A Proof of Lemma 2

Proof. Consider two supplies of arms T and B, such that for every arm generated in T there is an arm generated in B having exactly the same parameters except that the lifetime of the arm in T is timed and the arm in B is budgeted. Consider an execution of an algorithm AT using arms generated by T. An algorithm AB can executes exactly the same steps as AT using the corresponding arms generated by B. Algorithm AB can do that since it pulls an arm only when AT pulls the corresponding arm, an therefore arm in B will never expire before the corresponding arm in T expires. Clearly the execution of AB has the same reward as the corresponding execution of AT. □

B Proof of Lemma 3

Proof. An algorithm AD for the state-aware case imitates the execution of algorithm AS for the state-oblivious case as follows: (1) the first arm pull of AD is identical to that of AS, and (2) when AD receives a reward μ it decides with probability μ to consider it as 1 and otherwise as 0. It then follows the decision that algorithm AS does with this random value. The expected reward of algorithm AD is equal to that of algorithm AS. □

C Proof of Corollary 4

Proof. The arm payoff values are 1 and $1-\delta$, and the corresponding $\Gamma(\mu)$ values are:

$$\begin{aligned}\Gamma(1-\delta) &= 1-\delta+\delta p,\\ \Gamma(1) &= q(1-\delta+\delta p+1-p) < 1-\delta+\delta p\end{aligned}$$

Thus, $\Gamma(\mu^*) = 1-\delta+\delta p$. □

D Proof of Corollary 5

Proof. Let $0 \le \mu \le 1$. Now, $E[X] = 1/2$, $F(\mu) = \mu$, and $E[X \mid X > \mu] = \frac{1+\mu}{2}$ for the uniform distribution. Also the expected lifetime is $L = 1/p$. Hence,

$$\Gamma(\mu) = \frac{\frac{1}{2} + (1-\mu)(\frac{1}{p}-1)\frac{1+\mu}{2}}{1+(1-\mu)(\frac{1}{p}-1)}.$$

The maximum of this function is given by

$$\Gamma(\mu^*) = \frac{1-\sqrt{p}}{1-p}.$$

The optimal reward per timestep is 1, since we have infinite arms drawn from $U(0,1)$. Hence, the expected regret per timestep of any algorithm is bounded by

$$1 - \frac{1-\sqrt{p}}{1-p} = \frac{\sqrt{p}-p}{1-p} = \Omega(\sqrt{p}).$$

□

E Proof of Theorem 6

Proof. We partition the execution time of the algorithm into disjoint intervals, each interval corresponds to one execution of the while loop. Let s be a random variable denoting the length of an interval. Let $\tau(s)$ be a random variable denoting the number of fresh arm pulls during an interval, and $s-\tau$ denotes the number of repeat pulls to the arm with reward at least μ^*.

The expected reward per step during this interval is given by

$$\frac{1}{s}(\tau(s)E[X] + (s - \tau(s))E[X \mid X \geq \mu^*]). \tag{1}$$

Instead of evaluating the expectation of the above expression directly, we model the execution of the algorithm as a two state renewal process. In state I the algorithm is searching for an arm with reward at least μ^*, in state II it pulls that arm till it expires. The length of a state I interval has a geometric distribution with expectation $\frac{1}{1-F(\mu^*)}$. The length of a state II interval has a expectation $L-1$ (one pull of the arm was counted in the state I interval). The random variables corresponding to disjoint intervals are independent. We use the fundamental limit theorem for alternating renewal processes to obtain the fraction of time in the limit that the process spends in each of its two states:

Theorem 7 (Two-state renewal process)**.** *Suppose a system can have two possible states E_1 and E_2, with E_1 being the initial state. The system alternates between E_1 and E_2, spending discrete timesteps in each. Let the times spent in E_1 be given by the random variables X_j with a common distribution F_1, and in E_2 by the random variables Y_j with a common distribution F_2. Let all random variables be independent of each other (they can depend on the state). Let $E[X_j] = \mu_1 < \infty$ and $E[Y_j] = \mu_2 < \infty$. Then the asymptotic fraction of time spent in E_1 and E_2 are:*

$$f_1(t) \rightarrow \frac{\mu_1}{\mu_1 + \mu_2}, f_2(t) \rightarrow \frac{\mu_2}{\mu_1 + \mu_2}.[1]$$

Applying the above theorem we have that

$$\lim_{t\to\infty} \frac{\tau(s)}{s} = \frac{\frac{1}{1-F(\mu^*)}}{\frac{1}{1-F(\mu^*)} + L - 1},$$

and

$$\lim_{t\to\infty} \frac{s-\tau(s)}{s} = \frac{L-1}{\frac{1}{1-F(\mu^*)} + L - 1}.$$

Plugging these limits to the expectation (1) we get

$$\begin{aligned} & \lim_{t\to\infty} E[\frac{1}{s}(\tau(s)E[X] + (s - \tau(s))E[X \mid X \geq \mu^*])] \\ = \; & \frac{E[X] + (1 - F(\mu^*))(L-1)E[X|X > \mu^*]}{1 + (1 - F(\mu^*))(L-1)} \\ = \; & \Gamma(\mu^*). \end{aligned}$$

□

References

[1] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 2*. Wiley, 1971.