Bandits with non-binary action spaces

With applications to dowblink power control in wireless networks

May 20, 2022

Contents

1 Problem formulation

Consider one base station (BS) that serves N mobile devices. Time is slotted with slot-duration of Δ. The BS maintains a separate queue of data to be sent to each device. Data arrives to queue i in form of packets. The distribution of the number of packets arriving in a slot to queue i is assumed to be Poisson of rate $\lambda_i\Delta$. The channel condition of device i fluctuates over time and is modelled by the stochastic process $G_i(t)$ (here G stands for gain).

When queue i is served at rate r, an amount $r\Delta$ is removed from queue i. Assuming that packet sizes are exponentially distributed with rate μ_i, this implies that the probability a packet finishes transmission when served at rate r is $r\Delta$ for Δ sufficiently small. Also, we shall assume that packets arriving in a slot cannot be served in that slot

The BS has a total power budget of P_m in each slot that it can distribute over the devices. The service rate of a queue depends upon the power allocation as follows. If the BS allocates power P to a device when its gain is G, then the service rate of the queue of this device is:

$$R = G \cdot P.$$

For simplicity, the relation between service rate and the power is assumed to be linear. In practice, this relation can be more complicated.

Let $X_i(t)$ be the number of packets in the queue of device i in time-slot t, and let $P_i(t)$ be the power allocated by the BS to this device. Then, $X_i(t)$ is a discrete-time Markov chain with transitions

$$X_i(t+1) = X_i(t) + A_i(t) - S_i(t), \tag{1.1}$$

where $A_i(t) \sim \text{Poi}(\lambda_i\Delta)$ and $S_i(t)$ is a Bernoulli random variable with success probability $G_i(t)P_i(t)\Delta$. Assume $G_i(t)$ is also a discrete-time Markov chain (DTMC) whose transition matrix is known.

The state variable is the vector $[(X_i(t), G_i(t))]_{i=1}^N$ which is a DTMC when the power allocation is fixed over the sample path.

The BS has to decide how much power to allocate to each devices so as to minimize the mean sojourn time of tasks. The problem will be referred to as *downlink power control*, and can be formally written as an average-cost MDP:

$$\min \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} X_i(t) \tag{PC}$$

s.t.

$$\sum_{i=1}^{N} P_i(t) \le P_m, \ \forall t \tag{1.2}$$

The number of states in the states-space of (PC) grows exponentially in the number of devices. Computing optimal policies numerically in such cases is extremely challenging and practically infeasible.

The objective of this work is to find good heuristic policies for the (PC) similar to Whittle Index Policies (WIP). Bounds and asymptotic optimality type results will also be explored.

1.1 Fixed powers and fixed channel gains

As a first step, consider the case when there is no power control. i.e., the power allocation is fixed and is not part of the action space. Further, the channel gains

also are fixed. And finally, only one device can be selected in every slot. Then, (PC) simplifies to

$$\min \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} X_i(t) \tag{FP-FG}$$

s.t.

$$\sum_{i=1}^{N} a_i(t) \leq 1, \ \forall t, \tag{1.3}$$

$$a_i(t) \in \{0,1\}, \ \forall i,t, \tag{1.4}$$

where $a_i(t) = 1$ means device i is served in time-slot t. In the dynamics of the queue, only the success probability of $S_i(t)$ gets modified to $a_i(t)G_i(t)P_m\Delta$.

A popular heuristic for solving (FG) approximately is the Whittle index policy [**?**]. The basic idea is to relax the constraint (2.1) and analyse the dynamics of each individual device separately. That is, we find the optimal policy for the following MDP for device i:

$$\min \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left(X_i(t) - \lambda a_i(t)\right) \tag{FP-FG-IMDP}$$

s.t.

$$a_i(t) \in \{0,1\}, \ \forall i,t. \tag{1.5}$$

Here, the coupling between the devices is through the Lagrange multiplier λ. Observe that in any state, that is the value of X_i, if $\lambda = \infty$ then it is optimal to be active whereas if $\lambda = -\infty$, then it optimal to be inactive.

Now, suppose that, for every state x, as λ increases from $-\infty$ to ∞, there exists a unique $\lambda^*(x)$ such that for $\lambda \leq \lambda^*(x)$ it is optimal to be inactive and $\lambda \geq \lambda(x)$ it is optimal to be active, $\lambda^*(x)$ is called the Whittle index of x. The WIP is then to select the device with the highest Whittle index.

1.2 Roadmap for the internship

WIPs are mainly applied to problems in which the individual MDPs have a single dimensional state-space and the action set is binary as was the case in (FP-FG-IMDP). Our objectives are

1. Assume the channel gains are fixed but that power allocation is a control variable. This makes the action space non-binary. The objective is to conceive heuristics for this multi-action case.

2. The next step is to allow the gains to be dynamic. The state $(X_i(t), G_i(t))$ is now multidimensional. This problem will be subdivided into two:

   (a) Power allocation are fixed, i.e., the binary action case;

   (b) Power allocations are variable, i.e., the multi-action case.

