A learning algorithm for the Whittle index policy for scheduling web crawlers

Konstantin Avrachenkov, Vivek S Borkar

A learning algorithm for the Whittle index policy for scheduling web crawlers

Konstantin Avrachenkov
INRIA Sophia Antipolis,
2004, Route des Lucioles, B.P.93,
06902, Sophia Antipolis, France
E-mail: k.avrachenkov@inria.fr

Vivek S. Borkar
Department of Electrical Engineering,
Indian Institute of Technology Bombay,
Powai, Mumbai 400076, India.
E-mail: borkar.vs@gmail.com

***Abstract*—We revisit the Whittle index policy for scheduling web crawlers for ephemeral content proposed in Avrachenkov and Borkar, IEEE Trans. Control of Network Systems 5(1), 2016, and develop a reinforcement learning scheme for it based on LSPE(0). The scheme leverages the known structural properties of the Whittle index policy.**

I. Introduction

Reinforcement learning for approximate dynamic programming has been a popular approach to approximate policy evaluation or computation for Markov decision processes with large state spaces. A data driven iterative scheme based on real or simulated data, it offers the possibility of working around, to a significant extent, the so called curse of dimensionality which these problems suffer from. A further and very significant reduction in complexity is possible if one exploits the explicitly known structural properties, if any, of the problem. Recently there have been a few works that do so, e.g., exploiting threshold nature of the policy [11], [16], parametrized threshold policies [19], or index policies such as the Whittle policy for restless bandits [5]. In particular, [5] addressed the problem of learning the Whittle policy when Whittle indexability is known, but structure of the index is not explicitly known. This work addresses a situation when the latter is known in a parametric form, but the parameters are unknown, and the algorithm is required to operate online with streaming real time data. As a test case, we consider the specific problem of scheduling web crawlers for ephemeral content analyzed in [2] (see also [3], [14] for related work). We describe this problem in the next section and summarize the main results of [2]. Some preliminary analysis of the problem is given in section III. This is followed by the proposed learning scheme in section IV. Section V presents some supporting numerical experiments.

II. Problem formulation

Consider crawling decisions at discrete times $0, 1, 2, \cdots$. The content arrives in continuous time according to a Poisson process at site i with rate λ_i for $1 \leq i \leq N$. Content at site i more than M_i time units old is dropped. The interest in the content at site i less than M_i time units old decays exponentially with rate $c_i > 0$ [6], [20], [12]. Let $X_i(n)$ denote the cumulative interest in content at site i at time n. Consider ζ items arriving at site i during the time interval $[m, n]$, $n > m$, at (random) times $\{\tau(i), 1 \leq i \leq \zeta\}$. Then $\{\tau(i)\}$ are IID uniform on $[m, n]$ conditioned on $\zeta =$ (say) k, and

$$p_i(k) := P(\zeta = k) = \frac{(\lambda_i(n-m))^k}{k!} e^{-\lambda_i(n-m)}.$$

Define

$$\begin{aligned}
\alpha_i &= e^{-c_i}, \\
Z_i(n) &:= \sum_{\{m: n-1 \leq \tau_i(m) < n\}} \alpha_i^{n-\tau(m)}, \\
X_i(n) &= \sum_{m=n-M+1}^{n} \alpha_i^{n-m} Z_i(m), \\
\xi_i(n) &= Z_i(n) - \alpha_i^{M_i} Z_i(n - M_i), \\
u_i = u_i(\lambda_i, \alpha_i) &:= E\left[\xi_i(n)\right]
\end{aligned}$$

$$= \left(1-\alpha_i^{M_i}\right)\sum_k kp_i(k)\times \int_{n-1}^{n}\alpha_i^{n-s}ds$$
$$= \left((1-\alpha_i)(1-\alpha_i^{M_i})\right)\frac{\lambda_i}{c_i}.$$

