

Which Way of Exploring Is Better Is Decided Not by the Loss at One Moment but by How the Loss Grows

— per tenfold increase, UCB’s loss grows by **5.982350**, **3.151112** and **1.828572**, shrinking, while fixed-rate exploration’s growth rises to **8.360495** — the separator is whether one compares the loss at a chosen number of pulls or its growth per tenfold increase — [Paper 873]

Yuuki YAMAGISHI (山岸 勇輝)

Yamagishi Complexity Torus Lab, Japan

September 2026

YOI: <https://www.yamagishiyuuki.com/yoi/873>

— *From thinking to syncing.* —

Abstract

It is easy to assume that pulling each method the same number of times and comparing losses shows which way of exploring is better. This paper counts, with three arms, that the winner switches as the number of pulls is raised tenfold. No new theorem or law is claimed. First, at ten thousand pulls, fixed-rate exploration wins: exploring at rate 0.1 loses **122.396667**, and UCB, which adds a bonus to less-tried arms, loses **145.970000** (Section 2). Second, and this is the core. At a hundred thousand pulls they switch: the losses are **1023.296667** and **266.916667**, with fixed-rate exploration losing **3.833768** times as much as UCB (Section 2). Third, the switch is decided by how the loss grows: per tenfold increase, UCB’s loss grows by **5.982350**, **3.151112** and **1.828572**, shrinking, while fixed-rate exploration’s grows by **4.345181**, **5.090670** and **8.360495**, rising (Section 3). Fourth, stopping exploration makes the loss grow in proportion to time: pulling each arm once and then only the best average leaves 10 of 30 runs stuck on a non-best arm, grows by **9.914480** per tenfold, and loses **5004.796667** over a hundred thousand pulls (Section 3). Fifth, exploring at a fixed rate also grows in proportion to time: even after finding the best arm it keeps pulling others at rate 0.1, so its growth rises toward the 10 of proportional growth (Section 3). Sixth, the separator is whether one compares the loss at a chosen number of pulls or its growth per tenfold increase: comparing at one number of pulls decides the winner the moment that number is chosen (Section 4). **Keywords:** multi-armed bandit, UCB, epsilon-greedy, regret, exploration and exploitation.

Scope of this paper (scope note): No new theorem or law is claimed: the multi-armed bandit, exploring at a fixed rate (ϵ -greedy), adding a bonus to less-trying arms (UCB1), and that good methods lose only logarithmically, are all known. Trees are not entered: methods that use this bonus at every node of a tree, such as Monte Carlo tree search, are not entered. Real applications are not treated: ad selection and trial allocation are not the subject. No bounds are proved: theoretical upper and lower bounds on the loss are not treated; only averages over 30 runs with three arms are examined. Relation to earlier papers: Paper 387 showed that when to leave is not decided inside the patch, and in its scope note did not claim transfer to exploration and exploitation. This paper takes that seat, not by transferring the foraging model but by counting the balance between exploring and earning in a different model, the multi-armed bandit. What is added is laying out the losses of three methods tenfold from a hundred to a hundred thousand pulls, measuring that the winner switches between ten and a hundred thousand, separating methods whose tenfold growth shrinks from those whose growth rises, and placing the separator on loss at one number of pulls versus its growth.

1. Notation and the Question

Setting: three arms with win probabilities $\mu_1 = 0.6$, $\mu_2 = 0.5$ and $\mu_3 = 0.4$ are pulled one at a time. The loss up to T pulls is the sum of differences between the best arm's probability and that of the arm pulled.

$$\text{loss}(T) = \sum_{t=1}^T (\mu_1 - \mu_{a_t}), \quad \text{UCB choice : } a_t = \arg \max_i \left(\hat{\mu}_i + \sqrt{\frac{2 \ln t}{n_i}} \right)$$

$\hat{\mu}_i$ is the average so far and n_i the number of pulls. All three methods pull each arm once first; then the stopping method pulls only the best average, the fixed-rate method pulls a random arm with probability 0.1 and otherwise the best average. A hundred thousand pulls were repeated 30 times and averaged.

Question: does comparing losses after the same number of pulls show which way of exploring is better?

