Boletín de Estadística e Investigación Operativa
Vol. 26, No. 2, Junio 2010, pp. 124-133

Investigación Operativa

Multi-armed restless bandits, index policies, and dynamic priority allocation

José Niño-Mora

Department of Statistics
Carlos III University of Madrid

✉ jnimora@alum.mit.edu
http://alum.mit.edu/www/jnimora

Abstract

This paper presents a brief introduction to the emerging research field of multi-armed restless bandits (MARBs), which substantially extend the modeling power of classic multi-armed bandits. MARBs are Markov decision process models for optimal dynamic priority allocation to a collection of stochastic binary-action (active/passive) projects evolving over time. Interest in MARBs has grown steadily, spurred by the breadth of their possible applications. Although MARBs are generally intractable, a Lagrangian relaxation and decomposition approach yields a unifying design principle for heuristic priority-index policies, which are often tractable and nearly optimal, along with an upper bound on the optimal reward.

Keywords: Restless bandits, index policies, priorities, stochastic scheduling.
AMS Subject classifications: 90B36, 90C40, 90B05, 90B22, 90B18.

1. Introduction

This paper presents a short introduction to the emerging research field of multi-armed restless bandits, which are extensions of substantially expanded modeling power of classic multi-armed bandits. In recent years, they have attracted growing research attention, due both to the mathematical interest of the research challenges they raise and to the breadth of possible applications. The presentation highlights the key ideas involved as well as the historical development of the field, and is biased towards the author's view developed over the last decade of work in the area. For a more technical presentation, the reader is referred to the review paper Niño-Mora [18].

The remainder of the paper is organized as follows. Section 2 clarifies the origin of the term "multi-armed restless bandits," as it may sound surprising at

first sight. Sections 3 and 4 introduce the one-armed and the multi-armed restless bandit problem, respectively. Section 5 outlines the historical development of the field, emphasizing the research challenges addressed. Section 6 lists some applied models which have been successfully addressed via restless bandits. Finally, Section 7 concludes.

2. "Multi-armed restless bandits"?

When first heard, the term "multi-armed restless bandits" may conjure up disturbing images of dangerous outlaws plundering their way around, so it is in order to begin this discussion by clarifying its origin and meaning. Recall that, in English, the term "one-armed bandit" is used to refer to a slot machine, of the kind one finds in a casino, the "arm" being the lever that the gambler pulls after tossing in the coin. One can play a slot machine repeatedly, at discrete time periods, with such plays having a dual effect: to make the machine yield random rewards, and to change its state after each play, where the machine's state is displayed visually as a colorful assortment of symbols.

3. The one-armed restless bandit problem

Such a one-armed bandit provides a compelling metaphor for a *Markov decision process* (MDP) model of a generic dynamic and stochastic *project*, whose evolution over discrete time periods is controlled by a manager, who decides at the start of each period whether the project should be *active* (worked on) or *passive* (left idle) during the period. If at the start of period t the project occupies state $X(t) = i \in \mathbb{X}$ (where $\mathbb{X}$ denotes the project's state space) and is worked on, i.e., the *active action* $a(t) = 1$ is taken, it yields an immediate random reward with mean $R(i, 1)$, and its state moves to $X(t + 1) = j \in \mathbb{X}$ in a Markovian fashion with transition probability $p(i, j|1)$. If, on the other hand, the project is left idle, i.e., the *passive action* $a(t) = 0$ is taken, it yields an immediate reward with mean $R(i, 0)$, and its state moves to $X(t + 1) = j$ with probability $p(i, j|0)$. In a *classic bandit* model, a passive project's state remains frozen, so $p(i, i|0) \equiv 1$. In contrast, in a *restless bandit* model a passive project can change state, typically according to a different transition law than when the project is active.

