

Exploration and Exploitation Strategies for the K-armed Bandit Problem

by Alexander L. Strehl

K-armed Bandit Problem

At each time step one of k arms may be pulled. Each pull produces some reward generated by a fixed and bounded probability distribution depending only on the arm that was pulled. The goal is to maximize the total average reward received for a finite number of time steps.

The Greedy Strategy

Compute the sample mean of an arm A by dividing the total reward received from the arm by the number of times the arm has been pulled. At each time step choose the arm with highest sample mean.

Greedy Strategy Problems

It is very possible to pull an arm and receive a reward well below the actual mean payoff of the arm. The sample mean will thus be much less than the true mean and for this reason the arm may never be chosen again even if it has the highest true mean.

The Naive Strategy

Pull each arm an equal number of times.

Naive Strategy Problems

The Naive strategy makes no distinction between different arms. Therefore it will choose the worst arm as much as the best arm. The total average reward will therefore generally be non-optimal.

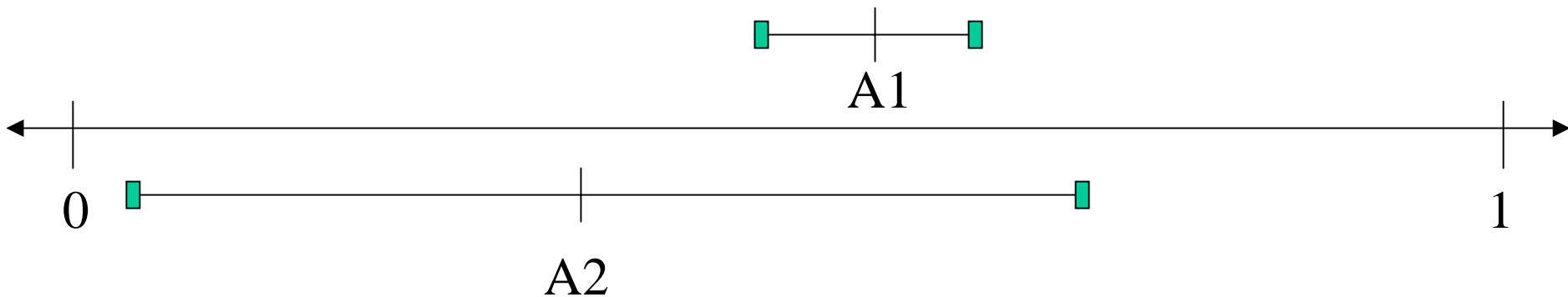
The γ -IE Strategy

The γ -IE strategy is a generalization of Kaelbling's¹ IE (Interval Estimation) Strategy. For each arm, calculate a confidence interval centered around the sample mean. Choose the arm whose confidence interval has the highest tail. Fong introduced the parameter γ , which affects the size of the confidence intervals and often improves the performance of the algorithm.

¹Kaelbling, L.P. (1993). *Learning in Embedded Systems*. Cambridge, MA: MIT Press

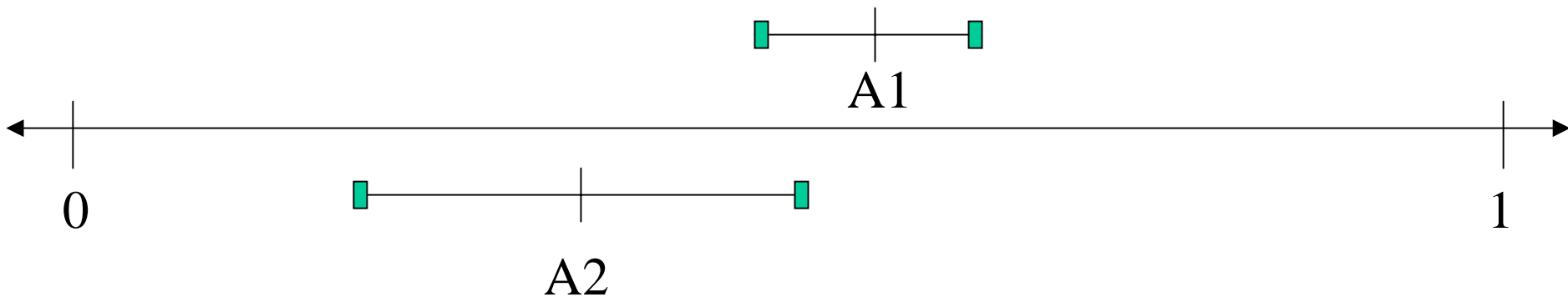
Example

Here we see that the γ -IE strategy is not the same as the greedy strategy. Arm A1 has a higher sample mean and we are very confident that the true mean is close to the sample mean. Arm A2 has a much lower sample mean but we are less confident. Using the γ -IE strategy we would choose to pull arm A2.



