CCN Interest Forwarding Strategy as Multi-Armed Bandit Model with Delays

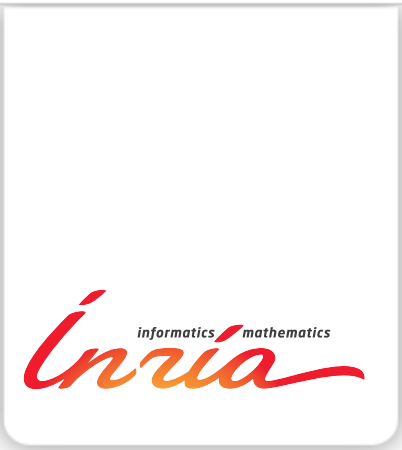
Konstantin Avrachenkov, Peter Jacko

CCN Interest Forwarding Strategy as Multi-Armed Bandit Model with Delays

Konstantin Avrachenkov, Peter Jacko

RESEARCH REPORT
N° 7917
March 2012
Project-Team Maestro

ISSN 0249-6399 ISRN INRIA/RR--7917--FR+ENG



CCN Interest Forwarding Strategy as Multi-Armed Bandit Model with Delays

Konstantin Avrachenkov*, Peter Jacko†

Project-Team Maestro

Research Report n° 7917 — March 2012 — 18 pages

Abstract: We consider Content Centric Network (CCN) interest forwarding problem as a Multi-Armed Bandit (MAB) problem with delays. We investigate the transient behaviour of the ε-greedy, tuned ε-greedy and Upper Confidence Bound (UCB) interest forwarding policies. Surprisingly, for all the three policies very short initial exploratory phase is needed. We demonstrate that the tuned ε-greedy algorithm is nearly as good as the UCB algorithm, the best currently available algorithm. We prove the uniform logarithmic bound for the tuned ε-greedy algorithm. In addition to its immediate application to CCN interest forwarding, the new theoretical results for MAB problem with delays represent significant theoretical advances in machine learning discipline.

Key-words: Information Centric Networks, Content Centric Networks, Interest Forwarding, Multi-Armed Model with Delays

* INRIA Sophia Antipolis, France, K.Avrachenkov@sophia.inria.fr
† BCAM – Basque Center for Applied Mathematics, Spain, jacko@bcamath.org

RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Routage des Intérêts dans CCN comme le Problème de Bandit-Manchot avec des Retards

Résumé : Nous considérons le routage des intérêts dans CCN (Content Centric Networking) comme le problème de bandit-manchot avec des retards. Nous etudions le comportement transitoire des politiques : ε-greedy, tuned ε-greedy et Upper Confidence Bound (UCB). Étonnamment, pour tous les trois politiques on a besoin d'un très court première phase exploratoire. Nous démontrons que l'algorithme tuned ε-greedy est presque aussi bon que l'algorithme UCB, le meilleur algorithme actuellement disponible. Nous établissons la limite uniforme logarithmique pour l'algorithme tuned ε-greedy. En outre de son application immédiate au routage des intérêts dans CCN, les nouveaux résultats théoriques pour le problème de bandit-manchot avec des retards représentent des avancées importantes dans la discipline l'apprentissage automatique.

Mots-clés : Information Centric Networks, Content Centric Networks, Routage des Intérêts, Problème du Bandit-Manchot avec des Retards

1 Introduction

There is a conceptual clash between expanding digital information dissemination and the host-based network architecture of the current Internet. To facilitate the dissemination of digital information, several Information-Centric Network (ICN) architectures have been proposed: TRIAD [6], DONA [10], CCN/NDN [8]. Since the CCN/NDN (Content-Centric Networking / Named Data Networking) proposal appears to be the most elaborate, we develop our contribution in the framework and within the terminology of CCN/NDN. For the sake of brevity, we shall refer to CCN/NDN as CCN. The main features of the ICN paradigm, and specifically CCN architecture, are that the content is addressed by a unique name and can have many identical cached copies. Any of such copies can be retrieved independently of its location. The content is typically divided into several small chunks. A chunk is also uniquely identified. A chunk of content is located and requested by forwarding of so-called interests. A user or a CCN router can forward interests to one or more neighbour CCN routers. Clearly, if there is no bandwidth limitation the most efficient way is to forward interests to all available neighbour routers. However, if there is a bandwidth limitation or the interest sender has to pay for the interest or/and delivered content, there can be better interest forwarding strategies than simple flooding.

In the present work we suggest to view the problem of optimal interest forwarding strategy as a Multi-Armed Bandit (MAB) problem. The MAB problem is a classical problem in machine learning discipline in which a decision maker finds an optimal balance between exploration and exploitation efforts. Here we adopt three well known algorithms from MAB literature: ε-greedy [12], tuned ε-greedy and UCB [1]. Our study brings advances to both networking and machine learning disciplines. We show that the MAB algorithms allow to detect the optimal router with very small number of interests sent to sub-optimal routers. The novelty from machine learning perspective is that we analyze the transient period of the MAB algorithms with delays. This is a very challenging topic with hardly any results available in the literature. In fact, we can only cite the work [4] on MAB with delay. However, the model in [4] is different from ours and there are many restrictive assumptions.

We expect that our MAB-based mechanisms can be integrated in the Interest Control Protocol (ICP) which regulates the pacing of the interests [3].

2 Model and interest forwarding strategies

We suppose that a CCN router or a user can forward interests to n CCN neighbour routers. We consider a discrete time model. The slot duration can be chosen equal to the minimal duration of packet generation at the MAC layer. Initially we assume that at each time slot $t \in \mathcal{T} := \{0, 1, 2, \dots\}$ the user can send only one interest to one of n CCN neighbour routers.

CCN routers reply with delays distributed according to discrete distribution functions $F_k(x)$, $k = 1, ..., n$, $x = 1, 2, ...$ with mean denoted by μ_k. Specifically, we assume that a chunk corresponding to the interest generated at the present slot and forwarded to the neighbour router k is delivered by router k after a random number of slots distributed according to the distribution function $F_k(x)$. Thus, we shall know the effect of the action taken at the time slot t only at the future time slot $t + X_k(t)$, where $X_k(t)$ is an i.i.d. random variable generated according to $F_k(x)$.