Then $\{Z_i(n)\}$ are i.i.d. and

$$X_i(n+1)=\alpha_i X_i(n)+\xi_i(n+1). \quad (1)$$

When not crawled, the state is not observed, so the mean dynamics becomes (with some abuse of notation)

$$X_i(n+1)=\alpha_i X_i(n)+u_i. \quad (2)$$

When crawled, $X_i(n)$ is instantaneously reset to zero at time n_+ and

$$X_i(n+1)=u_i. \quad (3)$$

Thus the observed dynamics is effectively the deterministic one given by (2)-(3): when you crawl, you do observe the actual random state at the end of the interval, but the state is instantly reset to zero, so that knowledge is no longer relevant. The objective is to maximize the average interest level

$$\liminf_{n\uparrow\infty}\frac{1}{n}\sum_{m=0}^{n-1}\sum_{i=1}^{N}X_i(n)\nu_i(n)$$

where $\nu_i(n)=1$ if location i is crawled at time n and 0 otherwise, subject to the constraint that only $N_0<N$ crawlers can be active at any time. This can be cast as a restless bandit problem that is Whittle indexable [2]. The Whittle index is given by [2]

$$\gamma_i(x,\lambda_i,\alpha_i) := \frac{1}{C_i}\Big[\zeta_i(x,\lambda_i,\alpha_i)((1-\alpha_i)x-u_i) + \left(\frac{1-\alpha_i^{\zeta_i(x,\lambda_i,\alpha_i)}}{1-\alpha_i}\right)u_i\Big] \quad (4)$$

where

$$\zeta_i(x,\lambda_i,\alpha_i):=\left\lceil \log^+_{\alpha_i}\left(\frac{u_i-(1-\alpha_i)x}{u_i}\right)\right\rceil. \quad (5)$$

Here $\lceil\cdots\rceil$ stands for 'the least integer exceeding $\cdots$' and we have rendered explicit the dependence on the unknown arrival rate λ_i and the unknown decay rate α_i via u_i. Let

$$\begin{aligned}\lambda &:= [\lambda_1,\cdots,\lambda_N]\in\mathcal{R}^N,\\ \alpha &:= [\alpha_1,\cdots,\alpha_N]\in(0,1)^N,\\ \Gamma &:= [\lambda:\alpha]\in\mathcal{R}^N\times(0,1)^N,\\ u(\Gamma) &:= [u_1(\lambda_1,\alpha_1),\cdots,u_N(\lambda_N,\alpha_N)]^T,\\ \mathcal{U} &:= \prod_{i=1}^{N}\left[u_i,\ \frac{u_i}{1-\alpha_i}\right].\end{aligned}$$

Anticipating the randomized policies we consider later, denote the reward under the Whittle index policy by $J(\Gamma):=\sum_{i=1}^N E_s\left[X_i(n)\right]$, where $E_s[\,\cdot\,]$ denotes the stationary expectation. Let $\pi_\Gamma(dx)$ denote the stationary distribution of $X(n)=[X_1(n),\cdots,X_N(n)],\ n\geq 0$. We shall denote by ∇^y the gradient w.r.t. the variable y.

The Whittle policy is as follows: At each time t, if the state is $x(t)=[x_1(t),\cdots,x_N(t)]$, rank order $\{\gamma_i(x_i(t),\lambda_i,\alpha_i)\}$ in decreasing order (resolving any ties arbitrarily) and render the top N_0 active, the rest passive.

III. The Whittle dynamics

For fixed values of $x_{-i}:=\{x_j, j\neq i\}$, the policy for the ith process is a threshold policy with threshold

$$T_i(x_{-i},\Gamma):=\gamma_i^{-1}(\cdot,\lambda_i,\alpha_i)\Big(\max_{j\neq i}\gamma_j(x_j,\lambda_j,\alpha_j)\Big),$$

with $1\leq i\leq N$. This will be monotone increasing in x_{-i} by virtue of the structural properties established in [2]. The policy is captured by the step function

$$I\{i\text{th process active}\}=I\{\nu_i(n)=1\}= =I\{x_i\geq T_i(x_{-i},\Gamma)\},$$

where $I\{\cdots\}$ is the indicator function. Following the approach of [16], [19], we approximate this function by