2 Power control with fixed channel gains

Assgume $G_i(t) = g_i$, $\forall t$, and that $P_m = 1$ Then, (PC) simplifies to

$$\min \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} X_i(t) \tag{FG}$$

s.t.

$$\sum_{i=1}^{N} P_i(t) \leq 1, \ \forall t, \tag{2.1}$$

$$P_i(t) \in [0,1], \ \forall i, t. \tag{2.2}$$

The success probability of $S_i(t)$ gets modified to $g_i P_i(t)\Delta$.

Assume first that power levels are discrete, i.e., $P_i(t) \in \{0, p_1, p_2, \ldots, P_m\}$. We shall follow the approach of [XLS21] who treat the multi-action case but for the finite time-horizon problem. Their main idea also starts by relaxing the constraint (2.1) just as in Whittle's method. However, the next step is what makes it different. Instead of decompsing (FG) into individual MDPs, they define a linear program (LP) with the occupation measures $\mu_n(s, a; t)$ for individual bandits[1] as the control variables.

The occupation measure is the probability of taking action a in state s at time t for bandit n for a given policy π.

The relaxed LP for (FG) for the finite-time horizon problem becomes

$$\min \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{(s_i,a_i)} X_i(t)\mu_i(s_i, p_i; t) \tag{OM-RLP}$$

s.t.

$$\sum_{i=1}^{N} \sum_{s_i} p_i \mu_i(s_i, p_i; t) \leq 1, \ \forall t, \tag{2.3}$$

$$\sum_{p_i} \mu_i(s_i, p_i; t) = \sum_{(s_i', p_i')} \mu_i(s_i', p_i'; t-1)\mathbb{P}(s_i', p_i', s_i), \ \forall t, n, s_i, \tag{2.4}$$

$$\sum_{p_i} \mu_i(s_i, p_i, 1) = \alpha_i(s_i), \ \forall s_i, \tag{2.5}$$

with (2.4) ensuring that the occupation measures respect the transition matrices of the individual bandits, and (2.5) imposes that the occupation measures be initialized at time 1 according to a given distribution α_i.

2.1 OMR index policy

In [XLS21], they propose an index policy derived from the solution of (OM-RLP) which is denote by μ_i^*. Define

$$\chi_i(s_i, p_i; t) = \frac{\mu_i^*(s_i, p_i, t)}{\sum_p \mu_i^*(s_i, p, t)}, \tag{2.6}$$

to be the probability of taking action p_i in state s_i of bandit i. Recall that in terms of our power allocation problem, this implies that device i is allocation power p_i in

[1] in our problem, bandits are the devices

time-slot t with probability $\chi_i(s_i, p_i; t)$ when its state (or the queue length) is s_i. If $\sum_p \chi_i(s, p; t) = 0$ for some state s, then $\chi_i(s_i, p_i; t) = 0$ for all p_i in this state for bandit i.

Directly choosing the power levels using the rule (2.6) may not result in an allocation that is compatible with (2.1) in every slot t. To ensure that this constraint is indeed satisfied they define the OMR index as

$$\psi_i(s_i; t) = \sum_{p_i \neq 0} \chi_i(s_i, p_i; t) p_i, \tag{OMRI}$$

which is the average reward if bandit i is activated in state s_i in slot t.

The OMR Index Policy proposed in [XLS21] works as follows. Arrange the bandits in decreasing order of the OMR index (OMRI). Starting from the first bandit, for bandit i choose action a_i with probability (2.6) and move onto the next bandit as long as the constraint (2.1) is satisfied.

2.2 Tasks for the internship

There are several questions that are worth investigating.

1. Can we determine similar heuristics from from solving individual relaxed LPs as is done in WIP? From a computational complexity point of view, solving individual LPs will be much more cheaper when N is large. Some property like asymptotic optimality will be needed as a justification.

2. How to obtain an OMRIP-like heuristic for the average-reward case? I think there is paper of Lasserre on computing the occupation measures for average-cost MDPs. Could that result be useful here?

3. Is there a better way of choosing the actions from the occupation measures?

4. Learning dynamics in the spirit of Section 3.4 in [XLS21] individual LPs.

2.3 Task 1: Individual LPs

The first task is to see how good a heuristic can be obtained from solving the occupation measures for the relaxed individual LPs. Numerical experiments will have to run to compare the quality of individual and global OMRIPs.

3 Heuristics for (PC)

This problem will be divided into two subproblems.

1. gains are independent and identically distributed from one slot to the other.
2. gains follow a DTMC. Can we do time-scale separation assuming gains change slowly compared to the queue-length and use results from above?

In principle, the approach of [XLS21] covers this case as well.

References

[XLS21] Guojun Xiong, Jian Li, and Rahul Singh. Reinforcement learning for finite-horizon restless multi-armed multi-action bandits. *CoRR*, abs/2109.09855, 2021.