We are interested in minimizing the expected number of interests sent to sub-optimal routers, or to sub-optimal arms in terminology of the multi-armed bandit framework [12]. The challenging novelty of our setting with respect to the classical multi-armed bandit problem formulation is that the cost becomes known to the decision maker with delays. In fact, the costs are the delays.

The optimal policy in the classical setting without delay is obtained by the Gittins index rule [5], which breaks the combinatorial complexity of the problem by computing the Gittins index

(a history-dependent function) for each router in isolation and then simply sending the interest at every slot to the router whose current Gittins index value is lowest. This result significantly reduces the dimensionality of the problem, but the evaluation of the Gittins index may still be computationally tedious, especially if the index depends on the whole history, not only on the last observed state. Moreover, the Gittins optimality result requires that the evolution of costs from routers be mutually independent, while the algorithms described below are efficient even for dependent arms [1].

Since strictly speaking optimal policy is very likely to be very complex even in the classical setting without delay, many researchers have proposed sensible policies and shown desirable properties of such policies [9, 1]. One desirable property of the multi-armed bandit problem policy is the uniform logarithmic bound on the number of sub-optimal arms chosen by the decision maker. We shall establish the uniform logarithmic bound for the tuned ε-greedy policy in the case of delayed information in Section 4.

In the present work we consider the following three algorithms: ε-greedy algorithm, tuned ε-greedy algorithm, and UCB (Upper Confidence Bound) algorithm. These are the most used multi-armed bandit algorithms, and in this paper we propose their generalizations to the setting with delayed information.

Let us formally describe each algorithm. The ε-greedy algorithm is the simplest algorithm. Its main drawback is that the expected number of sub-optimal arms grows linearly in time. A variant of ε-greedy algorithm was proposed in [12] for Markov Decision Process models without delay.

Denote by $T_k(t)$ the total number of interests sent to router k and answered up to the end of slot $t-1$.

Algorithm ε-greedy

1. (Initialization) Choose $t_0 \in \mathcal{T}$ and $\varepsilon \in (0,1)$. During the first t_0 slots keep sending interests to routers in round robin fashion or randomly to routers chosen according to the uniform distribution.

2. **at each time slot** $t \geq t_0$ **do**

3. For each router k, compute the average delay:

$$\bar{X}_k(t) =$$

$$\frac{1}{T_k(t)} \sum_{\tau=1}^{t} 1\{\text{interest sent to } k \text{ at } \tau \text{ and answered before } t\} X_k(\tau)$$

4. For each router k, set the index:

$$\nu_k(t) = \bar{X}_k(t).$$

5. with probability $1-\varepsilon$ send new interest to the router with the smallest index or with probability ε send new interest to a uniformly randomly chosen router.

6. **end for**

The tuned ε-greedy algorithm and UCB algorithm for models without delays have been proposed and analysed in [1]. Both the tuned ε-greedy and UCB algorithms have logarithmic bounds on the number of sub-optimal arms in the case of no delays [1].

Algorithm tuned ε-greedy

1. (Initialization) Choose $t_0 \in \mathcal{T}$ and $\varepsilon_0 \in (0, t_0)$. During the first t_0 slots keep sending interests to routers in round robin fashion or randomly to routers chosen according to the uniform distribution.
2. **at each time slot** $t \geq t_0$ **do**
3. For each router k, compute the average delay:

$$\bar{X}_k(t) =$$

$$\frac{1}{T_k(t)} \sum_{\tau=1}^{t} 1\{\text{interest sent to } k \text{ at } \tau \text{ and answered before } t\} X_k(\tau)$$

4. For each router k, set the index:

$$\nu_k(t) = \bar{X}_k(t).$$

5. with probability $1 - \varepsilon_0/t$ send new interest to the router with the smallest index and with probability ε_0/t send new interest to a uniformly randomly chosen router.
6. **end for**

Algorithm Upper Confidence Bound (UCB)

1. (Initialization) Choose $t_0 \in \mathcal{T}$ and $L > 0$. During the first t_0 slots keep sending interests to routers in round robin fashion or randomly to routers chosen according to the uniform distribution.
2. **at each time slot** $t \geq t_0$ **do**
3. For each router k, compute the average delay:

$$\bar{X}_k(t) =$$

$$\frac{1}{T_k(t)} \sum_{\tau=1}^{t} 1\{\text{interest sent to } k \text{ at } \tau \text{ and answered before } t\} X_k(\tau)$$

4. For each router k, set the index:

$$\nu_k(t) = \bar{X}_k(t) - \sqrt{\frac{L \ln(t)}{T_k(t)}}$$

   where L is so-called exploration parameter.
5. send new interest to the CCN router with the smallest index.
6. **end for**

In our case, since we minimize the cost, we should more appropriately call this algorithm the lower confidence bound algorithm. However, to make an explicit connection with the work [1] we shall continue to call it the UCB algorithm. In the previous works the UCB algorithm have shown slightly better performance than the tuned ε-greedy algorithm.

Parameters	Router 1	Router 2	Router 3
propagation delay	2	2	2
p parameter	0.8	0.7	0.6
r parameter	10	10	10
mean delay	4.5	6.29	8.67
std	1.77	2.47	3.33

Table 1: The values of parameters in the numerical example.

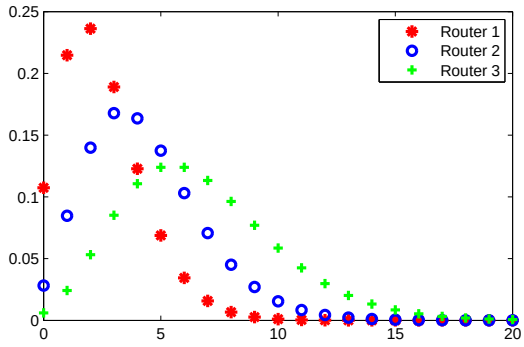


Figure 1: Negative binomial distributions in example.

