Agent Architectures

All agents (?) can be modeled by a set of functions:

- **Sensors:** states $\rightarrow$ percepts
- **Sensor Interpretation:** percepts + actions $\rightarrow$ memory
- **Policy:** memory $\rightarrow$ action
- **Effectors:** action $\rightarrow$ change in environment

Models of Optimal Agents: Fully-Observable Case

- Percept = State
- Optimal agent is mapping from states to actions
- No memory is needed

Models of Optimal Agents: Partially-Observable Case

- Percept provides partial information about state
  Let $h = (p_1, a_1, p_2, a_2, \ldots)$ be the percept-action history of the agent.
- Optimal agent is mapping from percept histories to actions.
  $\pi^* : H \rightarrow A$
- Other memory structures are possible. For example, the *belief state*
  Define $\mathbf{b}$ to be a probability distribution of all belief states:

$$\mathbf{b} = \mathbf{P}(s|h)$$

- The optimal agent is a mapping from belief states to actions: $\pi^* : B \rightarrow A$

Partially-Observable Markov Decision Process

- **Dynamics of environment:** $P(s'|s,a)$
- **Sensor model:** $P(o|s)$
- **Reward signal:** $P(r|s,a,s')$
- **Protocol:**

  At time t:

  - Agent observes o_t drawn according to $P(o_t|s_t)$
  - Agent performs action a_t
  - Environment changes to s_{t+1} according to $P(s_{t+1}|s_t,a_t)$
  - Agent receives reward r_{t+1} according to $P(r_{t+1}|s_t,a_t,s_{t+1})$

  Rewards are accumulated, but the agent does not actually see them.

Notational Differences

The book uses $\mathcal{P}_{s,s'}^{a} = P(s'|s,a)$

The book uses $\mathcal{R}_{s,s'}^{a} = R(s,a,s') = \sum_{r} P(r|s,a,s') \cdot r$

I also sometimes write this as $R(s'|s,a)$ because this is a natural notation in some cases.

Notes on the Optimal Policy

- **In MDPs, there is always a deterministic optimal policy**

  If there are ties in $Q^*(s, a)$, then the optimal policy may break those ties stochastically or by any fixed arbitrary scheme.

- **In POMDPs, the optimal policy may be stochastic.**

  Example: 2-armed bandit and knowledge that one of these bandits will pay off \$1 million within 10 trials.