- **bold**
- underline
- *italic*

A collection of notes from the textbook, *Reinforcement Learning* by Richard Sutton and Andrew Barto. Available at `http://incompleteideas.net/book/the-book.html`

# 1 Action selection

Most RL methods require some form of policy or action-value based action selection alogrithm.

- **Greedy Selection**: Choosing the best action.

$$A = \text{argmax}_a Q(a)$$

- **$\varepsilon$-greedy Selection**: Simple exploration with $\varepsilon$-probability.

$$A \leftarrow \begin{cases} \text{argmax}_a Q(a) & \text{with probability 1 - } \varepsilon \text{ (breaking ties randomly)} \\ \text{a random action} & \text{with probability } \varepsilon \end{cases}$$

- **Upper Confidence Bound (UCB)**: Takes into account the proximity of the estimate to being maximal and the uncertanty in the estimates. Does not perform well on large state spaces.

$$A_t = \underset{a}{\text{argmax}} \left[ Q_t(a) + c\sqrt{\frac{\ln t}{N_t(a)}} \right]$$

  Where:

  - $c > 0$ is the degree of exploration
  - $N_t(a)$ is the number of times that action $a$ has been selected prior to time $t$. If $N_t(a) = 0$, then $a$ is considered to be a maximizing action.

# 2 Performance Measures