Theorem 1 (Two-stage 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:*

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

Theorem 2. *Let every arm i expire exactly after L_i pulls, where $E[L_i] = L$ for all i. Let the number of arms k be infinite. Let $\bar{\mu}(t)$ denote the maximum mean reward that any algorithm for the deterministic mortal multi-armed bandit problem could obtain after running for t time steps, and* DetOpt*(t) denote the mean reward obtained by* DetOpt *after running for t time steps. Then, (1) $\lim_{t\to\infty} \bar{\mu}(t) = \mu^*$, and (2) $\lim_{t\to\infty} DetOpt(t) = \mu^*$.*

Proof. We prove (1) first. The proof has two parts: (a) we show that the optimal algorithm (OPT) either pulls an arm only once, or until it expires, and then (b) we show that the expected reward of such an algorithm tends to μ^* as $t \to \infty$.

After t timesteps, OPT makes a series of arm pulls, some t' of which are *fresh* arm pulls, that is, pulls on arms that have never been pulled before. These fresh arms could have existed from the beginning (*original* arms), or they could have appeared after some existing arm was pulled till expiry (*replacement* arms). Now, pulling a fresh replacement arm yields identical expected reward per timestep as pulling a fresh original arm, since the payoff and lifetime distributions are identical for both (as is the joint distribution of the two, because of their independence) and k is large enough that fresh original arms are always available. Thus, we can assume without loss of generality that OPT only pulls original arms.

Now, consider an algorithm A that simulates OPT, as follows. Every timestep, A increments a counter and then sees which arm OPT pulls. If OPT pulls a fresh arm, A does so as well, otherwise A merely records the arm that was pulled. Once the counter reaches the time horizon t, A pulls all the arms that it had recorded. Note that this sequence of arm pulls is valid: pulls of original arms can be arbitrarily reordered under budgeted death, as the timing of arm pulls has no effect on expiry. Also, the reward of A is identical to that of OPT.

Let μ' be the highest value among all the arms that OPT (and hence, A) did not pull to expiry. Clearly OPT should pull arms with value less than or equal to μ^* only once, otherwise A could exchange a repeat pull on some arm i of value $\mu^i \leq \mu'$ with an extra pull of the arm yielding μ' (both arms are unexpired, and their values known from the initial round of fresh arm pulls), thus performing better than OPT. Similarly, we can show that OPT must pull all arms with value greater than μ^* to expiry. Thus, OPT either pulls an arm only once, or until it expires, proving part (a).

Since arbitrary reordering of original arm pulls is allowed, we can think of OPT as alternating between two states: pulling fresh arms one by one till it finds one with value greater than μ', and repeatedly pulling this arm till it expires. The times spent in the first state are independent random variables with expectation $\frac{F(\mu')}{1-F(\mu')}$, while those in the second state are again independent

with expectation L. By Theorem 1, the fractions of time spent in each state are:

$$f_1 \to \frac{\frac{F(\mu')}{1-F(\mu')}}{\frac{F(\mu')}{1-F(\mu')} + L}, f_2 \to \frac{L}{\frac{F(\mu')}{1-F(\mu')} + L}.$$

Hence, the overall expected reward per timestep is asymptotically

$$\begin{aligned}
\lim_{t\to\infty} \bar{\mu}(t) &= \frac{\frac{F(\mu')}{1-F(\mu')} E[X|\ X \le \mu'] + L \cdot E[X|X > \mu']}{\frac{F(\mu')}{1-F(\mu')} + L} \\
&= \frac{E[X] + (1-F(\mu))(L-1)E[X|X>\mu]}{1 + (1-F(\mu))(L-1)} \\
&= \Gamma(\mu') \le \mu^*.
\end{aligned}$$

Where the last inequality follows from the definition of μ^*. Thus, $\lim_{t\to\infty} \bar{\mu}(t) = \mu^*$ for OPT, proving (1).

To prove (2) we notice that the algorithm is using the optimal threshold that maximizes $\Gamma(\mu)$ and thus matches the lower bound value. □

Lemma 3. *Any algorithm for the timed death case obtains at least the same expected reward per timestep in the budgeted death case.*

Proof. For any given time horizon t, at most t distinct arms can be pulled. Denote the lifetimes of these t arms by $L_1, L_2, \ldots, L_t$.