2. Loss by Number of Pulls

Computation 2.1: the losses of the three methods were computed tenfold from a hundred to a hundred thousand pulls.

method	100	1000	10000	100000
stop exploring	6.480000	54.736667	504.796667	5004.796667
explore at rate 0.1	5.533333	24.043333	122.396667	1023.296667
UCB	7.743333	46.323333	145.970000	266.916667

Observation 2.2

this is the core. The winner switches between ten and a hundred thousand pulls. Up to ten thousand pulls, exploring at rate 0.1 loses less than UCB: at a hundred thousand, UCB falls below, and exploring at rate 0.1 loses 3.833768 times as much. UCB was lower only at the hundred-thousand mark.

3. Growth per Tenfold

Computation 3.1: the factor by which the loss grows when the pulls are multiplied by ten was computed.

method	100 → 1000	1000 → 10000	10000 → 100000
stop exploring	8.447016	9.222276	9.914480
explore at rate 0.1	4.345181	5.090670	8.360495
UCB	5.982350	3.151112	1.828572

Observation 3.2

how the loss grows decides the switch. UCB's growth shrinks from 5.982350 to 1.828572, while exploring at rate 0.1 rises to 8.360495: UCB adds a larger bonus the fewer times an arm has been pulled, so bad arms are pulled less and less. Exploring at rate 0.1 keeps pulling bad arms at the same rate after finding the best, so its growth rises toward the 10 of proportional growth.

Observation 3.3

stopping exploration gets stuck by early luck. Pulling each arm once and then only the best average, 10 of 30 runs (0.333333) pulled a non-best arm the most: its growth per tenfold was 9.914480, and its loss over a hundred thousand pulls was 5004.796667, 18.750409 times UCB's.

4. The Separator

Observation 4.1

the separator is whether one compares the loss at a chosen number of pulls or its growth per tenfold increase. Compared at ten thousand, exploring at rate 0.1 wins; compared at a hundred thousand, UCB wins: comparing at one number of pulls decides the winner the moment that number is chosen. Comparing growth shows which will eventually win without choosing a number.

Observation 4.2

taking the seat of Paper 387.387 showed that when to leave is not decided inside the patch, and did not claim transfer to exploration and exploitation. Here, with a multi-armed bandit instead of a foraging model, we counted the balance between exploring and earning. In both questions, looking only inside the current arm or patch does not decide it.

Observation 4.3

what is not said. Tree search, real ads or trials, and theoretical bounds on the loss are not treated. What is shown is that which way of exploring is better is decided not by the number of pulls compared but by how the loss grows.

5. Census and Closing

1. loss is the sum of gaps to the best arm (Section 1): **three arms.**
2. **122.396667** against **145.970000** at ten thousand (Computation 2.1): **fixed rate** loses less.
3. **1023.296667** against **266.916667** at a hundred thousand (Computation 2.1): **the switch.**
4. ratio **3.833768** (Observation 2.2): **the core.**
5. UCB growth **1.828572** (Computation 3.1): **shrinking.**
6. fixed-rate growth **8.360495** (Computation 3.1): **rising.**
7. stopping growth **9.914480** (Computation 3.1): **proportional to time.**
8. **10 of 30 stuck** (Observation 3.3): **early luck.**

9. **stopping loses 18.750409 times** (Observation 3.3): **against UCB.**

10. **loss at one number, or growth** (Observation 4.1): **the separator; the seat of Paper 387.**

Closing: Which way of exploring is better is decided not by the loss at one moment but by how the loss grows. Pulling three arms ten thousand times, exploring at rate 0.1 loses 122.396667, less than UCB's 145.970000; at a hundred thousand the losses are 1023.296667 and 266.916667, a ratio of 3.833768, because UCB's growth per tenfold shrinks to 1.828572 while fixed-rate exploration's rises to 8.360495. Stopping exploration leaves 10 of 30 runs stuck on a non-best arm, growing 9.914480-fold per tenfold, in proportion to time: exploring at a fixed rate also piles up the exploration done after the best arm is found. Placed among the earlier papers: we took the seat of exploration and exploitation left open in Paper 387's scope note, with a multi-armed bandit rather than a foraging model. To be honest. There are three arms and 30 repetitions, and theoretical bounds on the loss are not treated.

References

1. Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*.
2. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press.
3. Yamagishi, Y. (2026). *When to Leave Is Not Decided Inside the Patch* [Paper 387]. YOI: <https://www.yamagishiyuuki.com/yoi/387>; DOI: 10.5281/zenodo.22699131

— *From thinking to syncing.* —

Yuuki YAMAGISHI (山岸 勇輝)

© 2026 Yamagishi Complexity Torus Lab, Japan · Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0)