Proceedings of Machine Learning Research vol 65:1–2, 2017

Sparse Stochastic Bandits

Joon Kwon JOON.KWON@ENS-LYON.ORG
CMAP, École polytechnique, Université Paris–Saclay

Vianney Perchet VIANNEY.PERCHET@NORMALESUP.ORG
CMLA, École Normale Supérieure Paris–Saclay
& Criteo Research, Paris

Claire Vernade CLAIRE.VERNADE@TELECOM-PARISTECH.FR
LTCI, Télécom ParisTech

Abstract

In the classical multi-armed bandit problem, d arms are available to the decision maker who pulls them sequentially in order to maximize his cumulative reward. Guarantees can be obtained on a relative quantity called regret, which scales linearly with d (or with $\sqrt{d}$ in the minimax sense). We here consider the *sparse case* of this classical problem in the sense that only a small number of arms, namely $s < d$, have a *positive* expected reward. We are able to leverage this additional assumption to provide an algorithm whose regret scales with s instead of d. Moreover, we prove that this algorithm is optimal by providing a matching lower bound – at least for a wide and pertinent range of parameters that we determine – and by evaluating its performance on simulated data.

Keywords: stochastic multi-armed bandit problem, regret, sparsity, UCB

We consider the classical stochastic multi-armed bandit problem with d "arms". Pulling arm $i \in [d] := \{1, ..., d\}$ at time t yields a reward $X_i(t) \in [-1, 1]$, the sequence $(X_i(t))_{t\geqslant 1}$ being assumed to be *i.i.d* and of expectation μ_i. This problem is well understood, and there exist algorithms minimizing the regret such that

$$\mathrm{Reg}(T) \lesssim \sum_{\substack{i\in[d]\\ \Delta_i>0}} \frac{\log(T)}{\Delta_i}, \quad \text{where } \Delta_i = \max_j \mu_j - \mu_i \,,$$

$\mathrm{Reg}(T)$ denotes the expected regret after T rounds, and $\lesssim$ indicates that the inequality holds up to some universal multiplicative or additive constants. We consider the *sparse* bandit problem where exactly $s > 1$ expectations are positive (wlog, we assume that they correspond to the first s indices of arms). We construct an anytime algorithm that leverages this a-priori knowledge to lower the linear dependency in d to s. Indeed, it guarantees

$$\mathrm{Reg}(T) \lesssim \sum_{\substack{i\in[s]\\ \Delta_i>0}} \Big(\frac{\log(T)}{\Delta_i} + \frac{\Delta_i \log(T)}{\mu_i^2}\Big).$$

We also prove that this algorithm is optimal, at least for a wide and pertinent range of parameters, by deriving an asymptotic matching lower bound.

. Extended abstract. Full version appears as [arXiv:1706.01383v1]

For instance, in the specific case where $\mu_1 = 1$ and for $2 \leqslant i \leqslant s$, $\mu_i = \mu_1 - \Delta := \mu \geqslant 1/2$ (and $\mu_i = 0$ for $i > s$), the guarantee of our algorithm boils down to

$$\mathrm{Reg}(T) \lesssim \max\left\{\frac{s\log(T)}{\Delta},\ \frac{s\Delta\log(T)}{\mu^2}\right\} = \frac{s\log(T)}{\Delta}.$$

On the other hand, our asymptotic, problem-dependent lower bound shows that the above performance is tight up to constant terms, as soon as $s \leqslant d/3$ since

$$\liminf_{T\to+\infty} \frac{\mathrm{Reg}(T)}{\log(T)} \geqslant \max\left\{\frac{s}{2\Delta},\ \frac{s\Delta}{2\mu^2}\right\} = \frac{s}{2\Delta}.$$

Acknowledgments

J. Kwon was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH. V. Perchet has benefitted from the support of the ANR (grant ANR-13-JS01-0004-01), of the *FMJH Program Gaspard Monge in optimization and operations research* (supported in part by EDF) and from the Labex LMH. C. Vernade was also partially supported by the Machine Learning for Big Data Chair at Télécom ParisTech.