

RL and Planning Under Uncertainty (ANU, Sem2, 2008)
Lab2: Bandit Algorithms and Poker
Lab Instructor: Scott Sanner

READ THE SUTTON AND BARTO BOOK, CHAPTER 5!!!

Agenda:

1. Play some SimplePoker: `java game.poker.SimplePokerDisplay` (like a one-round version of Texas Hold'em)
2. How to encode with stateless bandits? Two bandits per state (state = pair of visible cards). Player code creates *bet* and *fold* bandits on first visit to a state.
3. Run: `java game.poker.SimplePoker`. Plays a number of games with a given player, reports cumulative reward over fixed interval of trials.

Student lab work:

1. Look at `SimplePoker.main`. Try out the different players. `Player.AvgPlayer` performs poorly because only exploits, does not explore. `Player.NoisyPlayer` forces exploration, but have to get noise parameter right.
2. Implement `Player.BanditUCBPlayer` using UCB algorithm (see One-shot decision-making slides on web). No parameters to tune... how does it work?

For further reading:

===

Finite-time Analysis of the Multiarmed Bandit Problem

<http://homes.dsi.unimi.it/~cesabian/Pubblicazioni/ml-02.pdf>

The seminal paper on *uniform log regret bounds for bandits with upper confidence bound (UCB) algorithm.*

===

Multi-armed Bandits, Dynamic Environments, and Meta-bandits

<http://www.lri.fr/~nbaskiot/papier/MetaEve.pdf>

Among other ideas, defines a second level of bandits (meta-bandits) to model the exploration / exploitation tradeoff in *non-stationary* environments.

===

Multi-armed Bandit Problems with Dependent Arms

<http://www.machinelearning.org/proceedings/icml2007/papers/388.pdf>

Discusses the basic idea of handling dependent arms by clustering arms and using a two-level bandit policy (one for clusters, one for individual arms). In later work, extended to taxonomies of bandits (not just simple two-level hierarchy).

===

Experience-efficient Learning in Associative Bandit Problems

http://www.icml2006.org/icml_documents/camera-ready/112_Experience_Efficient.pdf

Talks about the associative bandit problem. One gets some additional information (e.g., user features and ad features) before deciding on an arm to pull. The problem is then to associate inputs with which arm to pull. Analyzes efficient approaches to the problem.

===

Bobby Kleinberg's Publications

http://www.cs.cornell.edu/~rdk/pubs_topic.html

Many interesting publications on bandit problems with large numbers of arms.

Bobby's PHD THESIS ABSTRACT: This thesis concerns an important class of online decision problems called generalized multi-armed bandit problems. Most existing algorithms for such problems were efficient only in the case of a small (i.e. polynomial-sized) strategy set. We extend the theory by supplying non-trivial algorithms and lower bounds for cases in which the strategy set is much larger (exponential or infinite) and the cost function class is structured, e.g. by constraining the cost functions to be linear or convex. As applications, we consider adaptive routing in networks, adaptive pricing in electronic markets, and collaborative decision-making by untrusting peers in a dynamic environment.

===

Contextual Recommender Problems

<http://storm.cis.fordham.edu/~gweiss/ubdm05/UBDM-Madani-14.pdf>

Discusses the basic idea of factoring the feature space for users and the feature space for ads, but provides no explicit algorithm. Nonetheless, a simple and good idea.