Non-monotonic Resource Utilization in the Bandits with Knapsacks Problem

Raunak Kumar (Cornell), Robert D. Kleinberg (Cornell)

NEURAL INFORMATION PROCESSING SYSTEMS

Model

We introduce a natural generalization of the (stochastic) **bandits with knapsacks** (BwK) model [1] by allowing **non-monotonic resource utilization**. This captures **resource renewal** in many applications of BwK, such as dynamic pricing.

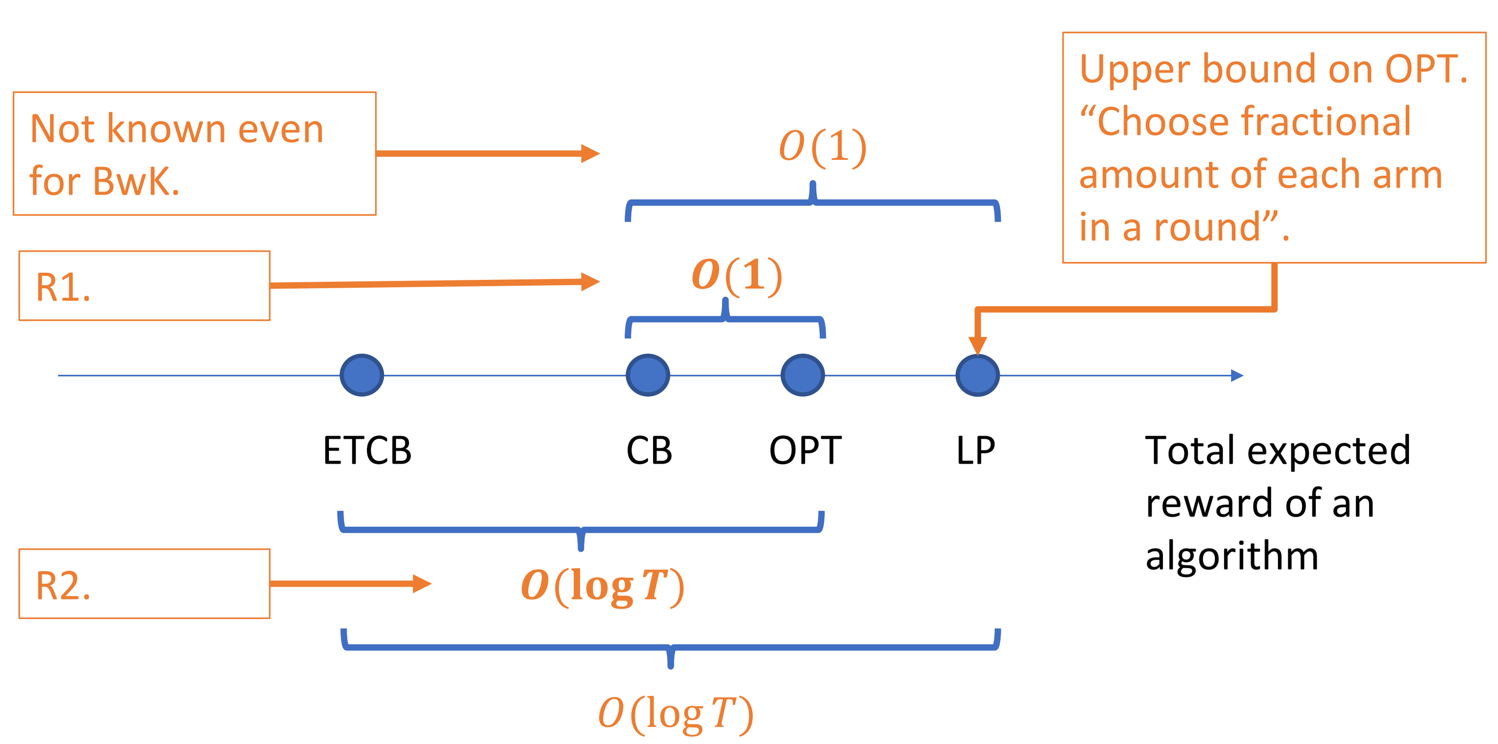
There are $\boldsymbol{k}$ **arms** and $\boldsymbol{m}$ **resources** each with initial budget B. In each round $t \in [T]$, if the budget of any resource is < 1, the algorithm must choose arm x^0 ("null arm"); otherwise, it may choose any arm. It observes an outcome sampled from the chosen arm's outcome distribution. The **outcome consists of a reward r_t and a drift $\mathbf{d}_{t,j} \in [-1,1]$** for each resource j. The budget of each resource is incremented by its drift: $B_{t,j} = B_{t-1,j} + d_{t,j}$.

