
Exploration and Exploitation Strategies for the k-armed Bandit Problem

Alex Strehl

Rutgers University

STREHL@CS.RUTGERS.EDU

Abstract

In this paper, we study several different methods that can be used in the k -armed bandit problem. Each method is considered under the PAO (probably approximately optimal) framework. The various approaches are also compared empirically. One new approach, based on a Laplace estimator method, is introduced and shown to have good performance.

1. Statement of Problem and Background

The k -armed bandit problem (also referred to as the hypothesis selection problem) has a rich history and has been studied extensively by both statisticians and computer scientists. In its simplest form, the problem is basically to pull (sample) one of k arms (hypotheses) at each time step so as to maximize the total average reward received over a finite number of time steps. Each arm, when pulled, produces some payoff or reward that is generated by a probability distribution. In this paper we assume the means of the rewards are between 0 and 1 and the range of possible payoffs are bounded from above and below.

1.1. Known Work and Results

Fong (1995) studied the k -armed bandit problem in the probably approximately optimal (PAO) framework (Geiner & Orponen, 1991). In that work, two sampling methods were considered: “Naive” and γ -IE. The Naive sampling method simply chooses each hypothesis an equal number of times (one after another). The γ -IE sampling strategy is a simple generalization of the IE (Interval Estimation) scheme introduced by Kaelbling (1993). At any point during the sampling procedure, it is possible to calculate a confidence interval ϵ_i for each hypothesis H_i such that we are relatively certain that the true mean of the hypothesis μ_i is within ϵ_i of the sample mean $\hat{\mu}_i$ (and vice versa).

This confidence interval is calculated using Hoeffding bounds. The γ -IE method chooses the hypothesis that maximizes $\hat{\mu}_i + \gamma\epsilon_i$. When $\gamma = 1$, the γ -IE method coincides with the original IE strategy. The main result of Fong’s work was to show that both methods are guaranteed to find a hypothesis with probability $(1 - \delta)$ that has mean within ϵ of the mean of the optimal hypothesis (the hypothesis with maximum mean). By definition, these are the PAO conditions (Geiner & Orponen, 1991). Fong (1995) proved that the number of samples needed for both methods, the sampling complexity, is $\Theta(\sum_{i=1}^k m(i))$, where $m(i) = 4 \frac{\Delta_i^2}{\epsilon^2} \max\{\ln \frac{k\pi^2}{3\delta}, 4 \ln \frac{8\delta_i^2}{\epsilon^2}\}$. In the previous formula, Δ_i is the range of possible payoffs for H_i (the difference between the maximum possible payoff and the minimum possible payoff). In particular, Fong showed that the number of samples needed to find this hypothesis is polynomial in the quantities $\frac{1}{\epsilon}$, $\ln \frac{1}{\delta}$, k , and $\sum_{i=1}^k \Delta_i^2$.

2. The Laplace Sampling Method

The Laplace method of sampling hypotheses chooses the hypothesis with greatest estimated mean $est[\mu_i] := \frac{z+s}{z+n}$, where s is the total reward received from sampling hypothesis H_i and n is the total number of times H_i has been sampled. In the estimation, z is chosen to be some number greater than 0. The idea behind this method is to initially assume that each hypothesis has been sampled z times and each time has produced maximum payoff of one unit. Intuitively, z can be considered a smoothing parameter that causes the estimates of hypothesis means to be optimistic, thus causing each hypothesis to be sampled enough times to make confident comparisons. A method very similar to the Laplace method was studied empirically by Sutton and Barto (1998).

2.1. Basic Theoretical Properties

Throughout this section, we make the additional assumption that all rewards received, when sampling any

hypothesis H , are strictly between zero and one. This case can be generalized to the case where negative rewards are allowed by letting $z_n = z + C(n)$, where $C(n)$ is a positive constant chosen so that after n trials $C(n) + s \geq 0$, if s is any possible total reward received from sampling the hypothesis n times. The estimated mean is then calculated using z_n rather than z . However, this may result in estimated mean rewards that are greater than one, so would require a simple revision of the Laplace method.

Proposition 1

Let n be the number of times a hypothesis H has been sampled and let s be the total reward accumulated over the n samples. If $n \geq \frac{z}{\epsilon} - z = \frac{z(1-\epsilon)}{\epsilon}$, then $est[\mu] := \frac{z+s}{z+n} \leq \hat{\mu} + \epsilon$, for any $\epsilon \in (0, 1)$.