Example Continued

After choosing Arm A2 a number of times the confidence interval shrinks. At this point we will now choose arm A1.



Laplace (or Optimistic) Strategy

Calculate an estimate to the mean of an arm using the formula:

$est := (Z+S)/(Z+N)$, where S is the total payoff of the arm, N is the number of times the arm was pulled, and Z is some positive constant.

At each time step choose the arm with highest estimated mean. This can be thought of as assuming each arm has been pulled Z additional

A strategy very similar to this is discussed by Sutton and Barto¹.

¹Sutton, Richard S. & Barto, Andrew G. (1998). *Reinforcement Learning*. Cambridge, MA: MIT Press

Laplace Strategy continued

The parameter Z can be chosen large enough so that the Laplace satisfies the PAO conditions (introduced by Greiner and Orponen). Define an ε -optimal arm to be an arm whose true mean is greater than $(\mu - \varepsilon)$, where μ is the mean of the optimal arm. For any ε and δ there exists a number Z such that the Laplace strategy is guaranteed to find an ε optimal arm with probability $(1 - \delta)$. Fong¹ has shown that the γ -IE and Naive strategies also satisfy the PAO conditions.

¹Fong, Philip W.L. (1995). A Quantitative Study of Hypothesis Selection. *Twelfth International Conference on Machine Learning*, 226-234, Tahoe City, California.

Experiments

Four experiments were carried out. Each experiment consists of two phases: Exploration (5000 trials) and Exploitation (10000 trials). During the exploration phase any arm may be chosen, during the exploitation phase the greedy algorithm is used. The exploitation phase is included to penalize methods that pick sub-optimal arms.

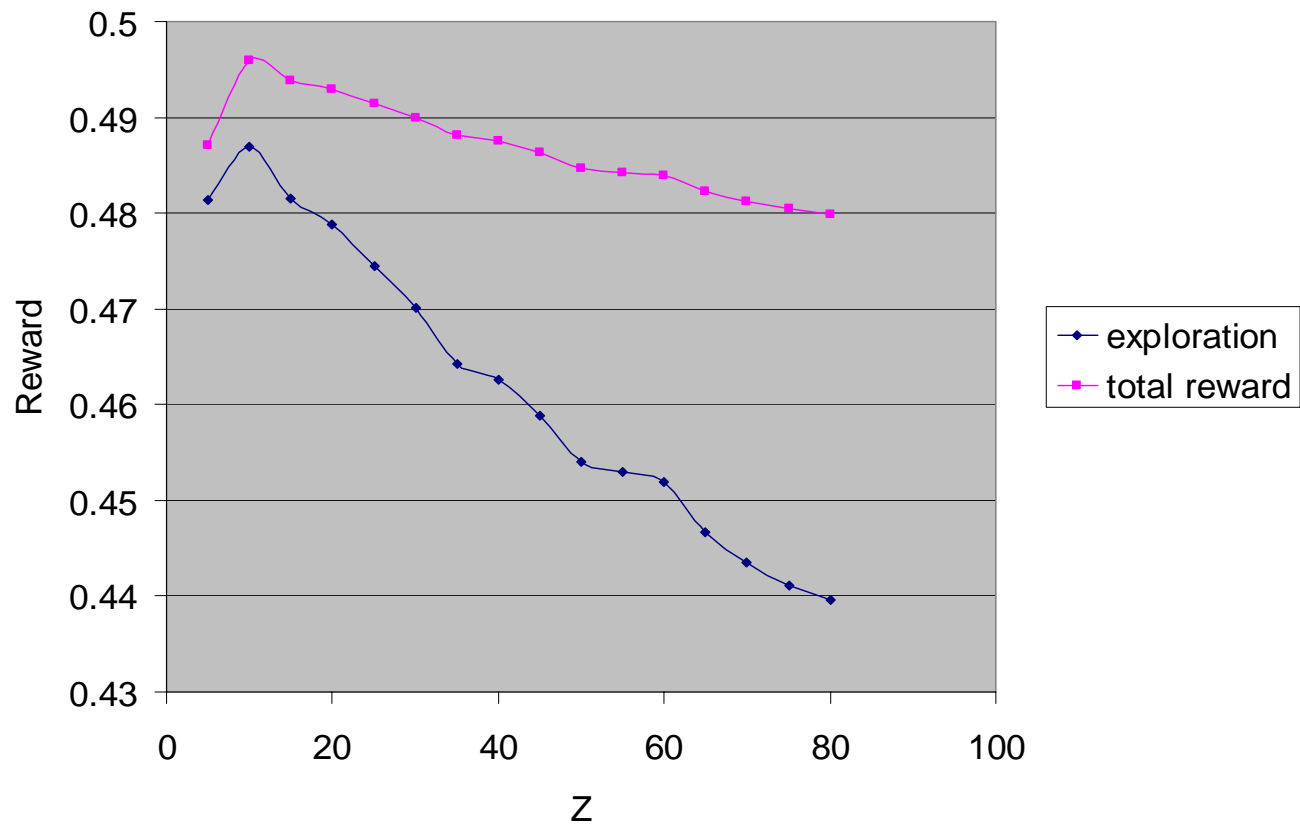
Each experiment was repeated 100 times and results averaged.

