ARTIFICIAL INTELLIGENCE

FOUNDATIONS OF COMPUTATIONAL AGENTS

11.3.4 Exploration and Exploitation

The Q-learning algorithm does not specify what the agent should actually do. The agent learns a Q-function that can be used to determine an optimal action. There are two things that are useful for the agent to do:

- **exploit** the knowledge that it has found for the current state s by doing one of the actions a that maximizes $Q[s,a]$.
- **explore** in order to build a better estimate of the optimal Q-function. That is, it should select a different action from the one that it currently thinks is best.

There have been a number of suggested ways to trade off exploration and exploitation:

- The ε-greedy strategy is to select the greedy action (one that maximizes $Q[s,a]$) all but ε of the time and to select a random action ε of the time, where $0 \leq \varepsilon \leq 1$. It is possible to change ε through time. Intuitively, early in the life of the agent it should select a more random strategy to encourage initial exploration and, as time progresses, it should act more greedily.
- One problem with an ε-greedy strategy is that it treats all of the actions, apart from the best action, equivalently. If there are two seemingly good actions and more actions that look less promising, it may be more sensible to select among the good actions: putting more effort toward determining which of these promising actions is best, rather than putting in effort to explore the actions that look bad. One way to do that is to select action a with a probability depending on the value of $Q[s,a]$. This is known as a **soft-max** action selection. A common method is to use a **Gibbs** or **Boltzmann distribution**, where the probability of selecting action a in state s is proportional to $e^{Q[s,a]/\tau}$. That is, in state s, the agent selects action a with probability

$$(e^{Q[s,a]/\tau})/(\sum_a e^{Q[s,a]/\tau})$$

  where $\tau>0$ is the **temperature** specifying how randomly values should be chosen. When τ is high, the actions are chosen in almost equal amounts. As the temperature is reduced, the highest-valued actions are more likely to be chosen and, in the limit as $\tau \to 0$, the best action is always chosen.
- An alternative is "optimism in the face of uncertainty": initialize the Q-function to values that encourage exploration. If the Q-values are initialized to high values, the unexplored areas will look good, so that a greedy search will tend to explore. This does encourage exploration; however, the agent can hallucinate that some state-action pairs are good for a long time, even though there is no real evidence for it. A state only gets to look bad when all its actions look bad; but when all of these actions lead to states that look good, it takes a long time to get a realistic view of the actual values. This is a case where old estimates of the Q-values can be quite bad estimates of the actual Q-value, and these can remain bad estimates for a long time. To get fast convergence, the initial values should be as close as possible to the final values; trying to make them an overestimate will make convergence slower. Relying only on optimism in the face if uncertainty is not useful if the dynamics can change, because it is treating the initial time period as the time to explore and, after this initial exploration, there is no more exploration.

It is interesting to compare the interaction of the exploration strategies with different choices for how α is updated. See Exercise 11.8.