Proof. $n \geq \frac{z}{\epsilon} - z \Rightarrow \frac{z}{z+n} \leq \epsilon \Rightarrow est[\mu] = \frac{z}{z+n} + \frac{s}{z+n} \leq \epsilon + \frac{s}{z+n} \leq \epsilon + \frac{s}{n}$ \square

Proposition 1 reveals that after a certain number of samples for hypothesis H , no matter how large z happens to be, we can be sure that the optimistic estimation $est[\mu]$ is no more than ϵ above the sample mean. In fact, we have that $\hat{\mu} \leq est[\mu] \leq \hat{\mu} + \epsilon$ if $n \geq \frac{z(1-\epsilon)}{\epsilon}$, since $\hat{\mu} \leq est[\mu]$ holds for all values of $n \geq 1$. To see this, note that $est[\mu] = \hat{\mu} + \frac{zn-zs}{z+n}$. However, the second term on the right hand side of this equation is non-negative, since z is positive and $s \leq n$.

Corollary 2

Let n be the number of times a hypothesis H has been sampled and let s be the total reward accumulated over the n samples. Then, as $n \rightarrow \infty$, the estimated mean $est[\mu] \rightarrow \hat{\mu} = \frac{s}{n}$.

Proof. Choose any $\epsilon \in (0, 1)$. Using proposition 1, we can find a value N such that if $n > N$ (take $N = \frac{z(1-\epsilon')}{\epsilon'}$ where ϵ' is between zero and ϵ), the difference between $est[\mu]$ and $\hat{\mu}$ is less than ϵ . \square

Basically, Corollary 2 indicates that the estimated mean $est[\mu]$ approaches the value of the sample mean $\hat{\mu}$ as the number of times we sample hypothesis H increases to infinity. An immediate consequence of this is that $est[\mu] \rightarrow \mu$ as $n \rightarrow \infty$, since the sample mean approaches the true mean in the limit as $n \rightarrow \infty$.

Since the payoffs for the different hypotheses are drawn from a probability distribution, they may have arbitrarily high variances. Hence, a practical sampling strategy must try each hypothesis some number of times just to get a good estimate of its mean before using this estimate to compare one hypothesis to another. Thus, we are interested in knowing if there

exists a value of z that will force our estimate of the mean of a hypothesis to be arbitrarily large until H has been sampled a certain number of times (regardless of the payoffs received from sampling H). The following proposition answers this question in the affirmative.

Proposition 3

Let n be the number of times a hypothesis H has been sampled and let s be the total reward accumulated over the n samples. For any positive integer n^* and for any $\epsilon \in (0, 1)$, if $z \geq \frac{n^*(1-\epsilon)}{\epsilon}$ then $est[\mu] \geq 1 - \epsilon$ whenever $n \leq n^*$.

Proof. $z \geq \frac{n^*(1-\epsilon)}{\epsilon} \Rightarrow z \geq \frac{n(1-\epsilon)}{\epsilon} \Rightarrow \epsilon z \geq n(1-\epsilon) \Rightarrow z + \epsilon z \geq z + n(1-\epsilon) \Rightarrow z \geq (1-\epsilon)(z+n) \Rightarrow est[\mu] \geq \frac{z}{z+n} \geq 1 - \epsilon$ \square

Fong(1995) proves that if each hypothesis H_i is sampled $m(i)$ times, where

$$m(i) = 4 \frac{\Delta_i^2}{\epsilon^2} \max\{\ln \frac{k\pi^2}{3\delta}, 4 \ln \frac{8\delta_i^2}{\epsilon^2}\}$$

then the hypothesis with maximum sample mean is with probability $(1 - \delta)$ greater than or equal to the maximum true mean of any of the k hypotheses minus ϵ . Using Proposition 3, we can guarantee that each hypothesis H_i will have estimated mean greater than or equal to $1 - \epsilon$ until it is chosen $m(i)$ times to be sampled. Thus, it would seem, by Proposition 1, that after some number of trials either we have sampled each hypothesis enough ($m(i)$) times or we have sampled one hypothesis enough times to guarantee that its estimated mean is very close to its sample mean, at which point it is either within ϵ of the optimal value (with probability $(1 - \delta)$) or it will never be sampled again until every other hypothesis has been sampled ($m(i)$) times. This is the main argument of the following theorem.

Theorem 4

Let $n^* := \max\{m(i)|i = 1, \dots, k\}$. If $z = \frac{n^*(1-\epsilon/2)}{\epsilon/2}$, then after $\frac{kn^*(1-\epsilon/2)^2}{(\epsilon/2)^2} + kn^*$ trials using the Laplace estimation strategy, the hypothesis with maximum sample mean will have mean within ϵ of the maximum mean with probability $(1 - \delta)$.