Let us incorporate into such a model a scalar parameter, denoted by λ, which models the *charge* incurred per active period (e.g., the charge per play in the slot machine), so the *mean net reward* earned when the project is worked on in state i is $R(i, 1) - \lambda$. The infinite-horizon λ-*charge one-armed restless bandit problem* is to find a policy π^* prescribing when to engage the project, which maximizes the expected total discounted net reward earned, where future rewards are geometrically discounted with factor $0 < \beta < 1$. In the infinite-

horizon version, on which we will focus, such a problem is formulated as

$$\max_{\pi\in\Pi} \mathsf{E}_{i_0}^{\pi}\left[\sum_{t=0}^{\infty}\Big\{R\big(X(t),a(t)\big)-\lambda\,a(t)\Big\}\beta^t\right]. \tag{3.1}$$

In formulation (3.1), Π denotes the class of *admissible policies*, among which an optimal policy is sought. Such a class consists of policies that make nonanticipative decisions, i.e., which base each action $a(t)$ on the project's state and action history $X(0),\ldots,X(t),a(0),\ldots,a(t-1)$. Further, $\mathsf{E}_{i^0}^{\pi}[\cdot]$ denotes expectation under policy π, conditional on the initial project state being equal to $X(0)=i^0$.

4. The multi-armed restless bandit problem

As for the term "multi-armed bandits," it is used as a metaphor for a *project portfolio*, the image being a collection of $N\geq 2$ one-armed bandits, of which the gambler is to choose at most $M\leq N$ to play at each time. Incorporating the project label $n=1,\ldots,N$ into the above single-project notation, we write, e.g., X_n, $R_n(i,a)$, $p_n(i,j|a)$, $X_n(t)$, and $a_n(t)$, with the obvious meaning. In such a setting, the project portfolio manager observes at the start of each period t the joint state $\mathbf{X}(t)=\big(X_n(t)\big)_{n=1}^N$, and takes a *joint action* $\mathbf{a}(t)=\big(a_n(t)\big)_{n=1}^N$, which must be based on the history of joint states and actions and satisfy $\sum_{n=1}^N a_n(t)\leq M$. The choice of action is based on adoption of a *scheduling policy* $\boldsymbol{\pi}$, which is to be taken from the resulting class $\boldsymbol{\Pi}(M)$ of *admissible scheduling policies*. The portfolio state's transition laws are determined by those of individual projects under the standard assumption that the latter are independent. As for the joint reward, it is assumed to be additive across projects. The infinite-horizon λ-*charge multi-armed restless bandit problem* (MARBP) is to find an admissible scheduling policy $\boldsymbol{\pi}^*$ prescribing which projects to engage at each time, if any, which maximizes the expected total discounted net reward earned. We formulate such a problem as

$$\max_{\boldsymbol{\pi}\in\boldsymbol{\Pi}(M)} \mathsf{E}_{\mathbf{i}^0}^{\boldsymbol{\pi}}\left[\sum_{t=0}^{\infty}\sum_{n=1}^{N}\Big\{R_n\big(X_n(t),a_n(t)\big)-\lambda\,a_n(t)\Big\}\beta^t\right], \tag{4.1}$$

where $\mathsf{E}_{\mathbf{i}^0}^{\boldsymbol{\pi}}[\cdot]$ denotes expectation under scheduling policy $\boldsymbol{\pi}$, conditional on the initial portfolio state being equal to $\mathbf{X}(0)=\mathbf{i}^0=(i_n^0)_{n=1}^N$.

5. A bit of bandit history

The classic multi-armed bandit problem has its roots in the area of *sequential design of experiments*, and in particular in the seminal works of Thompson [30], Robbins [29], and Bradt et al. [3]. Such works addressed, at increasing stages of development, the much-studied special case where engaging a project corre-

sponds to sampling from a Bernoulli population with unknown success probability, the goal being to find an optimal dynamic schedule for sampling from N such populations in a finite number of periods, where only one population can be sampled per period, to maximize the expected total number of successes. An MDP formulation is obtained via a Bayesian approach, where a project (population) state is given by the parameters of its posterior distribution.

