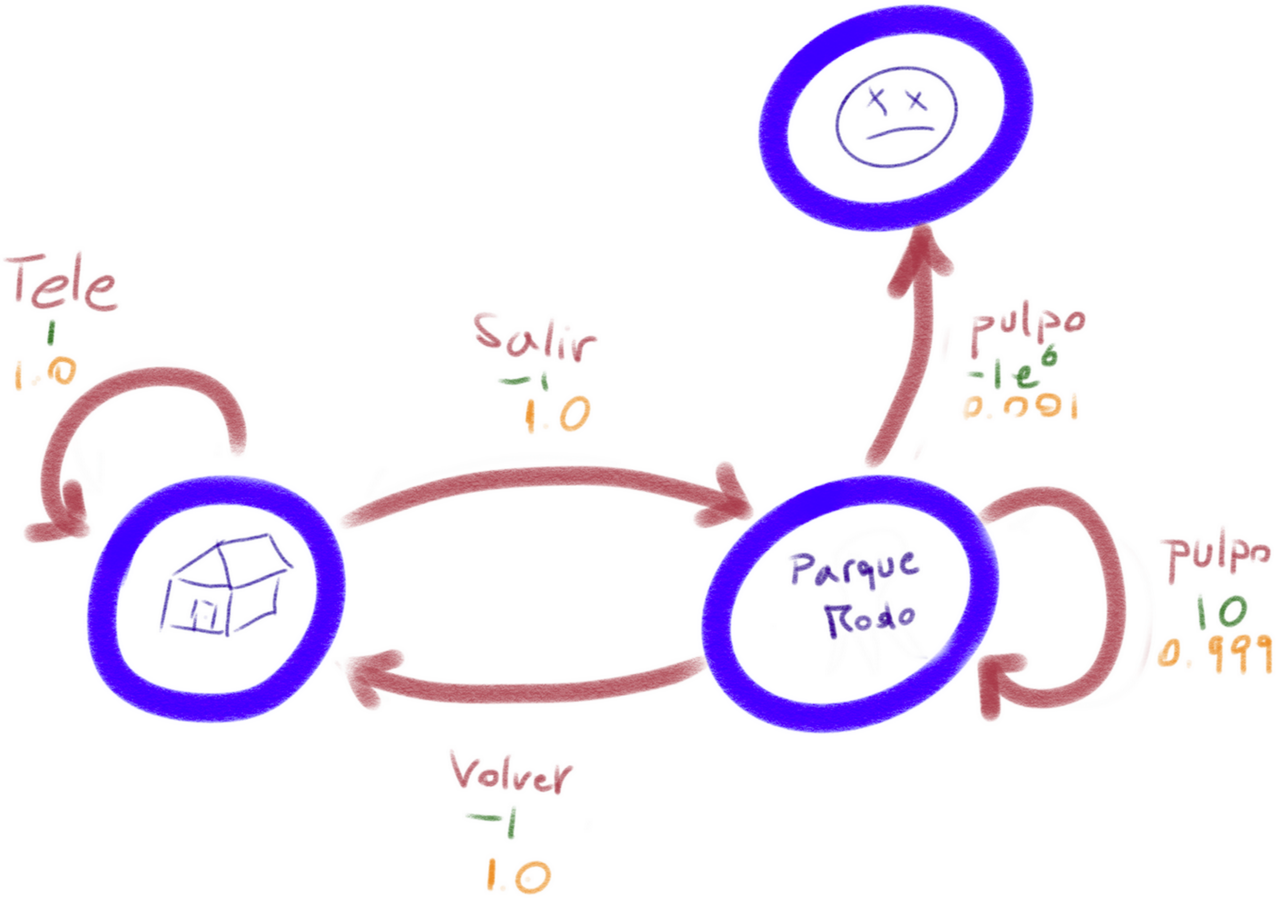