To get an idea of the performance of the above algorithms in the presence of delay, we provide a numerical example. In our numerical examples as the distribution of delay $F_k(x)$, we have taken the negative binomial distribution with deterministic shift. There are several reasons for this choice. The negative binomial distribution is quite versatile. With two parameters, we can easily choose any mean and variance, which have simple explicit expressions. The distribution shape can take diverse forms such as the shape of geometric distribution and the shape close to the normal distribution shape. The negative binomial distribution represents the distribution of a sum of geometrically distributed random variables. Since the waiting time distribution in many queueing systems is exponential or close to exponential, the negative binomial distribution represents well the response time of queueing systems in cascade. We introduce the deterministic shift to model the propagation delay. In Table 1 we present the parameters of our numerical example and in Figure 1 we plot the negative binomial distributions with the chosen parameters.

In Figure 2 we plot the fraction of interests sent to the optimal arm as a function of time for the three algorithms with Round Robin strategy employed in the initial phase. This numerical example demonstrates that despite the presence of delays, the three algorithms perform well. In particular, as in the case of no delay, the performances of the UCB and tuned ε-greedy algorithms are comparable and the ε-greedy algorithm performs not too badly. In the following sections we will focus on the analysis of these three algorithms from a mathematical point of view.

3 Analysis of initial exploratory phase

Let us now investigate the effect of the duration of the initial, purely exploratory, phase on the algorithm performance. We shall consider two possible initial strategies: the Round Robin (RR) strategy and the strategy when the arm chosen randomly with uniform probability (Uni). Note

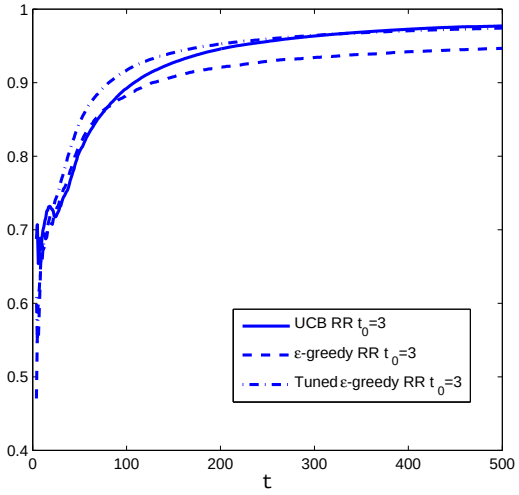


Figure 2: Comparison of MAB algorithms.

that in the Round Robin strategy the initial arm and the order are chosen randomly with uniform distribution.

In Figures 3-5 we plot the fraction of interests sent to the optimal arm for different durations ($t_0 = 3, 9, 30$) of the initial phase for different algorithms with different initial phase strategies.

A bit surprisingly, it turns out that it is better to set up very short duration of the initial phase. Another important observation that it is better to use the Round Robin initial strategy rather than the uniformly random strategy. This is intuitively expected as by using the Round Robin strategy we reduce the randomness. Below we provide theoretical explanation of these phenomena.

The initial phase $[0, t_0 - 1]$ is characterized by large exploration effort. Here we would like to provide an estimate for the period after which we can with high certainty rely on the choice of the best performing arm based on evaluated averages. Specifically, let us estimate the probability of choosing the best arm (denoted by $*$) given the arms are chosen independently before the end of the initialization phase.

Denote by I_t the arm chosen at time slot t. Assume first that arms are chosen randomly and independently during the initial phase with probability $p_j := E[1\{I_t = j\}]$, $j = 1, ..., n$. In the case of uniformly random strategy we have $p_j = 1/n$. Let further D be the maximum possible delay between choosing the arm and observing the realization ($D = 1$ corresponds to no delay) and

$$c_j := D^2 + \frac{\Delta_j}{2} D + \frac{\Delta_j}{2} p_* D,$$

where $\Delta_j = \mu_j - \mu_*$. Then, we have the following result.

Theorem 1 *If during the exploration phase we choose the arms randomly and independently with uniform distribution ($p_j = 1/n$), and at the end of the exploration period, at slot t_0, we choose the arm according to the estimated average, the probability of choosing the best arm is lower bounded by*

$$P[\bar{X}_*(t_0) < \min_{j \neq *} \bar{X}_j(t_0)]$$

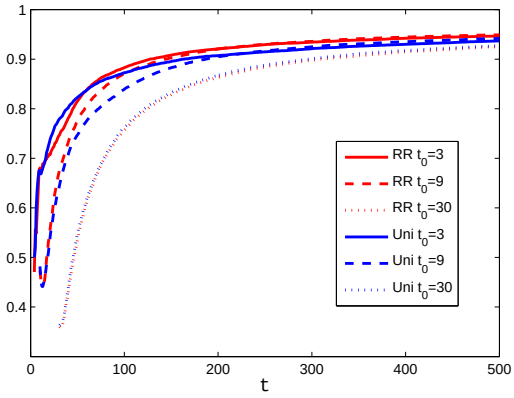


Figure 3: The effect of the initial phase duration and initial strategy: ε-greedy algorithm.

$$\geq \prod_{j \neq *} \left(1 - \exp\left(-\frac{\Delta_j^2/4(t_0 - D)^2}{2n^2 c_j^2 t_0}\right)\right)^2 \tag{1}$$

A strong point of the above result is that the derived lower bound is given in terms of exponential function, which means that starting from some value of t_0 the probability of success will be very high. However, the bound (1) can be loose. Therefore, next we suggest an approximation of the success probability based on the central limit theorem.

Also, it turns out that if the maximal delay is not too large, we do not introduce a large error by considering only realizations sent by the time $t_0 - D$. Then, by the time t_0 we observe reply from all sent interests.

Theorem 2 *If during the exploration phase we choose the arms randomly and independently with uniform distribution ($p_j = p_* = 1/n$), and if at the end of the exploration period, at slot t_0, we choose the arm according to the estimated average, the probability of choosing the best arm can be approximated as follows:*

$$P[\bar{X}_*(t_0 - D) < \min_{j \neq *} \bar{X}_j(t_0 - D)]$$

$$\approx \prod_{j \neq *} \Phi\left(\frac{\Delta_j p_j \sqrt{t_0 - D}}{2\sqrt{p_j Var(X_j) + \Delta_j^2 p_j(1 - p_j)/4}}\right)$$

$$\Phi\left(\frac{\Delta_j p_* \sqrt{t_0 - D}}{2\sqrt{p_* Var(X_*) + \Delta_j^2 p_*(1 - p_*)/4}}\right), \tag{2}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal random variable.