A typical application in such early literature is the optimal dynamic allocation of patients to clinical treatments with unknown success probabilities. In such a setting, the decision maker is faced with a difficult dilemma between *exploitation* (i.e., using with the next patient the treatment that appears to be better based on evidence gathered so far) and *exploration* or *learning* (i.e., trying out a treatment that currently appears to be inferior, but which might turn out to be the best, to obtain a more accurate belief estimate of its efficacy based on the observed outcome). Ever since, the classic multi-armed bandit problem stands as a paradigmatic model for such an exploitation versus exploration dilemma.

Being MDPs, the bandit problems outlined above are in principle amenable to solution by the standard *dynamic programming* (DP) technique, i.e., by formulating and solving numerically the corresponding DP equations. Yet, in the one-armed case, the resulting numerical solution neither exploits nor provides insights on the problem's special structure. In the multi-armed case, the DP formulation is hindered by the so-called *curse of dimensionality*, as the number of DP equations grows exponentially in the number of projects. This fact renders computationally intractable a conventional numeric DP approach for problems of the dimensions that arise in real applications.

Such a state of affairs prompted researchers early on to develop solution approaches that take advantage of special structure. Thus, Bradt et al. [3] first realized that the optimal solution to the classic finite-horizon undiscounted one-armed Bernoulli bandit problem is given by an *index policy*: there exists a scalar index $\lambda^*(i, s)$, which is a function of both the current project state i and the number of remaining periods s, such that it is optimal to engage the project when it occupies state i and s periods remain iff $\lambda^*(i, s) \geq \lambda$ (recall that λ is the charge incurred per active period). Bellman [1] extended such a result to the infinite-horizon discounted one-armed Bernoulli bandit problem, establishing the existence of an index $\lambda^*(i)$, which is a function of the state only, such that it is optimal to engage the project in state i iff $\lambda^*(i) \geq \lambda$.

Concerning the classic multi-armed bandit problem, it was long considered intractable, to the point that, to quote from Whittle [35],

> "it was formulated during the war, and efforts to solve it so sapped the energies and minds of Allied analysts that the suggestion was made that the problem be dropped over Germany, as the ultimate

instrument of intellectual sabotage."

Yet, the optimal policy for the infinite-horizon discounted version of such a problem, where one bandit can be played at a time, turned out to be remarkably simple, being based on building blocks that had been laid out long ago. In a landmark achievement, Gittins and Jones [8] first showed that the index that had been introduced in Bellman [1] to solve the one-armed bandit problem is the key to the solution of the multi-armed problem, by using it as a *priority index*. Consider the case where the charge per play is $\lambda = 0$. If $\lambda_n^*(\cdot)$ is the index of project n, then it is optimal to engage at each time a project whose current state's index value is highest. For a nonzero charge λ where the option to idle all projects is allowed, it is optimal to engage at each time a project of currently highest index, among those projects, if any, whose current index value exceeds λ. Nowadays, the classic bandit index $\lambda_n^*(\cdot)$ is known as the *Gittins index*. See the discussion paper Gittins [6] and the monograph Gittins [7] for thorough accounts of the subject. A variety of different proofs, yielding complementary insights, has later been found of the Gittins and Jones result. See, e.g., Whittle [36], Varaiya et al. [31], Weber [33], and Bertsimas and Niño-Mora [2]. It must be pointed out that, independently of Gittins and Jones, Klimov [10] presented a highly influential analogous result in the setting of the optimal scheduling of a multiclass M/G/1 queue with Bernoulli feedback. See the review article Niño-Mora [26].

As for the MARBP, although finding its optimal solution is in general computationally intractable, in a seminal paper Whittle [37] deployed a Lagrangian relaxation and decomposition approach that yields a heuristic priority-index policy, along with an upper bound on the optimal problem value. Whittle further conjectured and Weber and Weiss [34] established under certain conditions that such a *Whittle index* policy enjoys a form of asymptotic optimality. Such results focus on restless bandits under the (long-run) average criterion, rather than under the discounted criterion.