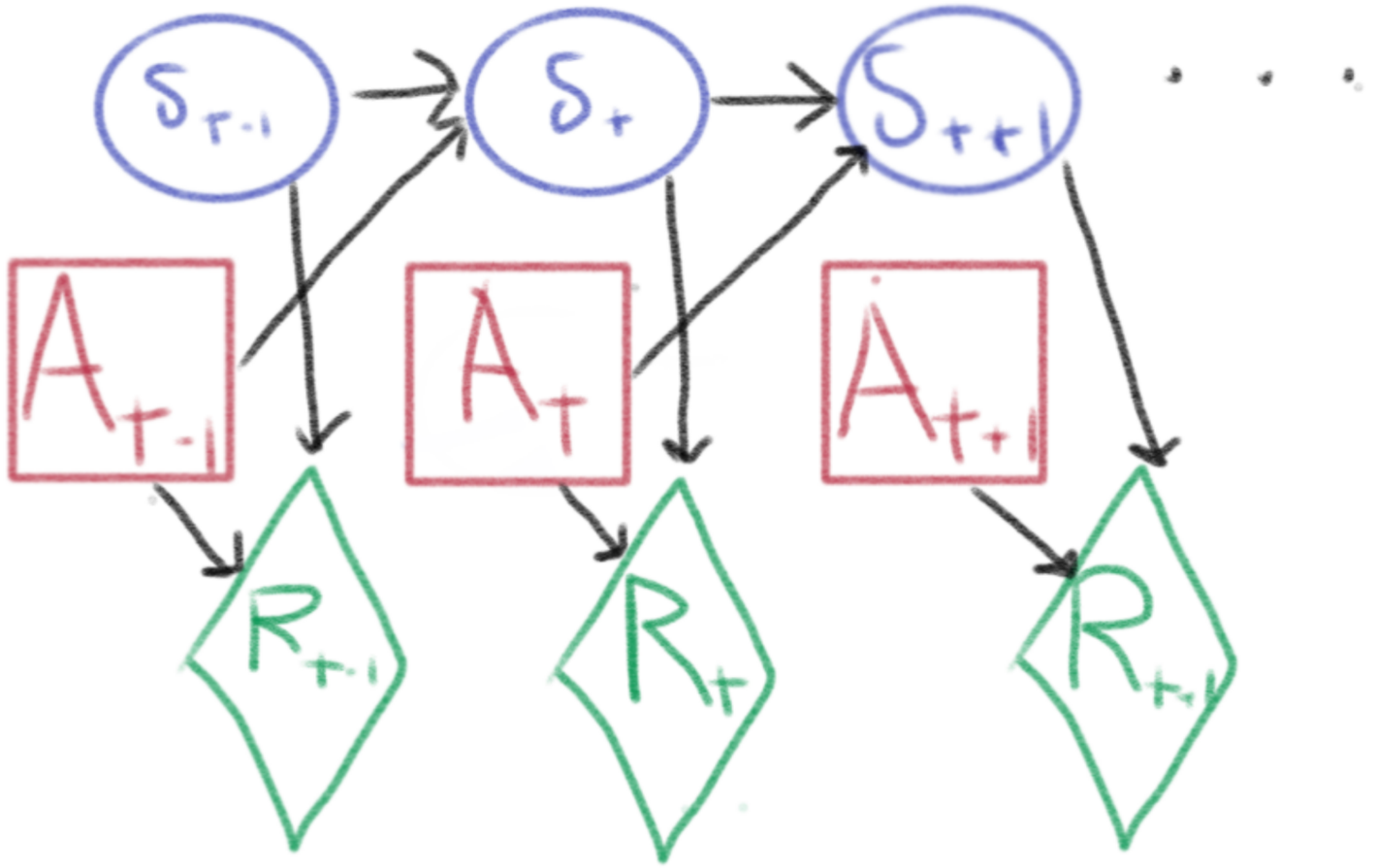
Reinforcement Learning

