Policy Iteration and Value Iteration
Proof of Convergence
**Value Iteration**

- **Algorithm**
  - we start with an arbitrary initial value function $V_0$
  - at each iteration $k$, we calculate $V_{k+1} = \mathcal{T}V_k$

- **Convergence**: show that $\lim_{k \to \infty} V_k = V^*$.

- **proof**

$$||V_{k+1} - V^*||_\infty = ||\mathcal{T}V_k - \mathcal{T}V^*||_\infty \leq \gamma ||V_k - V^*||_\infty \leq \ldots \leq \gamma^{k+1} ||V_0 - V^*||_\infty \longrightarrow 0$$
Algorithm

- we start with an arbitrary initial policy $\pi_0$
- at each iteration $k$, given the current policy $\pi_k$

**Policy Evaluation:** we calculate the value function $V^{\pi_k}$

**Policy Improvement:** we calculate the new policy $\pi_{k+1}$ as

$$\pi_{k+1}(s) \in \arg \max_{a \in A} \left[ r(s, a) + \gamma \sum_{s'} p(s'|s, a) V^{\pi_k}(s') \right]$$

Policy $\pi_{k+1}$ is greedy w.r.t. the value function $V^{\pi_k}$ (i.e., $\mathcal{T}^{\pi_{k+1}} V^{\pi_k} = \mathcal{T} V^{\pi_k}$)

- we stop when $V^{\pi_{k+1}} = V^{\pi_k}$. 
show that $V^{\pi_{k+1}} \geq V^{\pi_k}$

**proof:** from the definitions, we have

$$V^{\pi_k} = \mathcal{T}^{\pi_k} V^{\pi_k} \leq \mathcal{T} V^{\pi_k} = \mathcal{T}^{\pi_{k+1}} V^{\pi_k}$$

because of the monotonicity of $\mathcal{T}^{\pi_{k+1}}$, from $V^{\pi_k} \leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k}$, we may deduce

$$V^{\pi_k} \leq \mathcal{T}^{\pi_{k+1}} V^{\pi_k} \leq (\mathcal{T}^{\pi_{k+1}})^2 V^{\pi_k} \leq \ldots \leq \lim_{n \to \infty} (\mathcal{T}^{\pi_{k+1}})^n V^{\pi_k} = V^{\pi_{k+1}}$$
algorithm stops after a finite number of steps $q$ with the optimal policy $V^{\pi_q} = V^*$

**proof:** since there exists only a finite number of policies, the algorithm stops after a finite number of steps $q$ with $V^{\pi_q} = V^{\pi_{q+1}}$

\[
V^{\pi_q} = V^{\pi_{q+1}} = \mathcal{T}^{\pi_{q+1}} V^{\pi_{q+1}} = \mathcal{T}^{\pi_{q+1}} V^{\pi_q} = \mathcal{T} V^{\pi_q}
\]

so $V^{\pi_q}$ is a fixed point of $\mathcal{T}$. Since $\mathcal{T}$ has a unique fixed point, we may deduce that $V^{\pi_q} = V^*$, and thus, $\pi_q$ is an optimal policy.