Yet, the Whittle index has a limited scope, being only well-defined for the restricted class of so-called *indexable* restless bandits. To deploy the Whittle index policy in a particular model, one first needs to establish the *indexability* (i.e., existence of the index) of the underlying restless bandits. This can be done numerically on an instance by instance basis, although it is of course preferrable to establish indexability analytically for the model at hand, which typically involves imposing restricting conditions on model parameters. Such a state of affairs prompted Whittle [37] to write

> "One would very much like to have simple sufficient conditions for indexability; at the moment, none are known."

In a series of papers, the author has introduced and developed an approach that yields general sufficient indexability conditions. The first such conditions

for a finite-state restless bandit are presented in Niño-Mora [13], both under the discounted and the average criteria, being based on the framework of *partial conservation laws* (PCLs; see the review Niño-Mora [25]), also introduced there, along with an adaptive-greedy index-computing algorithm. Such results are extended in Niño-Mora [14] to restless bandits fed by a general resource, drawing on polyhedral methods of *linear programming* (LP), along the lines of the *polyhedral combinatorics* approach to combinatorial optimization. The scope of the PCL approach is expanded in Niño-Mora [17] to semi-Markov restless bandits on a finite or countably infinite state space, under discounted, average, and a new *mixed average-bias criterion*. The latter criterion overcomes limitations of the conventional average criterion, as certain relevant models that are not indexable relative to the average criterion are indexable relative to the average-bias criterion. A similar motivation prompted the author to introduce in Niño-Mora [15] indexability under the *bias criterion*. Such papers further clarify the economic interpretation of the Whittle index and the extensions introduced there, via the unifying concept of *marginal productivity* (MP) index. In short, a project's index $\lambda^*(i)$ measures the marginal value or productivity of engaging it when it occupies state i. The priority-index policy for a multi-armed problem allocates effort where it appears to be currently more productive, using the individual projects' MP indices as proxies of their current marginal productivities.

More recent work extends the scope of the PCL approach to indexability to the continuous state case. See Niño-Mora [22, 23], and Niño-Mora and Villar [24].

6. Applications

Besides the intrinsic mathematical interest of the challenging problems raised by research on restless bandits, the main driving force behind the fast-growing attention drawn by such a field is the realization by researchers that a wide variety of seemingly disparate applied problems fall within its scope. Formulating such problems as MARBPs allows investigators to deploy a unifying solution approach, which yields both an intuitively appealing and practical heuristic index policy of low complexity, and a bound on the optimal problem value, which can be used in practice to asses the policy's suboptimality gap. Although finding theoretical bounds on the latter remains an open research challenge, an increasing body of experimental evidence supports the view that, very often, the resulting index policies are nearly optimal and outperform previouly proposed index policies that had been devised via ad hoc arguments.

The growing list of applied stochastic models that have been addressed via restless bandit indexation includes the following: scheduling a multiclass make-to-stock queue (Veatch and Wein [32], Dusonchet and Hongler [4], Niño-Mora [17]); broadcast scheduling in information delivery systems (Raissi-Dehkordi and

Baras [28]); dynamic control of admission and routing to parallel queues (Niño-Mora [14, 19]); dynamic bandwidth allocation in a communication channel with delays (Ehsan and Liu [5], Niño-Mora [20]); dynamic routing of unmanned aerial vehicles with partial observations (Le Ny et al. [11]); bandits with switching costs (Niño-Mora [21]); dynamic scheduling of multiclass wireless transmissions (Goyal et al. [9], Niño-Mora [16]); dynamic scheduling of a multiclass queue with finite buffers (Niño-Mora [15]); opportunistic spectrum access (Niño-Mora [22, 23], Liu and Zhao [12]); multi-target tracking (Niño-Mora and Villar [24]); and finite-horizon indexation (Niño-Mora [27]).

7. Concluding remarks

The research field of multi-armed restless bandits is still in its early stages, and many challenging and relevant open problems remain to be solved. Among these one can mention the theoretical performance analysis of the priority-index policy, and the extension of the scope of restless bandit indexation to more complex models. New applications are awaiting to be successfully addressed via restless bandits.

Acknowledgments

This work has been supported in part by the Spanish Ministry of Education and Science via project MTM2007–63140.

