Proof of Gittins Index theorem

First, we prove that there always exist a **Markovian stationary policy** which is optimal. Markovian stationary policy:

- decision depends only on state: does not change with time
- mapping from state space to (distribution over) actions

$$a_t^* = \pi(s_{1,t}, \ldots, s_{N,t})$$

In fact, Gittins index suggests further that optimal policy is of form

$$a_t^* = \arg \max_{a=1,\ldots,N} G_a(s_{a,t})$$

Existence of Markovian stationary optimal policy is in fact true for more general "Markov Decision Processes".

Proof of Gittins Index theorem

First, we prove that there always exist a **Markovian stationary policy** which is optimal. Markovian stationary policy:

- decision depends only on state: does not change with time
- mapping from state space to (distribution over) actions

$$a_t^* = \pi(s_{1,t}, \ldots, s_{N,t})$$

In fact, Gittins index suggests further that optimal policy is of form

$$a_t^* = \arg\max_{a=1,\ldots,N} G_a(s_{a,t})$$

Existence of Markovian stationary optimal policy is in fact true for more general "Markov Decision Processes".

We want to maximize

$$\mathbb{E}[\sum_{t=1}^{\infty} \gamma^{t-1} r_{a_t}(s_{a_t,t})]$$

where $a_t = \pi_t(\mathcal{H}_t)$, for any history and time dependent policy π.
Let $s_t = (s_{1,t}, \ldots, s_{N,t})$. Define $r_a(s_t) := r_a(s_{a,t})$. For bounded rewards, and $\gamma < 1$, this is same as

$$\lim_{T \to \infty} \mathbb{E}[\sum_{t=1}^{T} \gamma^{t-1} r_{a_t}(s_t)]$$

(under above conditions, the limit exists, and by bounded convergence theorem, expectation and limit are interchangeable. In fact, if the reward is bounded, then the limit also exists for average reward criterion.)

Finite Horizon

Let us first consider any **finite horizon** T.
Define value of a state $V^{\pi}_{\ell,T}(s)$ as the expected discounted reward achieved by policy π given $s_\ell = s$, and policy π is followed from time ℓ to T.

$$V^{\pi}_{\ell,T}(s) = \mathbb{E}[\sum_{t=\ell}^{T} \gamma^{t-\ell} r_{\pi_t(H_t)}(s_t) | s_\ell = s]$$

Let us first consider any **finite horizon** T.
Define value of a state $V^{\pi}_{\ell,T}(s)$ as the expected discounted reward achieved by policy π given $s_\ell = s$, and policy π is followed from time ℓ to T.

$$V^{\pi}_{\ell,T}(s) = \mathbb{E}[\sum_{t=\ell}^{T} \gamma^{t-\ell} r_{\pi_t(H_t)}(s_t) | s_\ell = s]$$

Below we derive a recursive formula to compute V^{π}_{ℓ}: Let $a_\ell = \pi_\ell(\mathcal{H}_\ell)$.

$$\begin{aligned}
V^{\pi}_{\ell,T}(s) &= \mathbb{E}[\sum_{t=\ell}^{T} \gamma^{t-\ell} r_{\pi_t(H_t)}(s_t) | s_\ell = s] \\
&= \mathbb{E}[r_{a_\ell}(s) + \sum_{t=\ell+1}^{T} \gamma^{t-\ell} r_{\pi_t(H_t)}(s_t) | s_\ell = s] \\
&= \mathbb{E}[r_{a_\ell}(s) + \sum_{s'} \sum_{t=\ell+1}^{T} \gamma^{t-\ell} \mathbb{E}[r_{\pi_t(H_t)}(s_t) | s_{\ell+1} = s'] p_{a_\ell}(s, s') | s_\ell = s] \\
&= \mathbb{E}[r_{a_\ell}(s) + \sum_{s'} \gamma \mathbb{E}[\sum_{t=\ell+1}^{T} \gamma^{t-\ell-1} r_{\pi_t(H_t)}(s_t) | s_{\ell+1} = s'] p_{a_\ell}(s, s')] \\
&= \mathbb{E}[r_{a_\ell}(s) + \gamma \sum_{s'} V^{\pi}_{\ell+1,T}(s') p_{a_\ell}(s, s')]
\end{aligned}$$

Dynamic Programming for finite horizon

Claim: Let $V^*_{t,T}(s)$ be the value of state s from time t to T for optimal policy π^*, then

$$V^*_{\ell,T}(s) = \max_a \left\{ r_a(s) + \gamma \sum_{s'} V^*_{\ell+1,T}(s') p_a(s,s') \right\}$$

Note that this claim provides dynamic programming solution for finite horizon.

Also, above claim implies that optimal policy is a **Markovian** policy: only depends on current state and not past history. But for finite horizon, this is not stationary: value of next state depends on number of time steps remaining.

Claim: Let $V^*_{t,T}(s)$ be the value of state s from time t to T for optimal policy π^*, then

$$V^*_{\ell,T}(s) = \max_a \left\{ r_a(s) + \gamma \sum_{s'} V^*_{\ell+1,T}(s') p_a(s,s') \right\}$$

Note that this claim provides dynamic programming solution for finite horizon.

Also, above claim implies that optimal policy is a **Markovian** policy: only depends on current state and not past history. But for finite horizon, this is not stationary: value of next state depends on number of time steps remaining.

Here, we prove the dynamic programming equations. We prove that for any π, ℓ, s, $V^{\pi}_{\ell,T}(s) \leq V^*_{\ell,T}(s)$. Proof by backward induction on ℓ. For $\ell = T$, $V^{\pi}_{T,T}(s) = r_{a_T}(s) \leq \max_a r_a(s) = V^*_{T,T}(s)$.
Now, assume above claim is true for $\ell + 1$. Then,

$$\begin{aligned}
V^{\pi}_{\ell}(s) &= \mathbb{E}[r_{\pi(H_\ell)}(s) + \gamma \sum_{s'} V^{\pi}_{\ell+1}(s') p_{\pi(H_\ell)}(s,s')] \\
&\leq \max_a \left\{ r_a(s) + \gamma \sum_{s'} V^{\pi}_{\ell+1}(s') p_a(s,s') \right\} \\
&\leq \max_a \left\{ r_a(s) + \gamma \sum_{s'} V^{*}_{\ell+1}(s') p_a(s,s') \right\} \\
&= V^*_{\ell}(s)
\end{aligned}$$

Infinite Horizon

Now, define

$$V^{\pi}(s) = \lim_{T \to \infty} V^{\pi}_{1,T}(s)$$

. From the dynamic programminh equations, it is easy to see that $V^{*}_{\ell,T}$ depends only on the number of remaining time steps, i.e.,

$$V^{*}_{\ell,T}(s) = V^{*}(\ell - 1, T - 1) = \cdots = V^{*}(1, T - \ell + 1)$$

A rigorous proof can be obtained by induction on ℓ. Therefore, for any ℓ,

$$\lim_{T \to \infty} V^{*}_{\ell,T}(s) = \lim_{T \to \infty} V^{*}_{1,T-\ell+1}(s) = V^{*}(s)$$

Therefore, for infinite horizon:

$$V^{*}_{\ell,\infty}(s) = V^{*}(s) \quad = \quad \max_{a} \left\{ r_a(s) + \gamma \sum_{s'} V^{*}(s') p_a(s, s') \right\}$$

Above are called **Bellman equations**. Note that this implies that optimal policy is **Markovian and stationary**.