Experimental Results

Sampling Method	Exp. 1	Exp. 2	Exp. 3	Exp. 4
Naive	0.47	0.7692	0.4814	0.7665
IE	0.483	0.927	0.4901	0.798
Laplace	0.496	0.9883	0.4931	0.7996
Optimal Mean	0.5	1	0.5	0.8

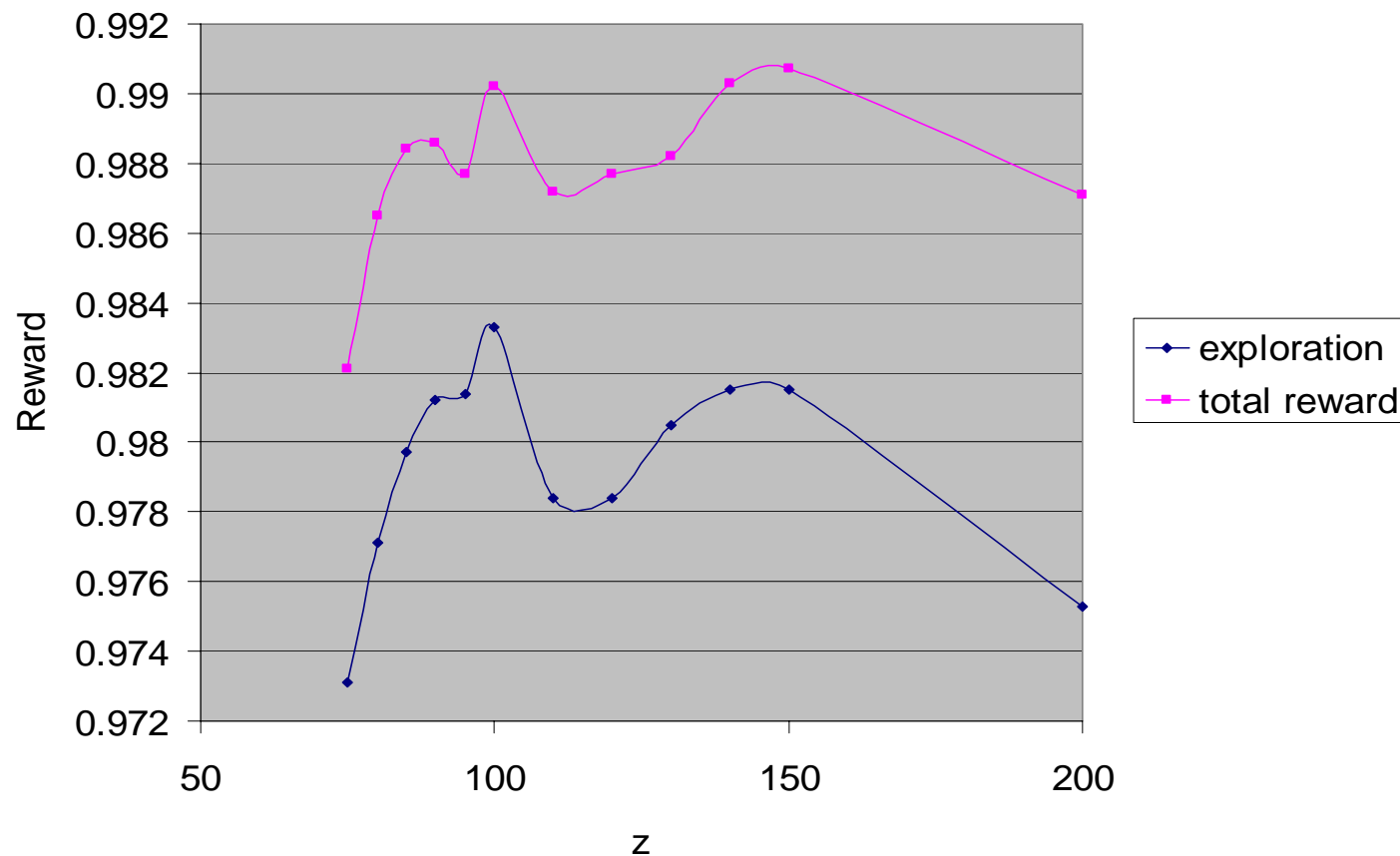
Tuning the Parameter Z

Experiment 1



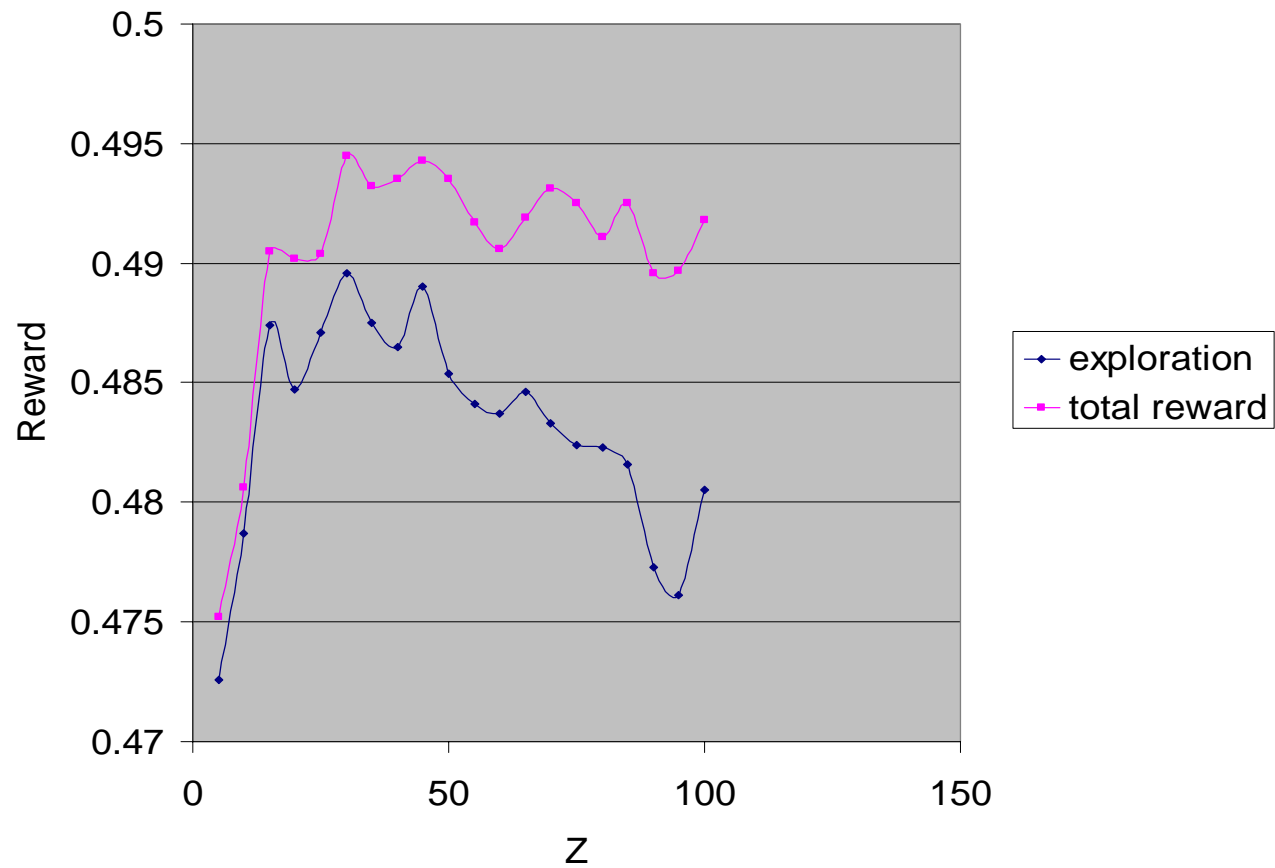
Tuning the Parameter Z cont.