References

[1] Bellman, R. (1956). A problem in the sequential design of experiments. *Sankhyā*, **16**, 221–229.

[2] Bertsimas, D. and Niño-Mora, J. (1996). Conservation laws, extended polymatroids and multiarmed bandit problems; a polyhedral approach to indexable systems. *Math. Oper. Res.*, **21**, 257–306.

[3] Bradt, R. N., Johnson, S. M., and Karlin, S. (1956). On sequential designs for maximizing the sum of n observations. *Ann. Math. Statist.*, **27**, 1060–1074.

[4] Dusonchet, F. and Hongler, M. -O. (2003). Continuous-time restless bandit and dynamic scheduling for make-to-stock production. *IEEE Trans. Robotics and Automat.*, **19**, 977–990.

[5] Ehsan, N. and Liu, M. (2004). On the optimality of an index policy for bandwidth allocation with delayed state observation and differentiated services. In: *Proc. INFOCOM 2004*, 1974–1983. IEEE.

[6] Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *J. Roy. Statist. Soc. Ser. B*, **41**, 148–177.

[7] Gittins, J. C. (1989). *Multi-armed Bandit Allocation Indices.* Wiley, Chichester (UK).

[8] Gittins, J. C. and Jones, D. M. (1974). A dynamic allocation index for the sequential design of experiments. In: Gani, J., Sarkadi, K., and Vincze, I., eds., *Progress in Statistics* (*European Meeting of Statisticians, 1972*), 241–266. North-Holland, Amsterdam (The Netherlands).

[9] Goyal, M., Kumar, A., and Sharma, V. (2006). A stochastic control approach for scheduling multimedia transmissions over a polled multiaccess fading channel. *Wireless Networks*, **12**, 605–621.

[10] Klimov, G. P. (1974). Time-sharing service systems. I. *Theory Probab. Appl.*, **19**, 532–551.

[11] Le Ny, J., Dahleh, M., and Feron, E. (2008). Multi-UAV dynamic routing with partial observations using restless bandit allocation indices. In: *Proc. 2008 American Control Conf.*, 4220–4225. IEEE.

[12] Liu, K. and Zhao, Q. (2008). A restless bandit formulation of opportunistic access: indexability and index policy. In: *Proc. 5th IEEE Conf. Sensor, Mesh and Ad Hoc Commun. Networks*, 1–5. IEEE.

[13] Niño-Mora, J. (2001). Restless bandits, partial conservation laws and indexability. *Adv. Appl. Probab.*, **33**, 76–98.

[14] Niño-Mora, J. (2002). Dynamic allocation indices for restless projects and queueing admission control: a polyhedral approach. *Math. Program.*, **93**, 361–413.

[15] Niño-Mora, J. (2006a). Marginal productivity index policies for scheduling a multiclass delay-/loss-sensitive queue. *Queueing Syst.*, **54**, 281–312.

[16] Niño-Mora, J. (2006b). Marginal productivity index policies for scheduling multiclass wireless transmissions. In: *Proc. NGI 2006, 2nd Euro-NGI Conf. Next Generation Internet Networks*, 342–349. IEEE.

[17] Niño-Mora, J. (2006c). Restless bandit marginal productivity indices, diminishing returns and optimal control of make-to-order/make-to-stock M/G/1 queues. *Math. Oper. Res.*, **31**, 50–84.

[18] Niño-Mora, J. (2007a). Dynamic priority allocation via restless bandit marginal productivity indices. *TOP*, **15**, 161–198. With discussion.

[19] Niño-Mora, J. (2007b). Marginal productivity index policies for admission control and routing to parallel multi-server loss queues with reneging. In: *Proc. NET-COOP 2007, 1st Euro-FGI Conf. Network Control and Optimization (Avignon, France)*, *Lect. Notes Comput. Sci.*, **4465**, 138–149. Springer, Berlin.