Markov Decision Process


Tele
1
1.0
Salir
-1
1.0
pulpo
-1e^6
0.001
Parque
Rodo
pulpo
10
0.999
Volver
-1
1.0


S_{t-1}
S_t
S_{t+1}
A_{t-1}
A_t
A_{t+1}
R_{t-1}
R_t
R_{t+1}

Objetivo: Maximizar Retorno

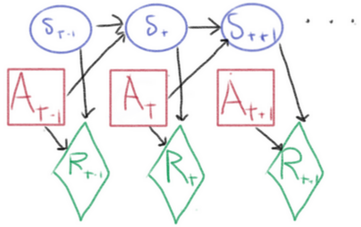
$$max_{a_1..a_t} G_T$$

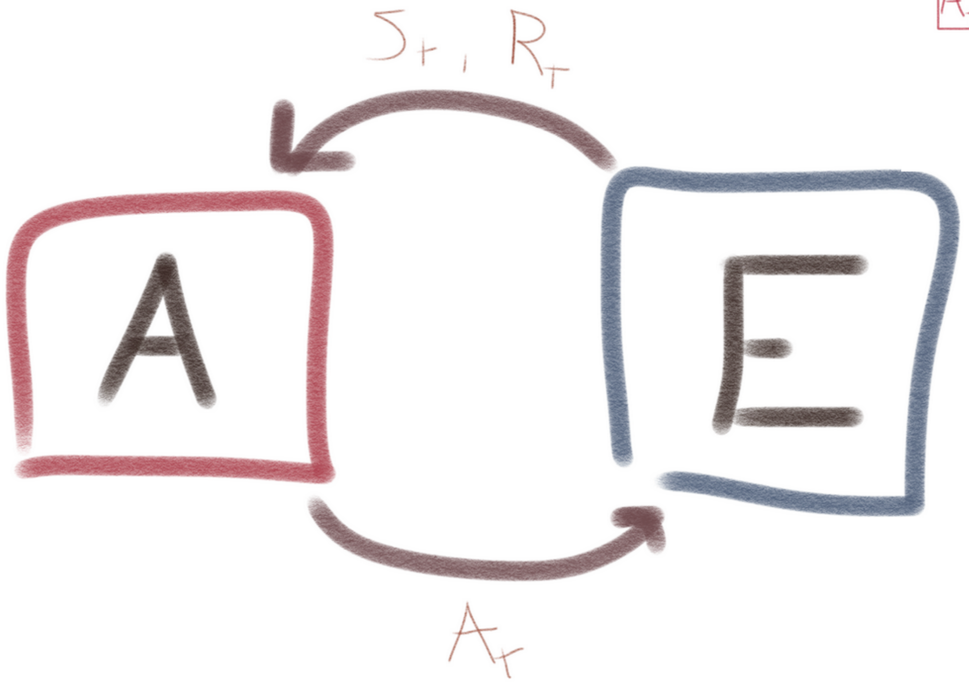
$$G_t = \sum_{t=1}^{t=T} \gamma^t R_t$$

$\gamma \in (0, 1]$

$T = Tiempo\ de\ terminación$

RL - El problema





Multi -Tragamonedas

Multi Armed Bandit

$$S_t \in \{\emptyset\}$$

$$A_t \in \{1, 2, 3\}$$

$$p(R_t|A_t) = \mathcal{N}(\mu_{A_t}, 1)$$

Multi Multi -Tragamonedas

Multi Multi Armed Bandit

$$S_t \in \{a, b, c\}$$

$$A_t \in \{1, 2, 3\}$$

$$p(R_t|A_t, S_t) = \mathcal{N}(\mu_{A_t, S_t}, 1)$$

$$\exists t :: p(S_t \neq S_{t+1}) > 0$$

$$\exists t :: p(S_{t+1}|S_t, A_t) \neq p(S_{t+1}|S_t)$$

a

b

c

Determinismo

$$S_t \in \{a, b, c\}$$

$$A_t \in \{1, 2, 3\}$$

$$p(R_t | A_t, S_t) = \mathcal{N}(\mu_{A_t, S_t}, 1)$$

a

b

c

$$p(S_{t+1} | S_t, A_t) \in \{0, 1\}$$

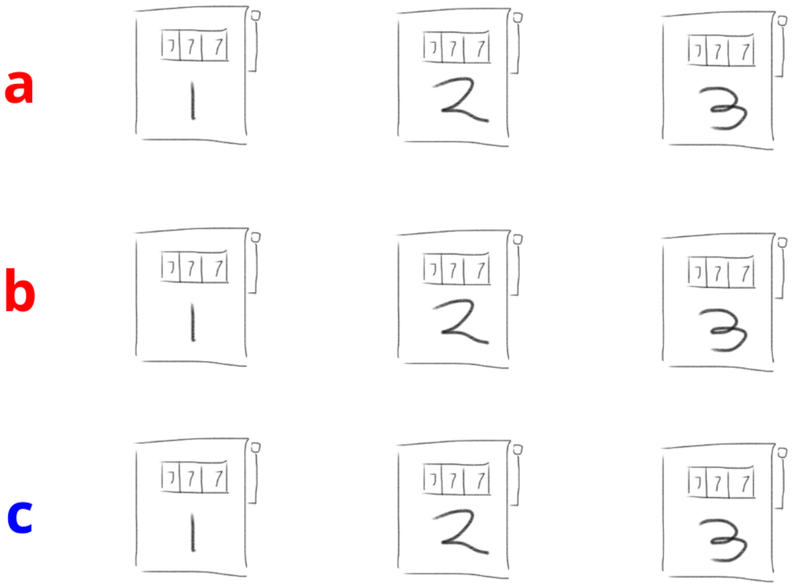
Observabilidad

$$HS_t \in \{a, b, c\}$$

$$S_t \in \{rojo, azul\}$$

$$A_t \in \{1, 2, 3\}$$

$$p(R_t|A_t, S_t) = \mathcal{N}(\mu_{A_t, S_t}, 1)$$