In the case when the Round Robin strategy is used in the initial phase, we can provide even sharper approximations.

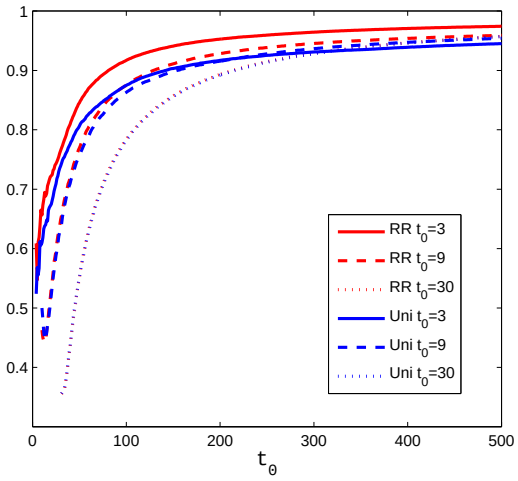


Figure 4: The effect of the initial phase duration and initial strategy: tuned ε-greedy algorithm.

Theorem 3 *If during the exploration phase we choose the arms according to the Round Robin strategy with the first arm and the order chosen randomly with the uniform distribution, and if at the end of the exploration period, at slot t_0, we choose the arm according to the estimated average, the probability of choosing the best arm can be approximated as follows:*

$$P[\bar{X}_*(t_0 - D) < \min_{j \neq *} \bar{X}_j(t_0 - D)]$$

$$\approx \prod_{j \neq *} \Phi\left(\Delta_j \sqrt{\frac{t_0 - D}{3(Var(X_*) + Var(X_j))}}\right). \tag{3}$$

We consider now our numerical example with truncated negative binomial distributions with $D = 15$. In Figure 6 we plot the approximations (2) and (3), which firstly confirm that it is enough to have a very short initial phase and secondly confirm our intuition that the Round Robin strategy is better than the random strategy.

One may be interested in rough estimation of the number of time slots after which using estimated averages the optimal arm will be selected with very high probability. We can provide recommendation for such number based on (3) and 2-sigma rule. If the arguments of the standard normal distribution function are equal to two, then respective probabilities are greater than 0.977. Thus, we conclude that after the time

$$T \geq D + 12\frac{Var(X_*) + \max_j Var(X_j)}{\min_j \Delta_j^2}, \tag{4}$$

with probability at least 0.977^{n-1}. In our numerical example, after 68 time slots the probability of choosing correctly the optimal arm is estimated to be more than 0.95.

4 Logarithmic bound for the tuned ε-greedy algorithm

In this section we finally prove that the regret (suboptimality) of employing the tuned ε-greedy algorithm can grow logarithmically in t, which is the same result as for the case without delay

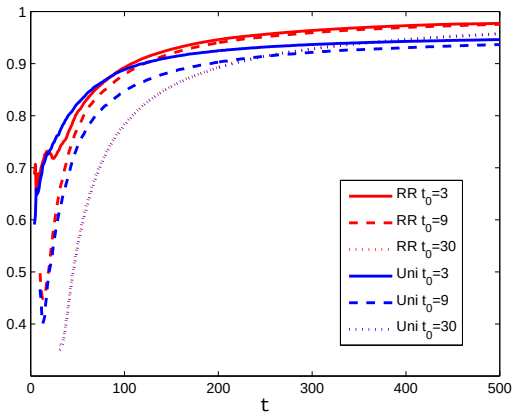


Figure 5: The effect of the initial phase duration and initial strategy: UCB algorithm.

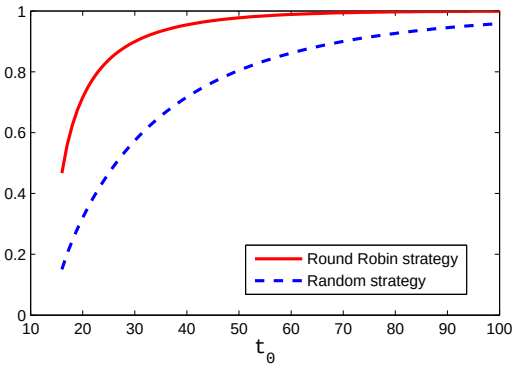


Figure 6: Approximations for the probability of choosing the optimal arm at the end of the initial phase.

(and known to be the best possible) [1].

Theorem 4 *Let observation delays follow arbitrary distributions bounded by D. Let $c >$ and $0 < d \leq min_{k:\mu_k>\mu_*}\Delta_k$, and let initial phase be run with the uniformly random strategy. For all $K > 1$ and for all delay distributions $F_1, \ldots, F_K$ with support in $[1, D]$, if algorithm tuned ε-greedy is run with input parameters $t_0 > cK/d^2$ and $\varepsilon_0 := cK/d^2$, then the probability that after any number $t \geq t_0$ of slots algorithm tuned ε-greedy chooses in slot $t+1$ a suboptimal arm j is at most*

$$2(D+1)\frac{c}{d^2}\ln\frac{td^2e^{1/2}}{cK}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{3c}{14d^2}}$$
$$+\frac{4(D+1)}{d^2}\exp\left\{\frac{D+1}{2}\right\}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{c}{2}}+\frac{c}{d^2(t+1)}.$$

This bound is logarithmic for c large enough (surely, if $c > 14/3D^2$), because the instantaneous suboptimality at any slot t large enough is of the order $(K-1)c/d^2t + o(1/t)$ for $t \to \infty$.

The main drawback is that we need to know a lower bound d, for the difference between the mean delays of the best and the strictly second-best arm.

5 Conclusion

The contribution of this paper is twofold. First, we have proposed tractable and well-performing interest forwarding algorithms for CCN networks. We have shown that the learning of the locations of chunks of interest is reasonable fast and logarithmically few interests are send sub-optimally, which means that the user and router resources are managed efficiently. Theoretical bounds show that the learning process is best achievable.