In this poster we focus on the special case of one resource. The paper deals with the general case of multiple resources.

Assumptions:
1. Null arm has zero reward, non-negative drift, and positive expected drift.
2. All arms have non-zero expected drifts.
3. Solution to the LP relaxation is unique.

Results

1. (**R1**) If we **know the true outcome distributions**, we design a policy, *ControlBudget* (CB), that has $O(1)$ instance-dependent regret with respect to OPT (total expected reward of the optimal solution).
2. (**R2**) If we **don't know the true outcome distributions**, we design a learning algorithm, *ExploreThenControlBudget* (ETCB), that has $O(\log T)$ instance-dependent regret with respect to OPT.



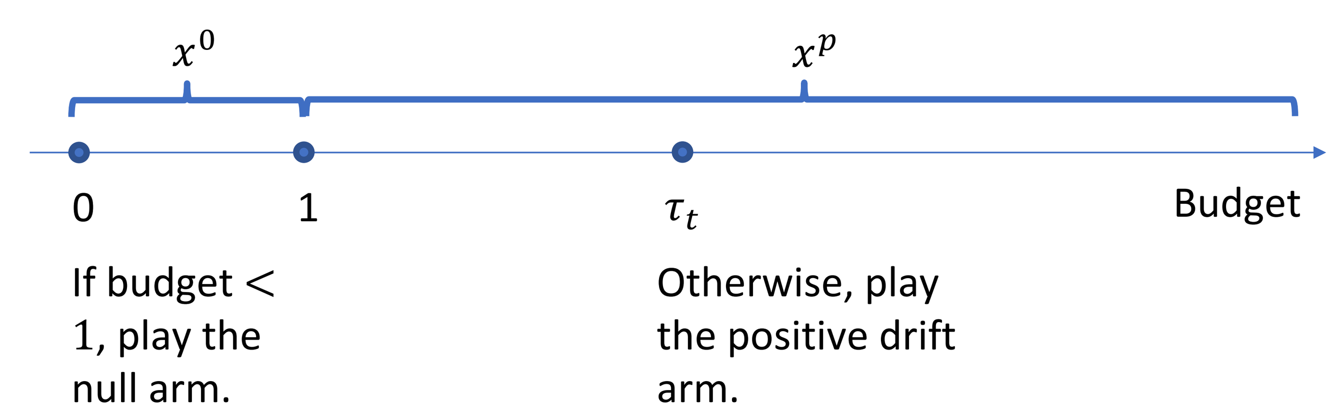
Learning Algorithm *ExploreThenControlBudget*

Learning algorithm, ETCB, proceeds in **two phases**: (1) explore in a round-robin fashion to find arms in the LP solution; (2) play the policy CB.

Policy *ControlBudget*

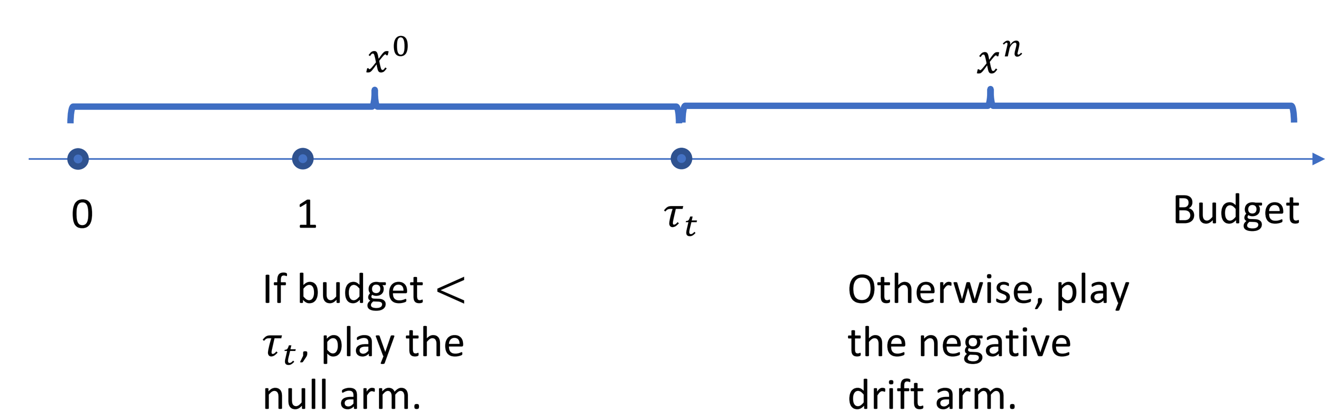
In the special case of one resource, the **solution** to the **LP** relaxation is **one of three types**:
1. One arm with positive drift.
2. Two arms: one negative drift and the null arm.
3. Two arms: one negative drift and one positive drift.