[20] Niño-Mora, J. (2007c). Marginal productivity index policies for scheduling multiclass delay-/loss-sensitive traffic with delayed state observation. In: *Proc. NGI 2007, 3rd Euro-NGI Conf. Next Generation Internet Networks*, 209–217. IEEE.

[21] Niño-Mora, J. (2008a). A faster index algorithm and a computational study for bandits with switching costs. *INFORMS J. Comput.*, **20**, 255–269.

[22] Niño-Mora, J. (2008b). An index policy for dynamic fading-channel allocation to heterogeneous mobile users with partial observations. In: *Proc. NGI 2008, 4th Euro-NGI Conf. Next Generation Internet Networks*, 231–238. IEEE.

[23] Niño-Mora, J. (2009). A restless bandit marginal productivity index for opportunistic spectrum access with sensing errors. In: *Proc. NET-COOP 2009, 3rd Euro-NG Conf. Network Control and Optimization*, Lect. Notes Comput. Sci., **5894**, 60–74. Springer, Berlin.

[24] Niño-Mora, J. and Villar, S. S. (2009). Multitarget tracking via restless bandit marginal productivity indices and Kalman filter in discrete time. In: *Proc. 2009 CDC/CCC, Joint 48th IEEE Conf. Decision and Control and 28th Chinese Control Conf.*, 2905–2910. IEEE.

[25] Niño-Mora, J. (2010a). Conservation laws and related applications. In: Cochran, J. J., ed., *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, New York. In press.

[26] Niño-Mora, J. (2010b). Klimov's model. In: Cochran, J. J., ed., *Wiley Encyclopedia of Operations Research and Management Science*. Wiley, New York. In press.

[27] Niño-Mora, J. (2010c). Computing a classic index for finite-horizon bandits. *INFORMS J. Comput.*, forthcoming.

[28] Raissi-Dehkordi, M. and Baras, J. S. (2002). Broadcast scheduling in information delivery systems. In: *Proc. GLOBECOM '02, Global Telecomm. Conf.*, 2935–2939. IEEE.

[29] Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, **58**, 527–535.

[30] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, **25**, 275–294.

[31] Varaiya, P. P., Walrand, J. C., and Buyukkoc, C. (1985). Extensions of the multiarmed bandit problem: the discounted case. *IEEE Trans. Automat. Control*, **30**, 426–439.

[32] Veatch, M. H. and Wein, L. M. (1996). Scheduling a multiclass make-to-stock queue: index policies and hedging points. *Oper. Res.*, **44**, 634–647.

[33] Weber, R. (1992). On the Gittins index for multiarmed bandits. *Ann. Appl. Probab.*, **2**,1024–1033.

[34] Weber, R. R. and Weiss, G. (1990). On an index policy for restless bandits. *J. Appl. Probab.*, **27**, 637–648.

[35] Whittle, P. (1979). Discussion on: "Bandit processes and dynamic allocation indices" [J. R. Statist. Soc. B, **41**, 148–177, 1979] by J. C. Gittins. *J. Roy. Statist. Soc. Ser. B*, **41**, 165.

[36] Whittle, P. (1980). Multi-armed bandits and the Gittins index. *J. Roy. Statist. Soc. Ser. B*, **42**, 143–149.

[37] Whittle, P. (1988). Restless bandits: Activity allocation in a changing world. In Gani, J., editor, *A Celebration of Applied Probability*, *J. Appl. Probab.*, **25A**, 287–298. Applied Probability Trust, Sheffield (UK).

About the author

José Niño-Mora is associate professor (*profesor titular*) at Carlos III University of Madrid (UC3M). He earned his PhD (1995) at the Massachusetts Institute of Technology (MIT) on a MEC/Fulbright fellowship. After stints at MIT (postdoc, 1995/96), the catholic University of Louvain at Louvain-la-Neuve (*CORE* and *Marie Curie* fellow, 1996/97), and Barcelona's Pompeu Fabra University (faculty member, 1997–2003), he joined the faculty of UC3M (2003) as a *Ramón y Cajal* researcher. His main research interests are in stochastic models of operations research, spanning theory, algorithms, and applications.