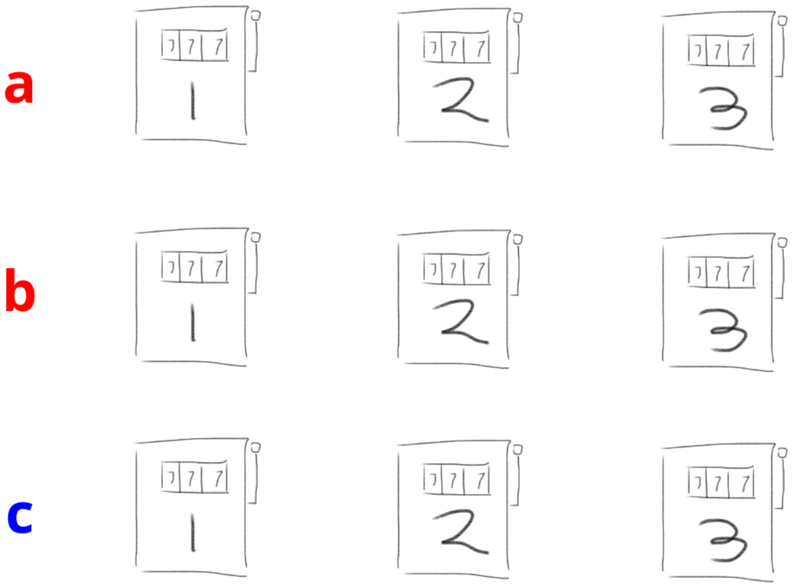
Observabilidad

$$HS_t \in \{a, b, c\}$$

$$S_t \in \{rojo, azul\}$$

$$A_t \in \{1, 2, 3\}$$

$$p(R_t|A_t, S_t) = \mathcal{N}(\mu_{A_t,S_t}, 1)$$

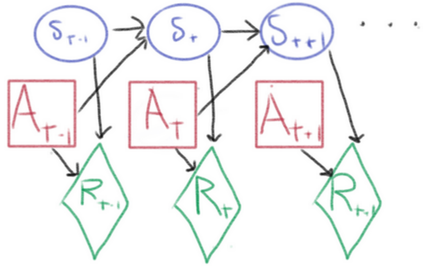


Algo Observable: $p(S_{t+1}|S_t, A_t) \neq p(S_{t+1}|A_t)$

Modelabilidad

$$p(S_{t+1}, R_{t+1} | S_t, A_t)$$

Conocida



Modelo conocido implica

1. Observable
2. Retorno conocido

Modelo no conocido parecido a

1. No observable
2. Estocastico
3. Adversario/Estrategico

Usualmente

No modelable

- Distribución desconocida
- Demasiados estados
- Estados/Acciones continuas (dificulta)

No observable

- El estado suele tener demasiadas variables
- Pero... en este punto hay esperanza: belief state
- Vamos a asumir "un grado de observabilidad"

Estocástico

- Porque el mundo lo es
- Porque el modelo determinista no es completo
- Porque es no observable
- Porque hay otros agentes

Valor, Política y Bellman

Valor

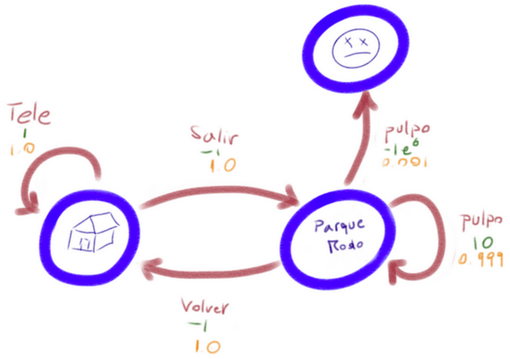
$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$v(s) = \mathbb{E}\left[\sum_{k=0}^{k=\infty} \left[\gamma^k R_{t+k+1} | S_t = s\right]\right]$$

