K$ according to the distribution $\mathbf{p}$.
  Pull arm M. Observe and incur cost c^i_M.
  if $g(A_M) \leq p_M$
   then $\widehat{\mathbf{C}}_M := \widehat{\mathbf{C}}_M + \frac{c^i_M}{p_M}$
   else $A_M := A_M + \frac{c^i_M}{p_M}$

Henceforth, we will work with the following specific choice of f. Let $\eta = \sqrt{\ln K / TK}$. For $z = (z_1, \ldots, z_K) \in \mathbb{R}^K$, and $j \in \{1, \ldots, K\}$, let

$$f_j(z) = \frac{e^{-\eta z_j}}{\sum_{\ell=1}^{K} e^{-\eta z_\ell}}.$$

We define our barrier function g by

$$g(x) = \max\left\{\eta, \frac{1}{K(1 + x/\theta)^{3/2}}\right\}, \quad \theta = \sqrt{KT \ln K}$$

Proof Overview

- First establish that the account value A_j^T rarely underestimates by much the contribution of “stepwise variance” to R_j
- Then show that the contribution of “stepwise expectations” to $R_j + A_j^T$ cuts off sharply at $O(\sqrt{TKlogK})$
- The result follows from these two claims

Theorem 1.1. *Let R denote the regret of the “Accounts” algorithm for the K-armed bandit, on any adaptively chosen cost sequence of length T. Then, for every $\alpha > 1$,*

$$\mathbf{Pr}\left(R \geq (\alpha + 7)\sqrt{TK \ln K}\right) \leq 1000K\sqrt{\alpha}\exp\left(-\frac{\sqrt{\alpha}\log K}{8}\right).$$

It follows that

$$\mathbf{E}(R) = O(\sqrt{TK \ln K}).$$

Notation

Let $R_j^i = \sum_{\ell=1}^{i} c_{M^\ell}^\ell - c_j^\ell$. denote the regret with regards to arm j at time i

Note that this gives us a new formula for the final regret, R, namely,

$$R = \max_{1 \le j \le K} R_j^T.$$

For $j \in \{1, \ldots, K\}$, let Φ_j denote the following function from $\mathbb{R}^K \to \mathbb{R}$.

$$\Phi_j(\mathbf{z}) := \frac{1}{\eta} \ln \frac{1}{f_j(\mathbf{z})} = \frac{1}{\eta} \ln \frac{\sum_{\ell=1}^{K} e^{-\eta z_\ell}}{e^{-\eta z_j}} = z_j + \frac{1}{\eta} \ln \left(\sum_{\ell=1}^{K} e^{-\eta z_\ell} \right)$$

This definition implies that, for each j, $\nabla \Phi_j = \mathbf{e}_j - f$, *i. e.*, for each i, j,

$$\frac{\partial}{\partial z_i} \Phi_j(\mathbf{z}) = \begin{cases} 1 - f_i(\mathbf{z}) & \text{if } i = j \\ -f_i(\mathbf{z}) & \text{otherwise.} \end{cases}$$

Notation continued

Define
$$\Delta R_j^i := R_j^i - R_j^{i-1}$$
$$\Delta \Phi_j^i := \Phi_j^i - \Phi_j^{i-1}$$
$$\Delta A_j^i := A_j^i - A_j^{i-1}.$$

Denote by H_i the history of the game prior to round i

$$Y_j^i := \Delta R_j^i + \Delta \Phi_j^i + \Delta A_j^i - \mathbf{E}\left(\Delta R_j^i + \Delta \Phi_j^i + \Delta A_j^i \,|\, \mathcal{H}_i\right),$$
$$Y = Y_j := \sum_{i=1}^{T} Y_j^i.$$

Note that Y_j^i is a martingale difference sequence

Proof, given

Lemma 4.1. *Let $1 \le j \le K$. Then, for every $\alpha \ge 1$,*

$$\mathbf{Pr}\left(Y_j - A_j^T > (\alpha+1)\sqrt{TK\ln K}\right) \le \left(\frac{16\sqrt{\alpha}}{\ln K} + \frac{128}{\ln^2 K}\right)\exp\left(-\frac{\sqrt{\alpha}\ln K}{8}\right)$$

Lemma 4.2.

