Optimal control, MDP w/ known dynamics

$$V^*(s) = \max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma V^*(s')\right)$$

$$\pi^*(s) = \arg\max_a$$ ——— " " ———

(1) Value Iteration

$$V_k^*(s) = \max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma V_{k-1}^*(s')\right)$$

$$\pi_k^*(s) = \arg\max_a$$ ——— " " ———

$$Q^*(s,a) = \sum_{s',r} p(s',r|s,a)\left(r + \gamma \max_{a'} Q^*(s',a')\right)$$

$$\pi^*(s) = \arg\max_a Q^*(s,a)$$

(2) Q-Value Iteration

$$Q_k^*(s,a) = \sum_{s',r} p(s',r|s,a)\left(r + \gamma \max_{a'} Q_{k-1}^*(s',a')\right),$$

$$\pi_k^*(s) = \arg\max_a Q_k^*(s,a)$$

Policy evaluation for a given $\pi(s)$

$$V^\pi(s) = \sum_{s',r} p(s',r|s,\pi(s))\left(r + \gamma V^\pi(s')\right) \rightarrow$$ solve system of $|S|$ linear equations

OR

$$V_k^\pi(s) = \sum_{s',r} p(s',r|s,\pi(s))\left(r + \gamma V_{k-1}^\pi(s')\right) \rightarrow$$ run until convergence.

(3) Policy Iteration

1. Policy evaluation for current policy π_j
2. Policy improvement (best action using one-step lookahead)

$$\pi_{j+1} = \arg\max_a \sum_{s',r} p(s',r|s,a)\left(r + \gamma V^{\pi_j}(s')\right)$$

repeat until convergence.