Case 1: Positive drift arm x^p.



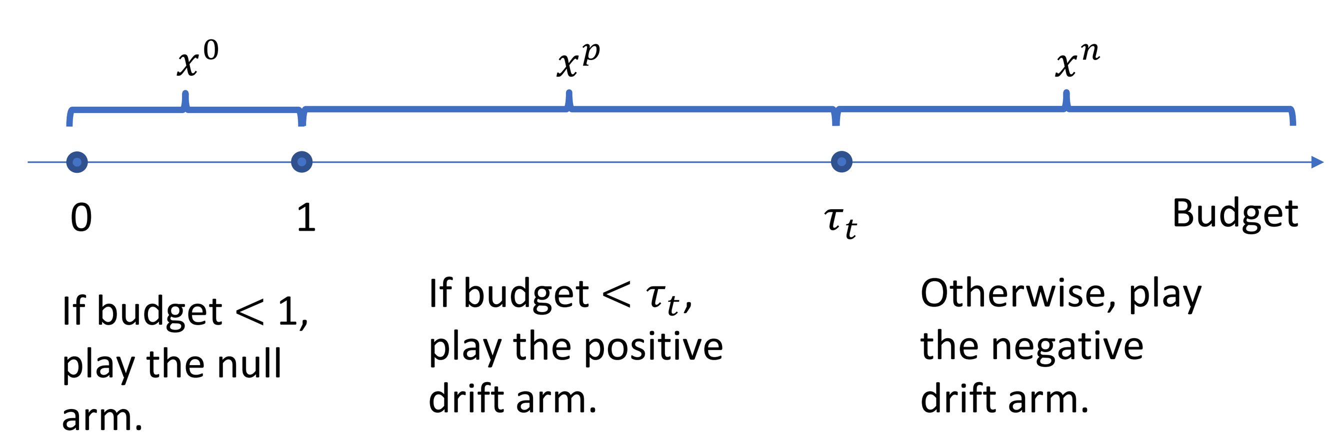
Case 2: Negative drift arm x^n and null arm x^0.

Threshold $\tau_t = c\log(T-t)$. c is a constant.



Case 3: Negative drift arm x^n and positive drift arm x^p.

Threshold $\tau_t = c\log(T-t)$. c is a constant.