$$\mathbf{Pr}\left(\exists j \sum_{i=1}^{T} \mathbf{E}\left(\Delta R_j^i + \Delta \Phi_j{}^i + \Delta A_j^i \,|\, \mathcal{H}_i\right) > 6\sqrt{TK\ln K}\right) \le \exp\left(\frac{-3\sqrt{TK\ln K}}{26}\right)$$

Assume wlog that $(\alpha+7)\sqrt{TKlnK} < T$, then by definition of Y_j

$$Y_j = \sum_{i=1}^{T} Y_j^i = R_j^T - R_j^0 + \Phi_j^T - \Phi_j^0 + A_j^T - A_j^0 - \sum_{i=1}^{T} \mathbf{E}\left(\Delta R_j^i + \Delta \Phi_j{}^i + \Delta A_j^i \,|\, \mathcal{H}_i\right)$$

Since $R_j^0 = A_j^0 = 0$ and $\Phi_j^0 - \Phi_j^T \le \Phi(0) = \frac{\ln K}{\eta} = \sqrt{TK\ln K}$, this implies

$$R_j^T \le Y_j - A_j^T + \sum_{i=1}^{T} \mathbf{E}\left(\Delta R_j^i + \Delta \Phi_j{}^i + \Delta A_j^i \,|\, \mathcal{H}_i\right) + \sqrt{TK\ln K}$$

Proof cont.

By lemmas 4.1 and 4.2, we have $R \leq \max_i R_j^T \leq (\alpha + 7)\sqrt{TK \ln K}$

And summing error probabilities completes the proof for the tail inequality

$$\mathbf{Pr}\left(R \geq (\alpha+7)\sqrt{TK\ln K}\right) \leq K\left(\frac{16\sqrt{\alpha}}{\ln K} + \frac{128}{\ln^2 K}\right)\exp\left(-\frac{\sqrt{\alpha}\ln K}{8}\right) + \exp\left(\frac{-3\sqrt{TK\ln K}}{26}\right).$$

$$\mathbf{Pr}\left(R \geq (\alpha+7)\sqrt{TK\ln K}\right) \leq \frac{200K\sqrt{\alpha}}{\ln K}\exp\left(-\frac{\sqrt{\alpha}\ln K}{8}\right)$$

To prove the upper bound on expectation, we note that, in general,

$$\mathbf{E}(R) \leq \mathbf{E}(\max\{R, 0\}) = \int_0^\infty \mathbf{Pr}(R \geq x)\mathrm{d}x.$$

The desired bound $\mathbf{E}(R) = O(\sqrt{TK\ln K})$ follows

Needed theorems

Theorem 5.1 (McDiarmid). *Suppose $X_1, \ldots, X_n$ is a martingale difference sequence, and b is an uniform upper bound on the steps X_i. Let V denote the sum of conditional variances,*

$$V = \sum_{i=1}^{n} \mathbf{Var}\,(X_i \mid X_1, \ldots, X_{i-1}).$$

Then, for every $a, v \geq 0$,

$$\mathbf{Pr}\left(\sum X_i \geq a \text{ and } V \leq v\right) \leq \exp\left(-\frac{a^2}{2v + 2ab/3}\right).$$

Theorem 5.2. *Suppose $X_1, \ldots, X_n$, V, are as in Theorem 5.1. Let B denote the maximum "conditional positive deviation,"*

$$B = \max_i \sup(X_i \mid X_1, \ldots, X_{i-1})$$

Then, for every $a, b, v \geq 0$,

$$\mathbf{Pr}\left(\sum X_i \geq a \text{ and } V \leq v \text{ and } B \leq b\right) \leq \exp\left(-\frac{a^2}{2v + 2ab/3}\right).$$

References

- [1] How to Beat the Adaptice Multi-Armed Bandit, 2006, V. Dani and T.P. Hayes, http://people.cs.uchicago.edu/~hayest/papers/AdaptiveBandit/AdaptiveBandit.pdf
- [2] V. Dani and T. P. Hayes. Robbing the bandit: less regret in online geometric optimization against an adaptive adversary. To appear in SODA 2006.
- [3] THE NONSTOCHASTIC MULTIARMED BANDIT PROBLEM∗, Auer et al, 2002