Valor

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$v(s) = \mathbb{E}\left[\sum_{k=0}^{k=\infty} \left[\gamma^k R_{t+k+1} | S_t = s\right]\right]$$



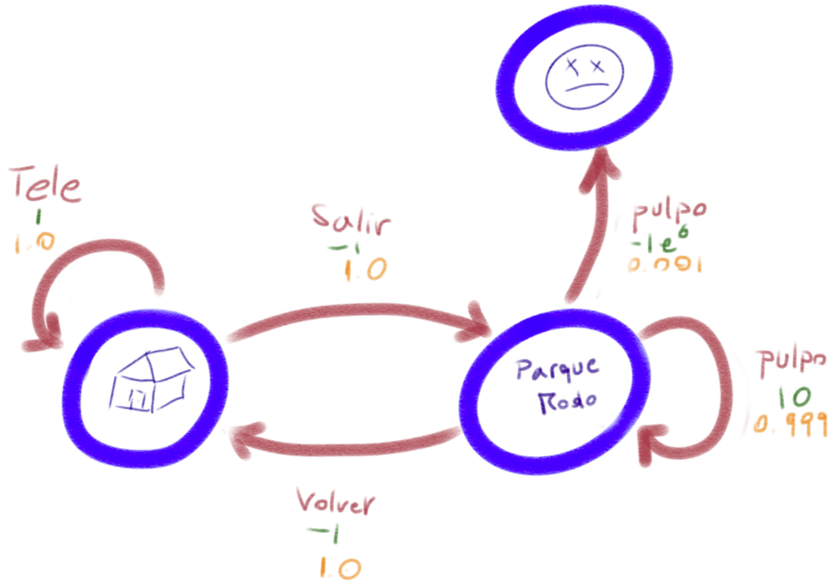
$$V(casa) = ?$$

$$V(PR) = ?$$

$$V(Muerto) = 0$$

Política

$$p(A_t = a|S_t = s) = \pi(a|s)$$



$$\pi(A_t = salir|S_t = casa) = ?$$

$$\pi(A_t = tele|S_t = casa) =$$

$$1 - \pi(A_t = salir|S_t = casa)$$

Valor Sujeto a una Política

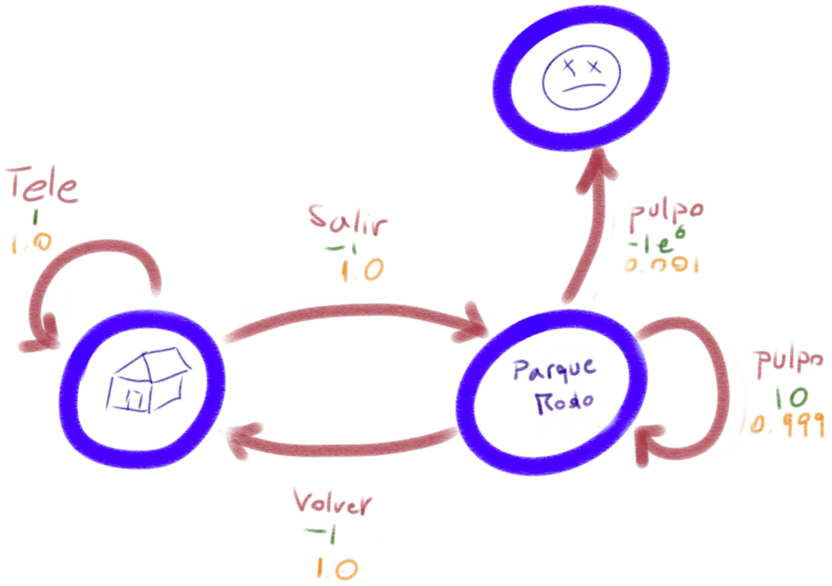
$$v_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{k=\infty} \left[\gamma^k R_{t+k+1} | S_t = s \right] \right]$$