Second, we have also contributed to the theory of the multi-armed bandit problem with delayed information. This is an important and challenging topic with few existing results. We have provided finite-time analysis of algorithms extended to this setting and showed that the deterioration of their performance due to delays is not significant. Perhaps surprisingly, there is no need to include a long exploratory phase, just a single datum from each arm is sufficient for an efficient performance of the algorithms.

Acknowledgement

We would like to thank Bruno Kauffmann, Luca Muscariello and Alain Simonian for stimulating discussions.

References

[1] P. Auer, N. Cesa-Bianchi and P. Fischer, "Finite-time analysis of the multiarmed bandit problem", *Machine Learning*, v.47, pp.235-256, 2002.

[2] G. Bennett, "Probability inequalities for the sum of independent random variables", *Journal of the American Statistical Association* 57, pp. 33-45, 1962.

[3] G. Carofiglio, M. Gallo, L. Muscariello, "ICP: Design and evaluation of an interest control protocol for content-centric networking", in Proceedings of IEEE INFOCOM Workshop on emerging design choices in name oriented networking, Orlando, USA, March 2012.

[4] S.G. Eick, "Gittins procedures for bandits with delayed responses", *Journal of the Royal Statistical Society. Series B (Methodological)*, v. 50(1), pp.125-132, 1988.

[5] J. C. Gittins, "Bandit processes and dynamic allocation indices", *Journal of the Royal Statistical Society, Series B*, v. 41(2), pp.148-177, 1979.

[6] M. Gritter and D.R. Cheriton, "An architecture for content routing support in the internet", in Proceedings of the USENIX Symposium on Internet Technologies and Systems, March 2001.

[7] W. Hoeffding, "Probability inequalities for sums of bounded random variables", *Journal of the American Statistical Association* 58, pp. 13-30, 1963.

[8] V. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs and R. Braynard, "Networking named content", in Proceedings of ACM CoNEXT 2009.

[9] T.L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules", *Advances in Applied Mathematics*, v.6(1), pp.4-22, 1985.

[10] T. Koponen, M. Chawla, B. Chun, A. Ermolinskiy, K. Kim, S. Shenker and I. Stoica, "A data-oriented (and beyond) network architecture", in Proceedings of ACM SIGCOMM 2007.

[11] D. Pollard, *Convergence of Stochastic Processes*, Springer-Verlag, 1984.

[12] R. Sutton and A. Barto, *Reinforcement learning: An Introduction*, MIT Press, 1998.

A Proofs

A.1 Proof of Theorem 1

We need to evaluate the following probability:

$$P[\bar{X}_{*,T_*(t)} < \min_{j\neq *} \bar{X}_{j,T_j(t)}] = P[\cap_{j\neq *}\{\bar{X}_{*,T_*(t)} < \bar{X}_{j,T_j(t)}\}]$$

$$= \prod_{j\neq *} P[\bar{X}_{*,T_*(t)} < \bar{X}_{j,T_j(t)}]$$

$$\geq \prod_{j\neq *} P[\{\bar{X}_{*,T_*(t)} < \mu_* + \frac{\Delta_j}{2}\} \cap \{\bar{X}_{j,T_j(t)} \geq \mu_j - \frac{\Delta_j}{2}\}]$$

$$= \prod_{j\neq *} P[\bar{X}_{*,T_*(t)} < \mu_* + \frac{\Delta_j}{2}] P[\bar{X}_{j,T_j(t)} \geq \mu_j - \frac{\Delta_j}{2}]. \quad (5)$$

Now let us estimate the probability $P[\bar{X}_{*,T_*(t)} < \mu_* + \frac{\Delta_j}{2}]$.

$$P[\bar{X}_{*,T_*(t)} < \mu_* + \frac{\Delta_j}{2}] = 1 - P[\bar{X}_{*,T_*(t)} \geq \mu_* + \frac{\Delta_j}{2}]$$

$$= 1 - P\left[\frac{\sum_{s=1}^t 1\{I_s = *\} X_*(s) 1\{s + X_*(s) \leq t\}}{\sum_{s=1}^t 1\{I_s = *\} 1\{s + X_*(s) \leq t\}} \geq \mu_* + \frac{\Delta_j}{2}\right]$$

$$= 1 - P\left[\sum_{s=1}^t 1\{I_s = *\}(X_*(s) - \mu_*) 1\{s + X_*(s) \leq t\} \geq \frac{\Delta_j}{2} \sum_{s=1}^t 1\{I_s = *\} 1\{s + X_*(s) \leq t\}\right]$$

$$= 1 - P\left[\sum_{s=1}^t 1\{I_s = *\}(X_*(s) - \mu_*) 1\{s + X_*(s) \leq t\}\right.$$

$$\left. - \frac{\Delta_j}{2} \sum_{s=1}^t (1\{I_s = *\} - p_*) 1\{s + X_*(s) \leq t\} \geq \frac{\Delta_j}{2} p_* \sum_{s=1}^t 1\{s + X_*(s) \leq t\}\right]$$

$$= 1 - P\left[\sum_{s=1}^t 1\{I_s = *\}(X_*(s) - \mu_*) 1\{s + X_*(s) \leq t\}\right.$$

$$\left. - \frac{\Delta_j}{2} \sum_{s=1}^t (1\{I_s = *\} - p_*) 1\{s + X_*(s) \leq t\} - \frac{\Delta_j}{2} p_* \sum_{s=1}^t (1\{s + X_*(s) \leq t\} - q_{*,t-s})\right.$$

$$\geq \frac{\Delta_j}{2} p_*(t - X_{*,max} + \sum_{i=1}^{X_{*,max}} q_{*,i}) \Bigg],$$

where $q_{*,i} := P[X_*(t) \leq i]$.

Next we define

$$Z_t := \sum_{s=1}^{t} 1\{I_s = *\}(X_*(s) - \mu_*)1\{s + X_*(s) \leq t\}$$

$$-\frac{\Delta_j}{2} \sum_{s=1}^{t} (1\{I_s = *\} - p_*)1\{s + X_*(s) \leq t\}$$

$$-\frac{\Delta_j}{2} p_* \sum_{s=1}^{t} (1\{s + X_*(s) \leq t\} - q_{*,t-s}).$$

It is a martingale (with respect to the sequence of the observed delays) with zero mean and bounded increment

$$|Z_t - Z_{t-1}| \leq c_t,$$

with $c_t = c$.