Proof. Let H_i be the first hypothesis to be sampled $\max\{\frac{z(1-\epsilon/2)}{\epsilon/2} = \frac{n^*(1-\epsilon/2)^2}{(\epsilon/2)^2}, n^*\}$ times. By Proposition 1, $est[\mu_i]$ will be within $\epsilon/2$ of the sample mean $\hat{\mu}_i$ and thus will be within $\epsilon/2$ of the true mean μ_i with probability $(1 - \delta)$. By Proposition 3, H_i will never be sampled again until every other hypothesis H_j is sampled $n^* \geq m(j)$ times, unless $\hat{\mu}_i$ is within $\epsilon/2$ of 1, in which case the true mean of H_j is necessarily within

ϵ of the maximum mean of any of the hypotheses with probability $1 - \delta$. Furthermore, no other hypothesis H_j will be sampled more than $\max\{\frac{n^*(1-\epsilon/2)^2}{(\epsilon/2)^2}, n^*\}$ times for the same reason (unless of course it happens to be within $\epsilon/2$ of 1, or all other hypotheses H_l have been sampled $m(l)$ times). \square

The upper bound proved in Theorem 4 is certainly not as small as $\sum_{i=1}^n m(i)$, which Fong (1995) proved to be an upper (and worst-case lower) bound for the number of trials needed for both the Naive and γ -IE strategies to provide PAO (Probably Approximately Optimal) hypotheses (as in Theorem 4). However, it is still polynomial in the quantities $\frac{1}{\epsilon}$, $\ln\frac{1}{\delta}$, k , and $\sum_{i=1}^k \Delta_i^2$, and we will see empirically that the Laplace estimator strategy generally performs very well compared with these two other methods.

2.2. Comparison with γ -IE

The Laplace and γ -IE sampling strategies have some features in common. Both have parameters that must be tuned specifically for each different setting of the hypothesis selection problem. In particular, the Laplace method requires a parameter $z \geq 1$ and the γ -IE method requires a parameter $\gamma \geq 1$. Both z and γ are typically chosen to be positive integers and both parameters control the trade-off between exploration and exploitation. The smaller the values of z and γ the more the respective strategies tend to concentrate on exploitation. In fact, if allowed, both methods approximate the greedy strategy arbitrarily well as $z \rightarrow 0$ for the Laplace method and as $\gamma \rightarrow 0$ for the γ -IE strategy. The greedy strategy is defined to be the strategy that chooses the hypothesis with highest sample mean where ties are broken arbitrarily. As $\gamma \rightarrow \infty$, the γ -IE strategy approaches the behavior of the Naive strategy. This is also basically true for the Laplace method as $z \rightarrow \infty$, except that if some hypothesis continues to produce rewards of one unit or greater, than that hypothesis will continue to be chosen. One seemingly practical advantage of the Laplace method is its simplicity. It only requires a simple calculation similar to the calculation of the sample mean, while the γ -IE method involves the calculation of a confidence interval. Thus, the Laplace method is simpler to program and typically runs much faster than the γ -IE method on the same settings of the problem.

3. Experimental Results

We have conducted four experiments comparing the three hypothesis selection methods: Naive, IE, and Laplace. The results indicate that the Laplace esti-

mator method will generally outperform the other two methods if z is set optimally. The results also give a hint as to the relationship between different values of z and the performance of the Laplace estimator strategy under this choice of z . Each experiment consisted of two phases: exploration and exploitation. During the exploration phase, the estimator could choose among any of the k hypotheses. After the exploration phase (5000 trials), the hypothesis with greatest sample mean is chosen for the duration of the exploitation phase (generally 10000 trials except in Experiment 2 where it is 5000). The average reward accumulated over the exploration phase as well as the total average reward was recorded. Each experiment was repeated 100 times and the results averaged. Among these 100 trials, the number of times a non-optimal hypothesis was chosen for the exploitation phase was also recorded. The first three experiments were modeled after the first three experiments described in Fong's (1995) paper.

Experiment 1 consisted of $k = 10$ hypotheses. The best hypothesis was H_1 , which had average payoff of 0.5, while all other hypotheses had average payoff of 0.4. The standard deviation (σ_i) for each hypothesis was 0.5.

Experiment 2 consisted of $k = 10$ hypotheses, where $\mu_i = \frac{i}{10}$ and $\sigma_i = 1$ for $i = 1, \dots, 10$. In this experiment the best hypothesis was H_{10} and the hypotheses get better as you move from H_1 to H_{10} . A good sampling strategy will generally try to concentrate on the higher paying hypotheses (H_9 and H_{10}) rather than the others. This not only ensures a higher average exploration reward but also gives the method a better chance of determining that H_{10} is actually better than H_9 which, due to the high variance, may be hard to do in only 5000 trials.

Experiment 3 consisted of $k = 3$ hypotheses, where $\mu_1 = 0.5$, $\mu_2 = 0.45$, $\mu_3 = 0.4$, $\sigma_1 = 1$, $\sigma_2 = .95$, and $\sigma_3 = .9$. In this experiment all three means are close together and all have very large differing variances.

