Partial Monitoring with Experts

Wouter M. Koolen

June 11, 2024

Supervisor

Prof.dr. Wouter M. Koolen
Department of Statistics
University of Twente

Keywords

Sequential Decision Making, Experts, Bandits, Partial Monitoring

Background The field of Online Learning studies sequential interactions between a learner and their environment. The learner sequentially chooses actions, and the environment determines the quality of each of the actions. The research goal is to understand how difficult such learning problems are, and to design efficient algorithms for the learner. For discussing the project, we need to elaborate on two dimensions of complexity in such problems:

1. The type of feedback. Available settings are "Full Information" where we see the quality of *all* actions, "Bandit Information" where we see the quality of the action chosen by the learner, and "Partial Monitoring" where the feedback follows a more general known structure.

2. The comparator class. The vanilla goal is to compete with the best fixed action. A more challenging yet structured problem is obtained when a class of policies (aka experts) is available. Now the goal is to compete with the best expert.

To illustrate these, let's instantiate the rates for the adversarial Bandit setting. Let's say we have K actions and M experts, and we play for T rounds. Then the EXP3 algorithm learns the overall best action at regret overhead $\sqrt{KT \ln K}$. The EXP4 algorithm by (Auer et al., 2002) learns the overall best expert at regret overhead $\sqrt{KT \ln M}$. These results are meaningful when the regret is $\ll T$. We see that we can successfully accommodate many experts (M occurs only logarithmically) as long as the number of actions is relatively small (these manifest as a factor $\sqrt{K}$).

Academic Content Partial monitoring studies feedback in general, extending both full and bandit information. This feedback model capture many practical problems, including Apple Tasting and auctions (Lattimore and Szepesvari, 2019, Chapter 37). Much is known for finite sets of actions (without experts). In particular, (Bartók et al., 2014) characterise the four possible exponents (on T) in the worst-case regret, and classify all feedback structures.

The thrust of this project is to develop theory and algorithms for partial monitoring with experts. We will revisit loss estimation procedures based on importance weighting (Auer et al., 2002), and we will take inspiration from the NeighborhoodWatch2 algorithm (Lattimore and Szepesvari, 2019, Chapter 37).

The project is envisaged to consist of mostly theoretical work, with only minor computational and empirical components.

References

Auer, P., N. Cesa-Bianchi, Y. Freund, and R. Schapire (2002). "The nonstochastic multiarmed bandit problem". In: *SIAM Journal of Computing* 32(1), pp. 48–77.

Bartók, G., D. P. Foster, D. Pál, A. Rakhlin, and C. Szepesvári (2014). "Partial monitoring – classification, regret bounds, and algorithms". In: *Mathematics of Operations Research* 4, pp. 967–997.

Lattimore, T. and C. Szepesvari (2019). *Bandit Algorithms*. Cambridge University Press.