# 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