Experiment 4 restricted the payoffs to zero and one, with $k = 3$, where $\mu_1 = 0.8$, $\mu_2 = 0.7$, and $\mu_3 = 0.6$. In this case, H_1 was the best hypothesis, paying a reward of 1 about eighty percent of the time.

In Experiments 1, 2, and 4, the Laplace method performed better than the other two methods, as can be seen in figure 1. In Experiment 3, the Laplace method performed on par with the γ -IE method. In general, the results suggest that the Laplace estimator strategy

Table 1. Experimental results comparing total average reward of different sampling methods.

SAMPLING METHOD	EXP. 1	EXP. 2	EXP. 3	EXP. 4
NAIVE	0.47	0.7692	0.4814	0.7692
γ -IE	0.483	0.927	0.4901	0.798
LAPLACE	0.496	0.9907	0.4931	0.7996

works at least as well as the γ -IE strategy.

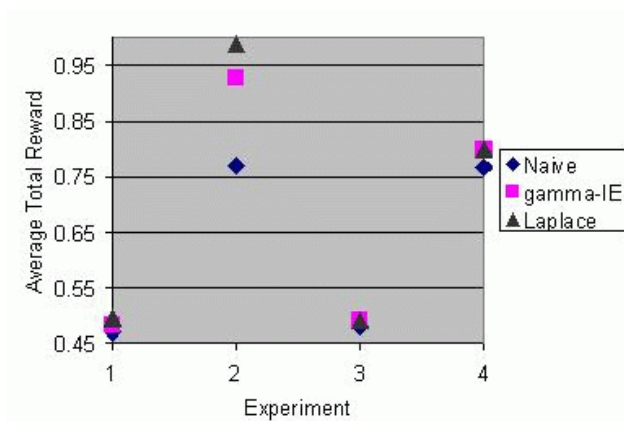


Figure 1. A Plot of the results for the four experiments. The points representing the results of the γ -IE and Laplace methods use the values of γ and z , respectively, that proved to provide the best results.

Although, for certain values of z , the Laplace estimator strategy was able to outperform both the γ -IE strategy and the Naive strategy, choosing a value for z is very important for the performance of the Laplace method. As can be seen in Figure 2, if z is too small or too large, the Laplace method doesn't perform optimally. Intuitively, it seems that if z is too small the method is not exploring enough and if z is too large it is waiting too long to exploit.

4. Future Work

It appears empirically that the Laplace estimator sampling strategy outperforms the other two strategies studied. However, it is much harder to prove theoretical bounds on the Laplace method that explains the empirical evidence. We are currently working on a more in-depth analysis, including average-case analysis, of the Laplace, γ -IE, and Naive methods in hope

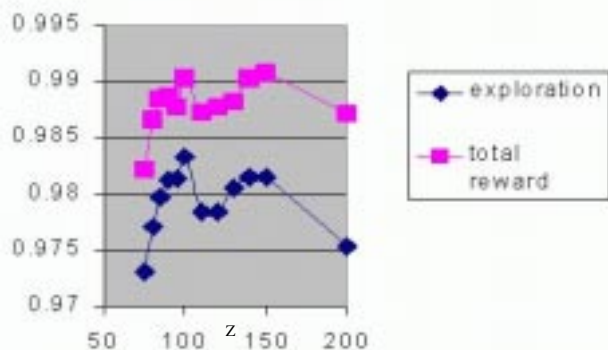


Figure 2. A Plot of the results for the Laplace method in Experiment 2. The lower line is the average exploration reward, while the upper line is the average total reward.

Table 2. A table of results for the Laplace method in Experiment 2. \mathcal{F} is the number of times (out of 100) that the method chose a non-optimal hypothesis.

z	AVG. EXPL. REW.	AVG. TOTAL REW.	\mathcal{F}
80	0.9771	0.9865	4
85	0.9797	0.9884	3
90	0.9812	0.9886	4
95	0.9814	0.9877	6
100	0.9833	0.9902	3
120	0.9784	0.9877	4
140	0.9815	0.9903	1
150	0.9815	0.9907	0
200	0.9753	0.9871	0
300	0.9767	0.9883	0
400	0.9705	0.9853	0
500	0.9671	0.9835	0

of such an explanation.

References

- Fong, Philip W. L. (1995). A Quantitative Study of Hypothesis Selection. *Twelfth International Conference on Machine Learning*, 226-234, Tahoe City, California.
- Greiner, Russell & Orponen, Pekka (1991). Probably Approximately Optimal Satisficing Strategies. *Artificial Intelligence*, 82, 21-44.
- Kaelbling, L. P. (1993). *Learning in Embedded Systems*. Cambridge, MA: MIT Press.
- Sutton, Richard S. & Barto, Andrew G. (1998). *Reinforcement Learning*. Cambridge, MA: MIT Press