MDP

$$\langle S, A, A, T, r, S_+, \gamma \rangle$$

$$T(s, a, s') = Pr[S_{t+1} = s' \mid S_t = s, A_t = a)$$

$$\sum_{s' \in S} T(s, a, s') = 1$$

```
Class MDP_sim :
    def estado_Inicial()
        ret S0
    def acciones_legales(s)
        ret A(s)
    def transición(s, a)
        ret s'
    def r(s, a, s')
        ret recompensa
    def terminal(s)
        ret es_terminal
```

TD(0)

Estimar el valor de la política

```
def TD0(mdp_sim, π, ε, num_episodios)
    V={0 for a in π.keys()}
    for _ in range(num_episodios)
        S = mdp_sim.estado_inicial()
        for _ in range numiter
            if mdp_sim.terminal(S)
                break
            a = π(S)
            S' = mdp_sim.trans(S)
            r = mdp_sim.reward(S, a, S')
            V = V[S]
            V[S] = V[S]+α·(r + mdp_sim.γ x V[S'] - V[S])
            Δ = max(Δ, abs(V - V[S]))
            S = S'
        If Δ < ε
            break
return V
```