Experiment 2



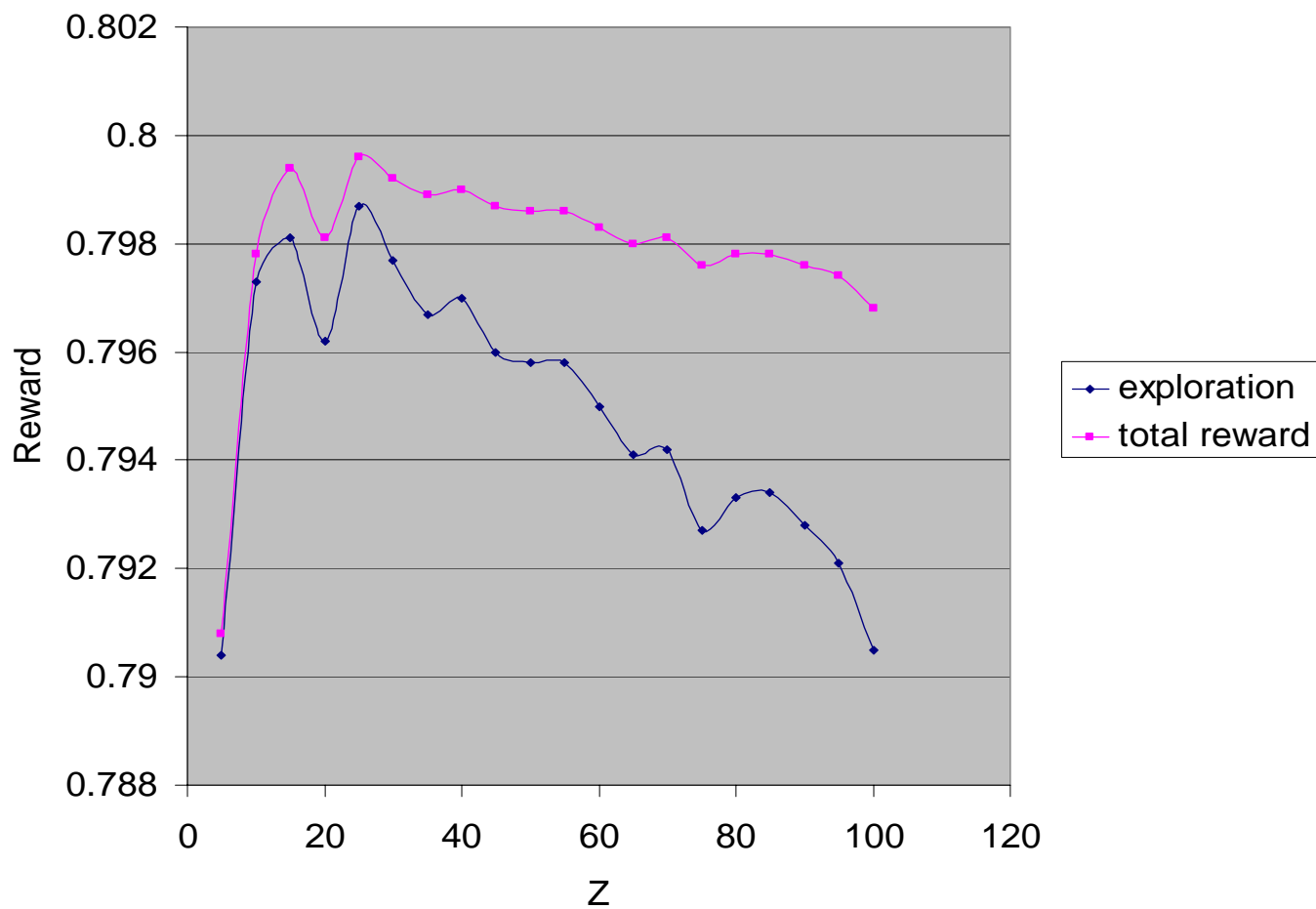
Tuning the Parameter Z cont.

Experiment 3



Tuning the Parameter Z cont.

Experiment 4



Conclusion

The Laplace Strategy is a simple alternative to the γ -IE Strategy. It has obtained slightly better empirical results. It is simpler in the sense that no confidence interval must be computed.

Future Work

- Average Case Analysis of these methods
- Calculation of optimal values of the parameters Z and γ based on the range and variance of the payoffs as well as the number of trials.