JMLR: Workshop and Conference Proceedings vol 40:1–2, 2015

Batched Bandit Problems

Vianney Perchet VIANNEY.PERCHET@NORMALESUP.ORG
Université Paris Diderot – INRIA

Philippe Rigollet RIGOLLET@MATH.MIT.EDU
Massachusetts Institute of Technology

Sylvain Chassang CHASSANG@PRINCETON.EDU
Princeton University

Erik Snowberg SNOWBERG@CALTECH.EDU
California Institute of Technology

Abstract

Motivated by practical applications, chiefly clinical trials, we study the regret achievable for stochastic multi-armed bandits under the constraint that the employed policy must split trials into a small number of batches. Our results show that a very small number of batches gives already close to minimax optimal regret bounds and we also evaluate the number of trials in each batch. As a byproduct, we derive optimal policies with low switching cost for stochastic bandits.

In practice, fixed costs and delays in the observation of outcomes make it prohibitively expensive to run clinical trials consisting of more than three or four batches of patients. The objective is to describe the regret achievable for two-armed bandits under the constraint of a small number M of batches, within which the likelihood of pulling each arm is constant. We study a class of *explore-then-commit (*ETC*) policies* policies parameterized by the partition of patients across batches. In the first batch, the policy randomizes uniformly between arms. At the end of each batch, a statistical test of performance is implemented. If it is conclusive, the supposed-to-be suboptimal arm is eliminated. If it is inconclusive, the policy keeps alternating between arms in the next batch.

For each batch size M, we describe ETC policies π^1 and π^2 achieving tight adaptive and minimax bounds on regret, where Δ is the gap in expected returns between arms and T the horizon:

$$
\begin{aligned}
R_T(\pi^1) &\lesssim \left(\frac{T}{\log(T)}\right)^{\frac{1}{M}} \frac{\overline{\log}(T\Delta^2)}{\Delta} \\
R_T(\pi^2) &\lesssim T^{\frac{1}{2-2^{1-M}}} \log^{\alpha_M}\left(T^{\frac{1}{2^M-1}}\right), \qquad \alpha_M \in [0, 1/4)\,.
\end{aligned}
$$

Thus losses from using few batches are small: π_1 attains the optimal adaptive rate $\log(T\Delta^2)/\Delta$ if $M = \Theta(\log(T/\log(T)))$; π^2 attains the optimal minimax rate $\sqrt{T}$ whenever $M = \Theta(\log\log T)$.

Tests on real and simulated data show that batch-optimized ETC policies with few batches perform well, even relative to more responsive strategies such as UCB2.

Acknowledments

Philippe Rigollet acknowledges the support of NSF grants DMS-1317308, CAREER-DMS-1053987, the Howard B. Wentz Jr. Junior Faculty award and the Meimaris family. Vianney Perchet received support from the French National Research Agency (ANR) Project GAGA: ANR-13-JS01-0004-01. Sylvain Chassang and Erik Snowberg acknowledge support from NSF grant SES-1156154.