$$f_i(x_i,x_{-i},\Gamma):=\frac{e^{\kappa_i(x_i-T_i(x_{-i},\Gamma)-0.5)}}{1+e^{\kappa_i(x_i-T_i(x_{-i},\Gamma)-0.5)}}$$

where $\kappa_i>0$. We treat this as the probability of picking 1, i.e., the active mode. The corresponding transition probabilities for the ith process are

$$p_i(y_i|x,\Gamma)=$$

$$\begin{cases} f_i(x_i, x_{-i}, \Gamma), & y_i = u_i, \\ 1 - f_i(x_i, x_{-i}, \Gamma), & y_i = \alpha_i x_i + u_i, \\ 0, & \text{otherwise.} \end{cases}$$

We shall use $p_i(y_i|x, \Gamma), p_i(y_i|x), p_i(y_i|x_i, x_{-i})$ interchangeably depending on the emphasis we want to put. Note that the actual operative policy is an index policy which deterministically picks at each time the N_0 sites to crawl. A probabilistic decision as above would violate the constraint. Thus the above is purely an approximation for analytic purposes. This is a common device in such learning schemes, employed so as to exploit the ease of optimization over continuous variables. If the scheme were a simulation-based off-line exercise, the probabilistic decision could be implemented, but this choice is not there for online schemes.

The transition probabilities are coupled, so the overall transition probability of the Markov chain is

$$p([y_1, \cdots, y_N]|[x_1, \cdots, x_N], \Gamma) := \\ := \prod_{i=1}^{N} p_i(y_i|x_i, x_{-i}, \Gamma).$$

We take a parametrized approximation for $T_i(x_{-i}, \Gamma)$ as $T_i(x_{-i}, \Gamma) \approx \sum_{j=1}^{s(i)} q_j^i(\Gamma)\psi_j(x_{-i})$. From now on we use the q_j^i's as surrogate parameters for Γ and suppress the Γ-dependence of q_j^i's and other entities henceforth. Let $U_i(x_i) := \{u_i, \alpha_i x_i + u_i\}$ and $U(x) := \prod_i U_i(x_i)$. The Poisson equation for this Markov chain is

$$V(x_1, \cdots, x_N) = \sum_{i=1}^{N} x_i - \beta \\ + \sum_{y \in U} \prod_{i=1}^{N} p_i(y_i|x_i, x_{-i}) V(y_1, \cdots, y_N), \quad (6)$$

where we have suppressed the dependence on the unknown parameter vector Γ. Note that under Whittle policy, the system trajectory is periodic and traverses only a finite subset of the state space. For such a case, the well-posedness of the Poisson equation, i.e., the existence of solution $(V(\cdot), \beta)$ with β uniquely characterized as $\sum_{i=1}^{N} E_s[X_i(n)]$ and $V(\cdot)$ unique up to an additive scalar, is easy to establish by a 'vanishing discount' argument. Arguing as in Proposition 1 of [10], we have, for $q^i = [q_1^i, \cdots, q_s^i]$, $q = [(q^1)^T : \cdots : (q^N)^T]^T$, and $\pi = \pi_\Gamma$,

$$\nabla^q \beta = \int \pi(dx) \left(\sum_{y \in U(x)} \nabla^q p(y|x) V(y) \right) \\ = \int \pi(dx) \Big(\sum_{y \in U(x)} p(y|x) \times \\ \nabla^q \log p(y|x) V(y) \Big). \quad (7)$$