Consider any one particular instantiation of $L_1, \ldots, L_t$. Consider an ordered list $\mathcal{S}$ of arm pulls under the timed death model. Suppose that under $\mathcal{S}$, arm i is pulled for the first time at $t_i^{(f)}$ and for the last time at $t_i^{(\ell)}$. Then, $L_i \ge t_i^{(\ell)} - t_i^{(f)}$. Now, $t_i^{(\ell)} - t_i^{(f)}$ is also the maximum number of times arm i could have been pulled, implying that arm i would be alive at time $t_i^{(\ell)}$ under the budgeted death model as well. Thus, identical pulls of arm i could have been performed under budgeted death. Indeed, $\mathcal{S}$ is a valid order of arm pulls under budgeted death, and yields the same reward. Thus, for any given instantiation of lifetimes $L_1, \ldots, L_t$, the optimal expected reward $E[R_{timed}(t, L_1, \ldots, L_t)]$ under timed death is less than or equal to that under budgeted death $E[R_{budgeted}(t, L_1, \ldots, L_t)]$.

Averaging over all possible lifetimes, we get

$$\begin{aligned}
E[R_{timed}(t)] &= \sum_{L_1=1}^{\infty} \cdots \sum_{L_t=1}^{\infty} E[R_{timed}(t, L_1, \ldots, L_t)] \cdot p(L_1, \ldots, L_t) \\
&\le \sum_{L_1=1}^{\infty} \cdots \sum_{L_t=1}^{\infty} E[R_{budgeted}(t, L_1, \ldots, L_t)] \cdot p(L_1, \ldots, L_t) \\
&\le E[R_{budgeted}(t)]
\end{aligned}$$

Dividing by time t to get the expected rewards per timestep $\mu_{timed}(t)$ and $\mu_{budgeted}(t)$, and then taking limits, yields the result.

$$\lim_{t\to\infty} \mu_{timed}(t) \le \lim_{t\to\infty} \mu_{budgeted}(t).$$

□

Corollary 4. *Consider a set of $k > 1$ arms drawn from a distribution where the payoff of new arms is 1 with probability $1/2k$ and $1-\delta$ otherwise, for some $\delta \in (0,1]$. Let the lifetime of an arm have geometrical distribution with $p = 1/k$, and expected lifetime $L = k$. Then, the expected regret is $\Omega(\delta t)$.*

Proof. The expected maximum reward in each step is at least

$$E[R_{max}] = (1-\delta)(1-\frac{1}{2k})^k + (1-(1-\frac{1}{2k})^k) = 1-\delta(1-\frac{1}{2k})^k \geq 1-\delta+\frac{\delta}{2}.$$

The expected reward of a random arm is

$$E[X] = \frac{1}{2k} + (1-\delta)(1-\frac{1}{2k}) = 1-\delta+\frac{\delta}{2k}.$$

Now, consider a bandit with infinitely many arms, whose first k arms are identical to the arms of the given finite-armed bandit, and the rest of the arms are drawn from an identical payoff distribution (i.e., payoff 1 with probability $1/2k$, and $1-\delta$ otherwise) and lifetime distribution. Since a bandit with only k arms is a more constrained version of the infinite-armed bandit above, the optimal expected reward per timestep for the k-armed bandit is bounded above by that for the infinite-armed bandit.

The optimal expected reward per timestep for the infinite-armed bandit can be found using Theorem 2. The only possible arm payoff values are 1 and $1-\delta$, and the corresponding $\Gamma(\mu)$ values are:

$$\begin{aligned}\Gamma(1) &= 1-\delta+\frac{\delta}{2k},\\ \Gamma(1-\delta) &= \frac{1-\delta+\frac{\delta}{2k}+\frac{k}{2k}}{1+k} \leq (1-\delta)\end{aligned}$$

Thus, $\Gamma(\mu^*) = 1-\delta+\frac{\delta}{2k} = E[R_{max}] - \frac{\delta}{2}\left(1-\frac{1}{k}\right)$. Hence, the expected regret in t steps is $\Omega(\delta t)$. □