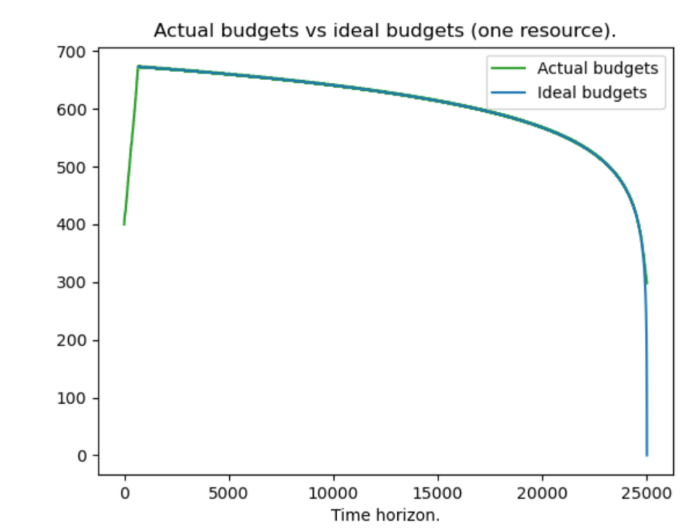
Regret Analysis Sketch

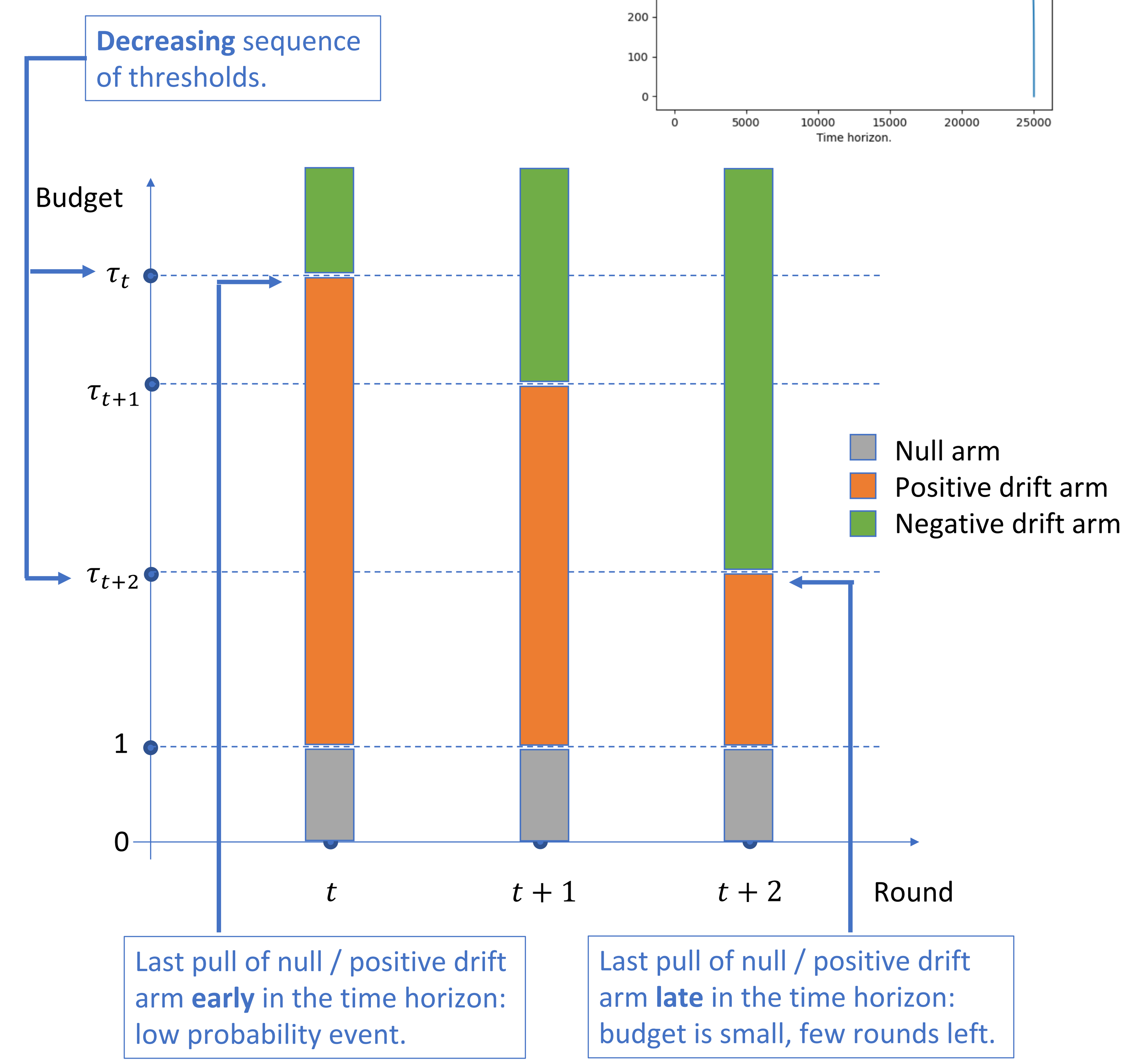
Case 1: Regret = expected number of null arm pulls. This is equal to number of visits to $[0,1)$ by a positive drift random walk, which is a constant.
Case 2: Similar to case 3.
Case 3: By properties of LPs, the regret is at most the sum of the expected number of null arm pulls and the expected leftover budget.

Lemma: Expected number of pulls of the null arm is a constant.

Budget < 1 if pulling x^p decreases budget: low probability event.

Lemma: Expected leftover budget is a constant.





References

[1] Badanidiyuru et al. (2018). "Bandits with Knapsacks." In: Journal of the ACM.

[2] Flajolet and Jaillet (2015). "Logarithmic Regret Bounds for Bandits with Knapsacks." In: arXiv.

[3] Li et al. (2021). "The Symmetry between Arms and Knapsacks: A Primal-Dual Approach for Bandits with Knapsacks." In: ICML.