We use this as the basis of our learning scheme described in the next section. Explicit calculation leads to

$$\begin{aligned} &\nabla^q \log p(y|x) \\ &= \sum_{i=1}^{N} \frac{\nabla^q p_i(y_i|x)}{p_i(y_i|x)} \\ &= \sum_{i=1}^{N} \Big(- \frac{I\{y_i = \alpha_i x_i + u_i\}}{p_i(\alpha_i x_i + u_i)} \times \\ &\qquad \nabla^q \left(\frac{e^{\kappa_i (x_i - \sum_j q_j^i \psi_j(x_{-i}) - 0.5)}}{1 + e^{\kappa_i (x_i - \sum_j q_j^i \psi_j(x_{-i}) - 0.5)}} \right) \\ &\quad + \frac{I\{y_i = u_i\}}{p_i(u_i|x)} \times \\ &\quad \nabla^q \left(\frac{e^{\kappa_i (x_i - \sum_j q_j^i \psi_j(x_{-i}) - 0.5)}}{1 + e^{\kappa_i (x_i - \sum_j q_j^i \psi_j(x_{-i}) - 0.5)}} \right) \Big) \\ &= \Big[\cdots, \sum_{i=1}^{N} \kappa_i \Big(\Big(\frac{I\{y_i = \alpha_i x_i + u_i\}}{p_i(\alpha_i x_i + u_i|x)} \\ &\qquad - \frac{I\{y_i = u_i\}}{p_i(u_i|x)} \Big) \times \\ &\qquad (\psi_j(x_{-i}) p_i(y_i|x)(1 - p_i(y_i|x))), \cdots \Big]^T, \end{aligned} \quad (8)$$

where we have exhibited the partial derivative w.r.t. q_j^i.

IV. The learning algorithm

Let $V(x) \approx \sum_{m=1}^{\ell} r(m)\phi_m(x)$ where the ϕ_m's are pre-selected features (e.g., polynomials). Let $\phi(x) :=$ the ℓ-vector $[\phi_1(x), \cdots, \phi_\ell(x)]^T$. We adapt the LSPE(0) algorithm from [23] as follows.

The parameter vector $r := [r_1, \cdots, r_\ell]^T$ is updated according to : for $t \geq 0$,

$$r(n+1) = r(n) + \gamma \bar{B}^{-1}(n)(\bar{A}(n)r(n) + \bar{b}(n)),$$

$$\eta(n+1) = \eta(n) + \frac{1}{n+1} \times \left(\sum_i X_i(n) - \eta(n) \right)$$

$$\bar{B}(n) = \bar{B}(n-1) + \frac{1}{n+1} \times [\phi(X(n))\phi(X(n))^T - \bar{B}(n-1)],$$

$$\bar{B}(-1) = \epsilon I,$$

$$\bar{A}(n) = \bar{A}(n-1) + \frac{1}{n+1} \times \Big(\phi(X(n)) \big(\phi(X(n+1))^T - \phi(X(n))^T \big) - \bar{A}(n-1) \Big),$$

$$\bar{b}(n) = \bar{b}(n-1) + \frac{1}{n+1} \times \Big(\phi(X(n)) \left(\sum_i X_i(n) - \eta_n \right) - \bar{b}(n-1) \Big),$$

and finally,

$$\bar{B}^{-1}(n) = \frac{n}{n-1} \left(\bar{B}^{-1}(n-1) - \frac{\bar{B}^{-1}(n)\phi(X(n))\phi(X(n))^T \bar{B}^{-1}(n-1)}{n-1+\phi(X(n))^T \bar{B}^{-1}(n-1)\phi(n)} \right).$$

The last iterate is simply the Sherman-Morrison formula to update the inverse. This is coupled with the stochastic gradient ascent for $\{q(n)\}$ based on (7) as follows:

$$q(n+1) = q(n) + a(n)\nabla^q \log p(X(n+1)|X(n)) \times \left(\sum_{i=1}^{\ell} r_i(n)\phi_i(X(n+1)) \right), \quad (9)$$

where $a(n) > 0$ satisfy

$$\sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty.$$

The term $\nabla^q \log p(X(n+1)|X(n))$ is given by (8). In the iterates for $\eta(n), \bar{B}(n), \bar{A}(n)$, we can replace the stepsize $1/(n+1)$ by $b(n)$ satisfying

$$\sum_n b(n) = \infty, \ \sum_n b(n)^2 < \infty, \ a(n) = o(b(n)).$$

An example is: for some $c \in (0,1)$,

$$a(n) = \frac{c}{\lceil \frac{n}{500} \rceil}, \ b(n) = \frac{c}{\lceil \frac{n}{1000} \rceil^{2/3}}.$$

V. Numerical experiments

We compare the performance of the proposed learning scheme with the performance of the Whittle index. For the comparison purposes we take the same numerical example as in [2]. The example has 4 arms and the parameters are given in Table I.

