# Reinforcement Learning Notes

## Gradient Bandit Algorithms

We have considered methods that estimate action values and use those estimates to select actions. This is often a good approach, but it is not the only one possible. In this section we consider learning a numerical *preference* for each action , which we denote . The larger the preference, the more often that action is taken, but the preference has no interpretation in terms of reward. Only the relative preference of one action over another is important; if we add 1000 to all the action preferences there is no effect on the action probabilities, which are determined according to a ***soft-max distribution*** (i.e., Gibbs or Boltzmann distribution) as follows:

(1)

Where we have introduced a new distribution function , for the probability of taking action at time . Initially, all action preferences are the same (e.g., ) so that all actions have an equal probability of being selected.

Let’s consider the case of two actions and ; . Then

and

which become , where . Here is is the sigmoid function. Then .

There is a natural learning algorithm for soft-max action preferences based on the idea of stochastic gradient descent. On each step, after selecting action and receiving the reward the action preferences are updated by:

, and

(2)

where is a step-size parameter, and is the average of the rewards up to but not including time t (with ), which can be computed incrementally similarly to the incremental update procedure for the *Multi-arm Bandit* problem. The term serves as baseline with which the reward is compared. If the reward is higher than the baseline, then the probability of taking in the future has increased, and if the reward is below baseline, then the probability has decreased. The non-selected actions move in the opposite direction.

## The Bandit Gradient Algorithm as Stochastic Gradient Ascent

One can get a deeper insight into the gradient bandit algorithm by understanding it as a stochastic approximation to gradient ascent. In exact gradient ascent, each action preference would be incremented in proportion to the increment’s effect on performance:

(3)

where the measure of the performance here is the expected reward:

and the measure of the increment’s effect is the *partial derivative* of this performance measure with respect to the action preference.