Thus, we can apply Azuma's inequality for martingales, which says that

$$P[Z_t \geq \lambda] \leq \exp\left(-\frac{\lambda^2}{2\sum_{s=1}^{t} c_s}\right)$$

and gives in our case

$$P[\bar{X}_{*,T_*(t)} < \mu_* + \frac{\Delta_j}{2}] \geq 1 - \exp\left(-\frac{\Delta_j^2/4p_*^2(t - X_{*,max})^2}{2c^2 t}\right). \tag{6}$$

Similarly, we have

$$P[\bar{X}_{j,T_j(t)} \geq \mu_j - \frac{\Delta_j}{2}] \geq 1 - \exp\left(-\frac{\Delta_j^2/4p_*^2(t - X_{*,max})^2}{2c^2 t}\right). \tag{7}$$

Substituting (6) and (7) into (5), we obtain the following results.

A.2 Proof of Theorem 2

Similarly to (5), we have

$$P[\bar{X}_{*,T_*(t-D)} < \min_{j \neq *} \bar{X}_{j,T_j(t-D)}]$$

$$\geq \prod_{j \neq *} P[\bar{X}_{*,T_*(t-D)} < \mu_* + \frac{\Delta_j}{2}] P[\bar{X}_{j,T_j(t-D)} \geq \mu_j - \frac{\Delta_j}{2}] \tag{8}$$

Define

$$Y_t = \sum_{s=1}^{t} \left(1\{I_s = *\}(X_{*,s} - \mu_*) - \frac{\Delta_j}{2}(1\{I_s = *\} - p_*)\right).$$

Then, we can use the Central Limit theorem to estimate the probability

$$P[\bar{X}_{*,T_*(t-D)} < \mu_* + \frac{\Delta_j}{2}] = P[Y_{t-D} < \frac{\Delta_j}{2} p_*(t - D)]$$

$$= P[\frac{Y_{t-D}}{\sqrt{(t-D)(p_* Var(X_*) + \Delta_j^2 p_*(1-p_*)/4)}} <$$
$$\frac{\Delta_j p_* (t-D)}{2\sqrt{(t-D)(p_* Var(X_*) + \Delta_j^2 p_*(1-p_*)/4)}}],$$

which gives

$$P[\bar{X}_{*,T_*(t-D)} < \mu_* + \frac{\Delta_j}{2}] \approx \Phi\left(\frac{\Delta_j p_* \sqrt{t-D}}{2\sqrt{p_* Var(X_*) + \Delta_j^2 p_*(1-p_*)/4}}\right), \tag{9}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Similarly, we obtain

$$P[\bar{X}_{j,T_j(t-D)} \geq \mu_j - \frac{\Delta_j}{2}] \approx \Phi\left(\frac{\Delta_j p_j \sqrt{t-D}}{2\sqrt{p_j Var(X_j) + \Delta_j^2 p_j(1-p_j)/4}}\right). \tag{10}$$

The substitution of (9) and (10) into (8) yields the result.

A.3 Auxiliary Material

In this subsection we state some results to be used in the proof of Theorem 4.

Let us first state the Chernoff-Hoeffding bound in a general form. This is called Hoeffding's inequality in [11, p. 191], citing [7].

Theorem 5 (Chernoff-Hoeffding bound) *Let $Y_1, Y_2, \ldots, Y_T$ be independent random variables with zero means and bounded ranges $a_t \leq Y_t \leq b_t$. Then, for each $\eta > 0$,*

$$\mathbb{P}[Y_1 + Y_2 + \cdots + Y_T \geq \eta] = \mathbb{P}[Y_1 + Y_2 + \cdots + Y_T \leq -\eta] \leq \exp\left\{-2\eta^2 / \sum_{t=1}^{T} (b_t - a_t)^2\right\}$$

Let us state also the Bennett's inequality [2] and its consequence, the Bernstein inequality.

Theorem 6 (Bennett's inequality) *Let $Y_1, Y_2, \ldots, Y_T$ be independent random variables with zero means and bounded ranges $-M \leq Y_t \leq M$. Write σ_t^2 for the variance of Y_t. Suppose $V \geq \sigma_1^2 + \cdots + \sigma_T^2$. Then, for each $\eta > 0$,*

$$\mathbb{P}[Y_1 + Y_2 + \cdots + Y_T \geq \eta] = \mathbb{P}[Y_1 + Y_2 + \cdots + Y_T \leq -\eta] \leq \exp\left\{-\frac{1}{2}\eta^2 V^{-1} B(M\eta V^{-1})\right\},$$

where $B(\lambda) := 2\lambda^{-2}[(1+\lambda)\log(1+\lambda) - \lambda]$, for $\lambda > 0$.

According to [11, p. 193]:

> The function $B(\cdot)$ is well-behaved: continuous, decreasing, and $B(0+) = 1$. When λ is large, $B(\lambda) \approx 2\lambda^{-1}\log\lambda$ in the sense that the ratio tends to one as $\lambda \to \infty$; the Bennett Inequality does not give a true exponential bound for η compared to V/M. For smaller η it comes very close to the bound for normal tail probabilities. Problem 2 shows that $B(\lambda) \geq (1 + \frac{1}{3}\lambda)^{-1}$ for all $\lambda > 0$.

Using the last bound, we get the Bernstein's inequality.

Theorem 7 (Bernstein's inequality) *Let $Y_1, Y_2, \ldots, Y_T$ be independent random variables with zero means and bounded ranges $-M \leq Y_t \leq M$. Write σ_t^2 for the variance of Y_t. Suppose $V \geq \sigma_1^2 + \cdots + \sigma_T^2$. Then, for each $\eta > 0$,*

$$\begin{aligned}\mathbb{P}[Y_1 + Y_2 + \cdots + Y_T \geq \eta] &= \mathbb{P}[Y_1 + Y_2 + \cdots + Y_T \leq -\eta] \\ &\leq \exp\left\{-\frac{1}{2}\eta^2/(V + \frac{1}{3}M\eta)\right\}.\end{aligned}$$

A.4 Proof of Theorem 4

Assume $t \geq cK/d^2$ (i.e., we are in the exploitation phase).