TABLE I. Data for numerical example

i	1	2	3	4
λ_i	250	250	250	250
c_i	0.7	0.35	0.7	0.21
M_i	∞	∞	∞	∞

The other parameters of the algorithm are as follows: $\kappa_i = 2$ and $\gamma = 0.05$. As approximating functions, we take:

$$T_i(x_{-i}) = q_0^i + q_1^i \sum\nolimits_{k \neq i} x_k + q_1^i (\sum\nolimits_{k \neq i} x_k)^2,$$

$$\phi = [x_1 \ \cdots \ x_N \ x_1^2 \ x_1 x_2 \ \cdots \ x_{N-1} x_N \ x_N^2].$$

While running the initial version of the algorithm we noticed that the algorithm quickly runs into numerical instability. We have identified two possible reasons for the instability. The first is that the value of the thresholds could reach small or large values, which poses numerical problems in evaluating the exponents of the logit distribution. The second issue, which probably comes from the deterministic nature of the original model, is that the matrix $\bar{B}$ becomes close to rank-deficient. To overcome the first issue, we project the value of the thresholds on the intervals $[u_i, u_i/(1-\alpha_i)]$. Note that according to the model definition, the state variables cannot go outside the intervals

$[u_i, u_i/(1-\alpha_i)]$, thus this is a legitimate step. To overcome the second issue, we replaced the matrix inversion $\bar{B}^{-1}$ with its Tikhonov regularization $(\bar{B}^T\bar{B} + \delta I)^{-1}\bar{B}^T$ for a small $\delta > 0$. These practical adjustments helped solve the numerical problems.

We have run the algorithm for $N = 10000$ steps. Note that the choice of κ_i controls the average value and dispersion of the number of engaged arms. With our choice $\kappa_i = 2$, we have on average one arm sampled and obtain the reward around 160, which is a value between the rewards that we could get by sampling constantly either the best or the second best arm. It is significantly smaller than what we can get by using Whittle index with one arm, viz., 283 [2]. On the other hand, our algorithm hardly uses any information about the system parameters, whereas the complete knowledge of the system parameters is needed when applying the Whittle index as in [2].

ACKNOWLEDGMENT

The authors were supported in part by the INRIA-DST project '*Machine Learning for Network Analytics*' administered by the Indo-French Centre for Promotion of Advanced Research (IFCPAR). VB was also supported by a J. C. Bose Fellowship from the Government of India. The work of KA is also supported by the PIA ANSWER project PIA FSN2 no.P159564-2661789/DOS0060094 between Inria and Qwant.

REFERENCES

[1] Agarwal, M.; Borkar, V. S. and Karandikar, A. (2008) "Structural properties of optimal transmission policies over a randomly varying channel", *IEEE Transactions on Automatic Control* 53(6), 1476-1491.

[2] Avrachenkov, A. and Borkar, V. S., "Whittle index policy for crawling ephemeral content", *IEEE Transactions on Control of Network Systems* 5(1), 2018, 446-455.

[3] Azar, Y.; Horvitz, E.; Lubetzky, R.; Peres, Y. and Shahaf, D. (2018) "Tractable near-optimal policies for crawling", *Proc. National Academy of Sciences* 115(32), 8099-8103.

[4] Bertsekas, D. P. (2012) *Dynamic Programming and Optimal Control, Vol. II* (4th ed.), Athena Scientific, Belmont, Mass.

[5] Borkar, V. S. (2018) "A reinforcement learning algorithm for restless bandits", *4th Indian Control Conference, Jan. 4-6, 2018, IIT Kanpur, India*, 89-94.