Corollary 5. *Assume arm payoffs are drawn from a uniform distribution, $F(x) = x (0 \leq x \leq 1)$. Consider the timed death case with parameter p $(0 < p < 1)$. Then the expected regret of any algorithm per time step is $O(\sqrt{p})$.*

Proof. Let $0 \leq \mu \leq 1$. 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

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

□

Lemma 6. *Let $\bar{\mu}^{sto}(t)$ and $\bar{\mu}^{det}(t)$ denote the respective maximum mean expected rewards that any algorithm for the deterministic and stochastic mortal multi armed bandit problems could obtain after running for t steps. Then $\bar{\mu}^{sto}(t) \leq \bar{\mu}^{det}(t)$.*

Proof. Let AS be an algorithm for stochastic arms and AD be an algorithm for deterministic arms. The algorithm AD imitates algorithm AS 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. Since this can be done for any algorithm AS, it implies that $\bar{\mu}^{sto}(t) \leq \bar{\mu}^{det}(t)$. □

Lemma 7. *The asymptotic expected reward per step of Algorithm* Stochastic *is at least*

$$\Gamma(\mu^* - \epsilon)\left(1 - O\left(\frac{\log L}{\mu^* L}\right)\right)$$

for any fixed $\epsilon > 0$.

Proof. We analyze the algorithm for $r = \max[\mu^*, \frac{2\log L}{\epsilon^2}]$ for some fixed $\epsilon > 0$.

The probability that the algorithm does not identify a good arm (an arm with expected reward at least μ^*) is bounded by $1/2$ (the probability that a Binomial random variable assumes a value strictly smaller than its expectation). Applying Azuma's inequality (Corollary 12.5 in page 304 of [2]) with $c = 1$ we show that the probability that an arm with expected reward smaller than $\mu^* - \epsilon$ will have at least $r\mu^*$ total reward in r probes is bounded by (one side bound) $e^{-r\epsilon^2/2} \leq 1/L$.

Using similar argument as in the proof of Theorem 2 we show that the the asymptotic expected reward per step of Algorithm B is at least

$$\frac{2rE[X] + (1 - F(\mu^*))(1 - 1/L)(L - 2r)E[X|X > \mu^* - \epsilon]}{2r + (1 - F(\mu^*))(L - 2r)}.$$

The enumerator can be written as

$$E[X] + (1 - F(\mu^*))(L-1)E[X|X > \mu^* - \epsilon] + (2r-1)E[X] - \frac{2r}{L}(1 - F(\mu^*))(L-1)E[X|X > \mu^* - \epsilon]$$
$$\geq (E[X] + (1 - F(\mu^*))(L-1)E[X|X > \mu^* - \epsilon])(1 - \frac{2r}{L}).$$

The denominator can be written as

$$1 + (1 - F(\mu^*))(L-1) + (2r-1)F(\mu*) \leq (1 + (1 - F(\mu^*))(L-1))(1 + \frac{2r-1}{(1 - F(\mu^*))(L-1)}).$$

Thus we get

$$\begin{aligned}
& \frac{2rE[X] + (1 - F(\mu^*))(1 - 1/L)(L - 2r)E[X|X > \mu^* - \epsilon]}{2r + (1 - F(\mu^*))(L - 2r)} \\
= \; & \frac{E[X] + (1 - F(\mu^*))(L-1)E[X|X > \mu^* - \epsilon]}{1 + (1 - F(\mu^*))(L-1)}\left(1 - O(\frac{\log L}{(1 - F(\mu^*)\mu^* L})\right) \\
= \; & \Gamma(\mu^* - \epsilon)\left(1 - O(\frac{\log L}{(1 - F(\mu^*)\mu^* L})\right)
\end{aligned}$$

for any fixed $\epsilon > 0$. □

In the case of arms with rewards distribution uniform in $[0, 1]$ and lifetime distributed geometric with parameter $1/k$, the expected regret per step is still $O(1/\sqrt{k})$.

References

[1] Feller, W. (1971). *An introduction to probability theory and its applications, volume 2*. Wiley.

[2] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.