Let $\overline{X}_{j,s}$ be the sample mean of observed rewards if arm j was chosen s times conditioned on the delay distribution. Let $\overline{X}_{j,s,u}$ be the sample mean of observed rewards if arm j was chosen s times using u observations.

Let I_t denote the arm chosen at time t. Let $S_j(t)$ denote the number of times arm j was chosen in the first t slots (given a policy). Then we have

$$\mathbb{P}\left[I_{t+1} = j\right] = (1 - \varepsilon_{t+1})\,\mathbb{P}\left[\overline{X}_{j,S_j(t)} \geq \max_{k \neq j} \overline{X}_{k,S_k(t)}\right] + \frac{\varepsilon_{t+1}}{K}.$$

(Here we should have either put $>$ or defined a tie-breaking rule.)

If $j \neq *$ (where $*$ denotes the best arm), then we can bound it by

$$\begin{aligned}\mathbb{P}\left[I_{t+1} = j\right] &\leq \mathbb{P}\left[\overline{X}_{j,S_j(t)} \geq \overline{X}_{*,S_*(t)}\right] + \frac{\varepsilon_{t+1}}{K} \\ &\leq \mathbb{P}\left[\overline{X}_{j,S_j(t)} \geq \mu_j + \frac{\Delta_j}{2}\right] + \mathbb{P}\left[\overline{X}_{*,S_*(t)} \leq \mu_* - \frac{\Delta_j}{2}\right] + \frac{\varepsilon_{t+1}}{K}.\end{aligned}$$

Let now $U_{j,s}(t)$ denote the number of observed realizations from choosing arm j in the first t slots given that it was chosen $s = S_j(t)$ times (given a policy). Let us study the first term next.

$$\begin{aligned}&\mathbb{P}\left[\overline{X}_{j,S_j(t)} \geq \mu_j + \frac{\Delta_j}{2}\right] = \sum_{s=1}^{t} \mathbb{P}\left[S_j(t) = s \text{ and } \overline{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right] \\ &= \sum_{s=1}^{t} \mathbb{P}\left[S_j(t) = s \mid \overline{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right] \mathbb{P}\left[\overline{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right] \\ &= \sum_{s=1}^{t} \mathbb{P}\left[S_j(t) = s \mid \overline{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right] \sum_{u=1}^{s} \mathbb{P}\left[U_{j,s}(t) = u \text{ and } \overline{X}_{j,s,u} \geq \mu_j + \frac{\Delta_j}{2}\right] \\ &= \sum_{s=1}^{t} \mathbb{P}\left[S_j(t) = s \mid \overline{X}_{j,s} \geq \mu_j + \frac{\Delta_j}{2}\right] \sum_{u=1}^{s} \mathbb{P}\left[U_{j,s}(t) = u \mid \overline{X}_{j,s,u} \geq \mu_j + \frac{\Delta_j}{2}\right] \mathbb{P}\left[\overline{X}_{j,s,u} \geq \mu_j + \frac{\Delta_j}{2}\right].\end{aligned}$$

Assuming that $\mathbb{P}\left[\overline{X}_{j,s,u} \geq \mu_j + \frac{\Delta_j}{2}\right] > 0$, then, for $1 \leq u \leq s$,

$$\mathbb{P}\left[U_{j,s}(t) = u \mid \overline{X}_{j,s,u} \geq \mu_j + \frac{\Delta_j}{2}\right] \begin{cases} = 0, & \text{if } s - D > u, \\ \leq 1, & \text{if } s - D \leq u, \end{cases}$$

because there can be at most D non-observed realizations of the chosen arms.

So,

$$\sum_{u=1}^{s}\mathbb{P}\left[U_{j,s}(t)=u \mid \overline{X}_{j,s,u}\geq\mu_j+\frac{\Delta_j}{2}\right]\mathbb{P}\left[\overline{X}_{j,s,u}\geq\mu_j+\frac{\Delta_j}{2}\right]$$
$$\leq\sum_{u=\max\{1,s-D\}}^{s}\mathbb{P}\left[\overline{X}_{j,s,u}u\geq\mu_j u+\frac{\Delta_j u}{2}\right]$$
$$\leq\sum_{u=\max\{1,s-D\}}^{s}\exp\left\{-2\left(\frac{\Delta_j^2u^2}{2^2}\right)/u\right\}=\sum_{u=\max\{1,s-D\}}^{s}\exp\left\{-\left(\frac{\Delta_j^2u}{2}\right)\right\},$$

where the last inequality is due to the Chernoff-Hoeffding bound.

Upperbounding the last geometric sum by a sum of constants equal to the first term, we further have

$$\sum_{u=1}^{s}\mathbb{P}\left[U_{j,s}(t)=u \mid \overline{X}_{j,s,u}\geq\mu_j+\frac{\Delta_j}{2}\right]\mathbb{P}\left[\overline{X}_{j,s,u}\geq\mu_j+\frac{\Delta_j}{2}\right]$$
$$\leq\exp\left\{-\frac{\Delta_j^2}{2}\max\{1,s-D\}\right\}(D+1).$$

