Convergence of value iteration

Theorem: Value iteration converges to optimal value: $\hat{V} \rightarrow V^\star$

Proof: For any estimate of the value function $\hat{V}$, we define the Bellman backup operator $B : \mathbb{R}^{|\mathcal{S}|} \rightarrow \mathbb{R}^{|\mathcal{S}|}$

$$B\,\hat{V}(s) = R(s) + \gamma \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s,a)\,\hat{V}(s')$$

We will show that Bellman operator is a *contraction*, that for any value function estimates $V_1,\ V_2$

$$\max_{s \in \mathcal{S}} |B V_1(s) - B V_2(s)| \leq \gamma \max_{s \in \mathcal{S}} |V_1(s) - V_2(s)|$$

Since $B V^\star = V^\star$ (the contraction property also implies existence and uniqueness of this fixed point), we have:

$$\max_{s \in \mathcal{S}} \left| B\,\hat{V}(s) - V^\star(s) \right| \leq \gamma \max_{s \in \mathcal{S}} \left| \hat{V}(s) - V^\star(s) \right| \implies \hat{V} \rightarrow V^\star$$

Proof of contraction property:

$$
\begin{aligned}
&|BV_1(s) - BV_2(s)| \\
&= \gamma \left| \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V_1(s') - \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V_2(s') \right| \\
&\le \max_{a \in \mathcal{A}} \left| \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V_1(s') - \sum_{s' \in \mathcal{S}} P(s'|s,a)\, V_2(s') \right| \\
&= \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s'|s,a)\, |V_1(s') - V_2(s')| \\
&\le \gamma \max_{s \in \mathcal{S}} |V_1(s) - V_2(s)|
\end{aligned}
$$

where third line follows from property that

$$
|\max_x f(x) - \max_x g(x)| \le \max_x |f(x) - g(x)|
$$

and final line because $P(s'|s,a)$ are non-negative and sum to one

Value iteration convergence

How many iterations will it take to find optimal policy?

Assume rewards in $[0, R_{\max}]$, then

$$V^\star(s) \leq \sum_{t=1}^{\infty} \gamma^t R_{\max} = \frac{R_{\max}}{1-\gamma}$$

Then letting V^k be value after kth iteration

$$\max_{s \in \mathcal{S}} |V^k(s) - V^\star(s)| \leq \frac{\gamma^k R_{\max}}{1-\gamma}$$

i.e., we have linear convergence to optimal value function

But, time to find optimal policy depends on separation between value of optimal and second suboptimal policy, difficult to bound