CONTENTS

1 REINFORCEMENT LEARNING

1.1 DEFINITIONS

define u^t

1.2 BELLMAN EQUATIONS

Definition 1 (Value function)**.** A **value function** maps states to a policy's expected return from that point. A **state-value function** (Q function) does the same, but with state-action pairs.

$$V^\pi(s) := \mathbb{E}_\pi\left(u^t \mid s^t = s\right)$$
$$Q^\pi(s,a) := \mathbb{E}_\pi\left(u^t \mid s^t = a, a^t = a\right)$$

Note $V^\pi(s) = \mathbb{E}_{a\sim\pi}\left(Q^\pi(s,\cdot)\right) = \sum_a \pi(a|s)Q^\pi(s,a)$.

A policy π^* is optimal if its value functions are optimal:

$$V^*(s) = \max_\pi V^\pi(s) \text{ for all } s$$
$$Q^*(s,a) = \max_\pi Q^\pi(s,a) \text{ for all , } s, a.$$

Both V^π and Q^π can be rewritten recursively, and in that form are called the **Bellman equations**. By expanding $u^t = r^t + \gamma u^{t+1}$ and applying the law of total expectation over states s^{t+1} yields the following.

Proposition 1. Value and state-value functions satisfy the following Bellman equations:

$$V^\pi(s) = \sum_{s',a} \pi(a|s)\mathcal{T}(s'|s,a)\left(\mathcal{R}(s,a,s') + \gamma V^\pi(s')\right),$$
$$Q^\pi(s) = \sum_{s'} \mathcal{T}(s'|s,a)\left(\mathcal{R}(s,a,s') + \gamma V^\pi(s')\right)$$
$$= \sum_{s'} \mathcal{T}(s'|s,a)\left(\mathcal{R}(s,a,s') + \gamma \sum_{a'} \pi(a'|s)Q^\pi(s',a')\right).$$

Consider the system of equations given by the Bellman equation for $V^\pi(s_i)$ for each $s_i \in \mathcal{S}$. This system of $|\mathcal{S}|$ equations can be solved by any usual linear system solver, e.g. Gaussian elimination.

We can make the Bellman equations express how optimal value functions behave by introducting a max term (which also makes them nonlinear).

Note 1 (Bellman optimality equations)**.**

$$V^*(s) = \max_a \sum_{s'} \mathcal{T}(s'|s,a)\left(\mathcal{R}(s,a,s') + \gamma V^*(s)\right),$$

$$Q^*(s,a) = \sum_{s'} \mathcal{T}(s'|s,a)\left(\mathcal{R}(s,a,s') + \gamma \max_{a'} Q^*(s',a')\right).$$

1.3 DYNAMIC PROGRAMMING

The DP approach to computing an optimal policy is to calculate $V^\pi(s)$ for all states, then back out an optimal policy by always moving to the value-maximizing next state.

There are two main types:

- **Policy iteration:** switch between policy evaluation & policy improvement, producing a sequence

$$\pi^0 \to V^0 \to \pi^1 \to V^1 \to \cdots \to \pi^* \to V^*$$

- **Value iteration:** directly measure a value function and back out an optimal policy at the end, producing a sequence

$$V^0 \to V^1 \to \cdots \to V^*$$

In policy iteration, given a policy π, we could use Gaussian elimination on the system of Bellman equations to calculate $V^\pi(s)$ for all $s \in \mathcal{S}$, but this is $\mathcal{O}(|\mathcal{S}|)$. And in the case of value iteration, the system of Bellman optimality equations is nonlinear anyway.

Instead, we use **bootstrapping** to recursively back out a value function. For policy iteration, fix π, then we can compute V^π as the limit of the sequence $V^0 \to V^1 \to \cdots$ defined by

$$V^{i+1(s)} \leftarrow \sum_{s',a} \pi(a|s)\mathcal{T}(s'|s,a)\left(\mathcal{R}(s,a,s') + \gamma V^i(s')\right).$$

Definition 2. A (γ)-**contraction map** on a complete metric space (X,d) is a map $\phi : X \to X$ such that

$$d(\phi(x), \phi(y)) \leq \gamma\, d(x,y)$$

for all $x, y \in X$, where $\gamma \in [0,1)$ is constant.

Theorem 1 (Contraction Mapping Principle)**.** A contraction map ϕ has a unique fixed point x^*

that can be computed as the limit of successive applications of ϕ to an arbitrary starting point x^0:

$$x^0 \to \phi(x^0) \to \phi(\phi(x^0)) \to \cdots \to x^*.$$

Proof. See Marsden & Hoffman pg. 301. □

Proposition 2. For a fixed policy π, the sequence $\{V^i\}$ defined above converges to V^π.

Proof. Consider the Bellman operator

$$\phi : V \mapsto \sum_{s',a} \pi(a|s)\mathcal{T}(s'|s,a)\left(\mathcal{R}(s,a,s') + \gamma V(s')\right).$$

This is a contraction map in the max norm, since for any value functions V, W,

$$\begin{aligned}
\|\phi V - \phi W\|_\infty &= \max_s |(\phi V)(s) - (\phi W)(s)| \\
&= \gamma \max_s \left| \sum_{s',a} \mathcal{T}(s'|s,a)\pi(a|s)\left(V(s') - W(s')\right) \right| \\
&\leq \gamma \max_s \sum_{s',a} \mathcal{T}(s'|s,a)\pi(a|s)\,|V(s') - W(s')| \\
&\leq \gamma \|V - W\|_\infty \max_s \sum_{s',a} \mathcal{T}(s'|s,a)\pi(a|s) \\
&= \gamma \|V - W\|_\infty.
\end{aligned}$$

The last line is because $\sum_{s',a} \mathcal{T}(s'|s,a)\pi(a|s) = 1$ for all s.

Then by the contraction mapping principle, there is a unique fixed point V gotten by starting with an arbitrary value function V^0 and repeatedly applying ϕ. This fixed point satisfies the Bellman equation for V^π by construction, so it's our desired V^π. □

Note 2. The above proof shows that in the finite MDP setting, the Bellman equation for V^π has a unique solution.

This gives us a method for policy evaluation (the step $\pi^i \to V^i$), but we still need to produce a better policy π^{i+1}.

Finish