This bound therefore gives us

$$\mathbb{P}\left[\overline{X}_{j,S_j(t)}\geq\mu_j+\frac{\Delta_j}{2}\right]$$
$$\leq(D+1)\sum_{s=1}^{t}\mathbb{P}\left[S_j(t)=s \mid \overline{X}_{j,s}\geq\mu_j+\frac{\Delta_j}{2}\right]\exp\left\{-\frac{\Delta_j^2}{2}\max\{1,s-D\}\right\}$$
$$\leq(D+1)\sum_{s=1}^{\infty}\mathbb{P}\left[S_j(t)=s \mid \overline{X}_{j,s}\geq\mu_j+\frac{\Delta_j}{2}\right]\exp\left\{-\frac{\Delta_j^2}{2}\max\{1,s-D\}\right\}$$
$$\leq(D+1)\exp\left\{-\frac{\Delta_j^2}{2}\right\}\sum_{s=1}^{D}\mathbb{P}\left[S_j(t)=s \mid \overline{X}_{j,s}\geq\mu_j+\frac{\Delta_j}{2}\right]$$
$$+(D+1)\sum_{s=D+1}^{\lfloor L\rfloor}\mathbb{P}\left[S_j(t)=s \mid \overline{X}_{j,s}\geq\mu_j+\frac{\Delta_j}{2}\right]\exp\left\{-\frac{\Delta_j^2}{2}(s-D)\right\}$$
$$+(D+1)\sum_{s=\lfloor L\rfloor+1}^{\infty}\mathbb{P}\left[S_j(t)=s \mid \overline{X}_{j,s}\geq\mu_j+\frac{\Delta_j}{2}\right]\exp\left\{-\frac{\Delta_j^2}{2}(s-D)\right\}$$

where

$$L:=\frac{1}{2K}\sum_{s=1}^{t}\varepsilon_s.$$

(The first term is new, the second and third ones correspond to the original ones with $L=x_0$.)

Note that if $\lfloor L\rfloor\geq D$, then the above decomposition of the sum in the last step in fact holds as equality. In case $\lfloor L\rfloor<D$, the second term is zero and some of the summands appear both in the first and in the third term, therefore the inequality holds.

The sum of the first and second terms can be upperbounded by

$$(D+1)\sum_{s=1}^{\lfloor L\rfloor}\mathbb{P}\left[S_j(t)=s \mid \overline{X}_{j,s}\geq \mu_j+\frac{\Delta_j}{2}\right]$$

omitting the exponential terms (≤ 1), which is further upperbounded (as before) by

$$(D+1)\sum_{s=1}^{\lfloor L\rfloor}\mathbb{P}\left[S_j^{\mathrm{R}}(t)\leq s \mid \overline{X}_{j,s}\geq \mu_j+\frac{\Delta_j}{2}\right]\leq (D+1)L\,\mathbb{P}\left[S_j^{\mathrm{R}}(t)\leq L\right].$$

Using the Bernstein inequality, we have (a slightly tighter upperbound than before)

$$\mathbb{P}\left[S_j^{\mathrm{R}}(t)\leq L\right]\leq \exp\{-\frac{3}{14}L\}$$

and for $t\geq cK/d^2$, we lowerbound L as before,

$$L\geq \frac{c}{d^2}\ln\frac{td^2e^{1/2}}{cK}.$$

Therefore, the sum of the first and second terms can be upperbounded by

$$(D+1)\frac{c}{d^2}\ln\frac{td^2e^{1/2}}{cK}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{3c}{14d^2}}.$$

As before, the third term can be upperbounded by

$$\frac{2(D+1)}{\Delta_j^2}\exp\left\{-\frac{\Delta_j^2}{2}(\lfloor L\rfloor-D)\right\}=\frac{2(D+1)}{\Delta_j^2}\exp\left\{\frac{\Delta_j^2}{2}D\right\}\exp\left\{-\frac{\Delta_j^2}{2}\lfloor L\rfloor\right\}$$

omitting the probability term (≤ 1) and using $\sum_{s=K+1}^{\infty}e^{-\alpha s}\leq \frac{1}{\alpha}e^{-\alpha K}$. Further, using $\lfloor L\rfloor\geq L-1$, this can be upperbounded by

$$\frac{2(D+1)}{\Delta_j^2}\exp\left\{\frac{\Delta_j^2(D+1)}{2}\right\}\exp\left\{-\frac{\Delta_j^2}{2}L\right\}$$

and further by

$$\frac{2(D+1)}{d^2}\exp\left\{\frac{D+1}{2}\right\}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{c}{2}}.$$

So, we have

$$\mathbb{P}\left[\overline{X}_{j,S_j(t)}\geq \mu_j+\frac{\Delta_j}{2}\right]\leq (D+1)\frac{c}{d^2}\ln\frac{td^2e^{1/2}}{cK}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{3c}{14d^2}}+\frac{2(D+1)}{d^2}\exp\left\{\frac{D+1}{2}\right\}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{c}{2}}.$$

In fact, the same upperbound holds for $\mathbb{P}\left[\overline{X}_{*,S_*(t)}\leq \mu_*-\frac{\Delta_j}{2}\right]$.

Finally, we have $\varepsilon_t=cK/d^2t$, therefore

$$\mathbb{P}\left[I_{t+1}=j\right]\leq 2(D+1)\frac{c}{d^2}\ln\frac{td^2e^{1/2}}{cK}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{3c}{14d^2}}+\frac{4(D+1)}{d^2}\exp\left\{\frac{D+1}{2}\right\}\left(\frac{cK}{td^2e^{1/2}}\right)^{\frac{c}{2}}+\frac{c}{d^2(t+1)}.$$

Contents

1 Introduction **3**

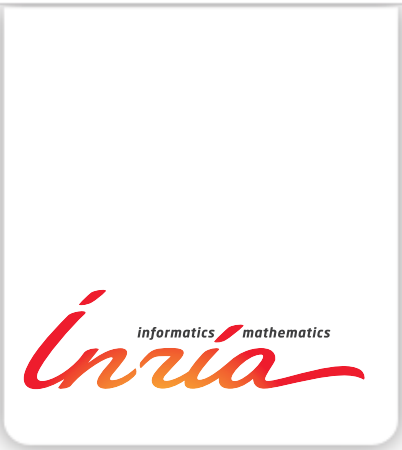
2 Model and interest forwarding strategies **3**

3 Analysis of initial exploratory phase **6**

4 Logarithmic bound for the tuned ε-greedy algorithm **9**

5 Conclusion **11**

A Proofs **12**
A.1 Proof of Theorem 1 12
A.2 Proof of Theorem 2 13
A.3 Auxiliary Material 14
A.4 Proof of Theorem 4 15



RESEARCH CENTRE
SOPHIA ANTIPOLIS – MÉDITERRANÉE

2004 route des Lucioles - BP 93
06902 Sophia Antipolis Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399