Si

$\pi(salir|casa) = 0,$

cuanto es

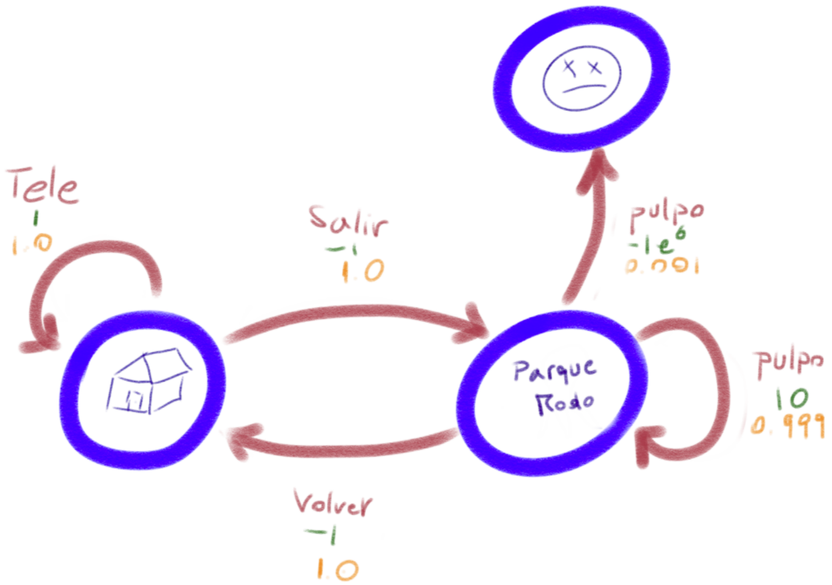
$v_\pi(casa)$?



Valor Sujeto a una Política

$$v_\pi(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{k=\infty} \left[\gamma^k R_{t+k+1} | S_t = s \right] \right]$$

Si $\pi(pulpo|PR) = 1$,
cuanto es $v_\pi(PR)$?



Valor Óptimo

$$v_*(s) = \max_\pi v_\pi(s)$$

Bellman

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

Bellman

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$v_*(s) = \max_a \left(\mathbb{E}\left[G_t | S_t = s, A_t = a\right]\right)$$

Bellman

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$v_*(s) = \max_a \left(\mathbb{E}\left[G_t | S_t = s, A_t = a\right]\right)$$

$$v_*(s) = \max_a \left(\mathbb{E}_{\pi_*}\left[G_t | S_t = s, A_t = a\right]\right)$$

Bellman

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$v_*(s) = \max_a \left(\mathbb{E}\left[G_t | S_t = s, A_t = a\right]\right)$$

$$v_*(s) = \max_a \left(\mathbb{E}_{\pi_*}\left[G_t | S_t = s, A_t = a\right]\right)$$

$$v_*(s) = \max_a \left(\mathbb{E}_{\pi_*}\left[R_t + \gamma G_{t+1} | S_t = s, A_t = a\right]\right)$$

Bellman

$$v(s) = \mathbb{E}\left[G_t | S_t = s\right]$$

$$v_*(s) = \max_a \left(\mathbb{E}\left[G_t | S_t = s, A_t = a\right]\right)$$

$$v_*(s) = \max_a \left(\mathbb{E}_{\pi_*}\left[G_t | S_t = s, A_t = a\right]\right)$$

$$v_*(s) = \max_a \left(\mathbb{E}_{\pi_*}\left[R_t + \gamma G_{t+1} | S_t = s, A_t = a\right]\right)$$

$$v_*(s) = \max_a \left(\mathbb{E}_{\pi_*}\left[R_t + \gamma v_*\left(S_{t+1}\right) | S_t = s, A_t = a\right]\right)$$

Aprendiendo una Política Óptima

Calcular $v_\pi(s)$

$$v_*(s) = \max_a \left(\mathbb{E}_{\pi_*} \left[R_t + \gamma v_* \left(S_{t+1} \right) | S_t = s, A_t = a \right] \right)$$

Si conocemos $p(S_{t+1}, R_t | S_t, A_t)$, programación dinámica:

$$v_\pi(s) = \max_a \sum_{s', r} p(S_{t+1} = s', R_t = r | S_t = s, A_t = a) \left[r + \gamma v_\pi(s') \right]$$

Si no queremos pensar mucho, Montecarlo:

$$v_\pi(s) = \frac{\sum G_t 1_{S_t = s}}{1_{S_t = s}}$$

Si queremos mejorar, bootstrapping usando la experiencia:

$$v_\pi(s) = R_t + v_\pi(s')$$

Más Alla del Aprendizaje Puro

Un programa agente como política incial

Un estimado inicial de la función de valor

Un modelo (aproximado) del entorno para hacer planning

Un conjunto de datos para aprendizaje off-policy

Un esquema de recompensas más granular

Diseño de la señal de estado