[6] Goyal, A.; Bonchi, F. and Lakshmanan, L. V. (2010) "Learning influence probabilities in social networks", *Proc. ACM WSDM*, 241250.

[7] Jacko, P. *Dynamic priority allocation in restless bandit models*, Lambert Academic Publishing, 2010.

[8] Larranaga, M., Ayesta, U. and Verloop, I.M. , "Stochastic and fluid index policies for resource allocation problems." *Proceedings of IEEE Conference on Computer Communications (INFOCOM), Hong Kong, Apr. 26 - May 1, 2015,* IEEE, 2015.

[9] Liu, K., and Zhao, Q. , "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access", *IEEE Trans. Info. Theory*, 56(11), 2010, 5547-5567.

[10] Marbach, P. and Tsitsiklis, J. N., "Simulation-based optimization of Markov reward processes", *IEEE Transactions on Automatic Control* 46(2), 2001, 191-209.

[11] Massaro, A.; De Pellegrini, F. and Maggi, L. (2019) "Optimal trunk-reservation by policy learning", *Proceedings of IEEE Conference on Computer Communications (INFOCOM), Paris, 15-19 April 2019,* IEEE, 2019.

[12] Moon, T.; Chu, W.; Li,L.; Zheng, Z. and Chang, Y. (2011) "Refining recency search results with user click feedback", ArXiv Preprint arXiv:1103.3735.

[13] Nino-Mora, J. and Villar, S. S., "Sensor scheduling for hunting elusive hiding targets via Whittle's restless bandit index policy." *5th International Conference on Network Games, Control and Optimization (NetGCooP)*, Paris, Oct. 12-14, 2011.

[14] Nino-Mora, J. (2014) "A dynamic page-refresh index policy for web crawlers", *Proceedings of International Conference on Analytical and Stochastic Modeling Techniques and Applications (ASMTA 2014), Budapest, Hungary, June 30-July 2, 2014*, 44-60.

[15] Ny, J.L., Dahleh, M. and Feron, E. , "Multi-UAV dynamic routing with partial observations using restless bandit allocation indices." *Proceedings of American Control Conference, Seattle, June 11-13, 2008,* pp. 4220-4225, 2008.

[16] Roy, A.; Borkar, V.; Chaporkar, P. and Karandikar, A. (2019) "A structure-aware online learning algorithm for Markov decision processes", *Proceeding of VALUETOOLS 2019 : The 12th EAI International Conference on Performance Evaluation Methodologies and Tools, Palma, Spain, March 12 - 15, 2019*, 71-78.

[17] Raghunathan, V., Borkar, V.S., Cao, M. and Kumar, P.R. (2008) "Index policies for real-time multicast scheduling for wireless bradcast systems." *Proceedings of the IEEE Conference on Computer Communications (INFOCOM), Phoenx, AZ, Apr. 13-18, 2008* IEEE, 2008.

[18] Ruiz-Hernandez, D. (2008) *Indexable Restless Bandits*, VDM Verlag.

[19] Roy, A.; Borkar, V.; Chaporkar, P. and Karandikar, A. (2019) "Low complexity online radio access technology selection algorithm in LTE-WiFi HetNet", *IEEE Trans. on Mobile Computing*, online (early access).

[20] Lefortier, D.; Ostroumova, L.; Samosvat, E. and

Serdyukov, P. (2013) “Timely crawling of high-quality ephemeral new content”, *Proceedings of ACM Conference on Information and Knpwledge Management (CIKM), Oct. 27-Nov. 1, 2013, San Francisco*, pp. 745750.

[21] Sutton, R. and Barto, A. (2018) *Reinforcement Learning: An Introduction* (2nd ed.), MIT Press, Cambridge, Mass.

[22] Whittle, P. , “Restless bandits: activity allocation in a changing world”, *Journal of Applied Probability Vol. 25: A Celebration of Applied Probability*, pp. 287-298, 1988.

[23] Yu, H. and Bertsekas, D. P., “Convergence results for some temporal difference methods based on least squares”, *IEEE Transactions on Automatic Control* 54(7), 2009, 1515-1531.