UNIVERSITY OF OXFORD

Machine Learning
Michaelmas Term 2017
Week 5

Practical 3 : Multi-armed Bandits

Note *You only need to do this practical if you wish to submit assessed work based on this practical. If you submit assessed work based on dimensionality reduction or clustering; then you may skip this practical. However, I strongly recommend you to at least try running simulations for Uniform Exploration, ϵ-Greedy, UCB1 and Successive Elimination. If you intend to submit assessed work based on this practical, then you should either do both Task 1 and 2, or (just) Task 3.*

Task 1

You will run simulations to understand the performance of the various multi-armed bandit algorithms. You may assume for simplicity that all reward distributions are Bernoulli. Thus, to construct an instance you need to choose the mean rewards for each of the arms. I recommend starting with a small number of arms, e.g. $K = 10$ and focusing on the dependence on T rather than the dependence on number of bandit arms.

After you've tried setting mean rewards randomly, try choosing the mean rewards that are used in the lower-bound construction. Even though the multiplicative constants in the theoretical bounds are loose, if you plot the logartihm of cumulative regret at time t vs t, then you should (eventually) see a slope of $1/2$.

Task 2

See how Thompson sampling performs compared to the various algorithms used in Task 1.

To generate "adversarial data", use the following simple model. Start with two different set of mean rewards, say $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$. You will initially draw rewards according to one of them, say $\boldsymbol{\mu}$, so reward of arm a will have mean μ_a. This can continue for several rounds. However at each round, with some probability, p, you switch to rewards according to $\boldsymbol{\mu}'$, and then in this new "environment" the mean reward for arm a will be μ'_a. After some rounds (again with probability p), you may switch back to rewards using $\boldsymbol{\mu}$. Thus, the environment is no longer stationary.

Run simulations to see how the algorithms designed for stationary environments work in this setting and compare that to the performance of algorithms designed for an adversarial setting.

You may have to fiddle around with the various parameters to see interesting behaviour.

Task 3

Have a look at one of the following papers (or pick your own bandit-related paper, if you have something in mind) and try to reproduce the simulations in the paper. You probably want to first work with generated data before trying to experiment with datasets used in some of these papers.

Note that you will have to spend some time reading parts of the papers to understand precisely what the experimental setup is. The papers also propose different algorithms; you should understand the algorithms themselves, though I recommend skipping proofs in the first instance.

The papers below are *not* listed in any particular order of preference.

- An Empirical Evaluation of Thompson Sampling
- One Practical Algorithm for Both Stochastic and Adversarial Bandits
- Multi-Armed Bandit Algorithms and Empirical Evaluation
- Thompson Sampling: An Asymptotically Optimal Finite Time Analysis