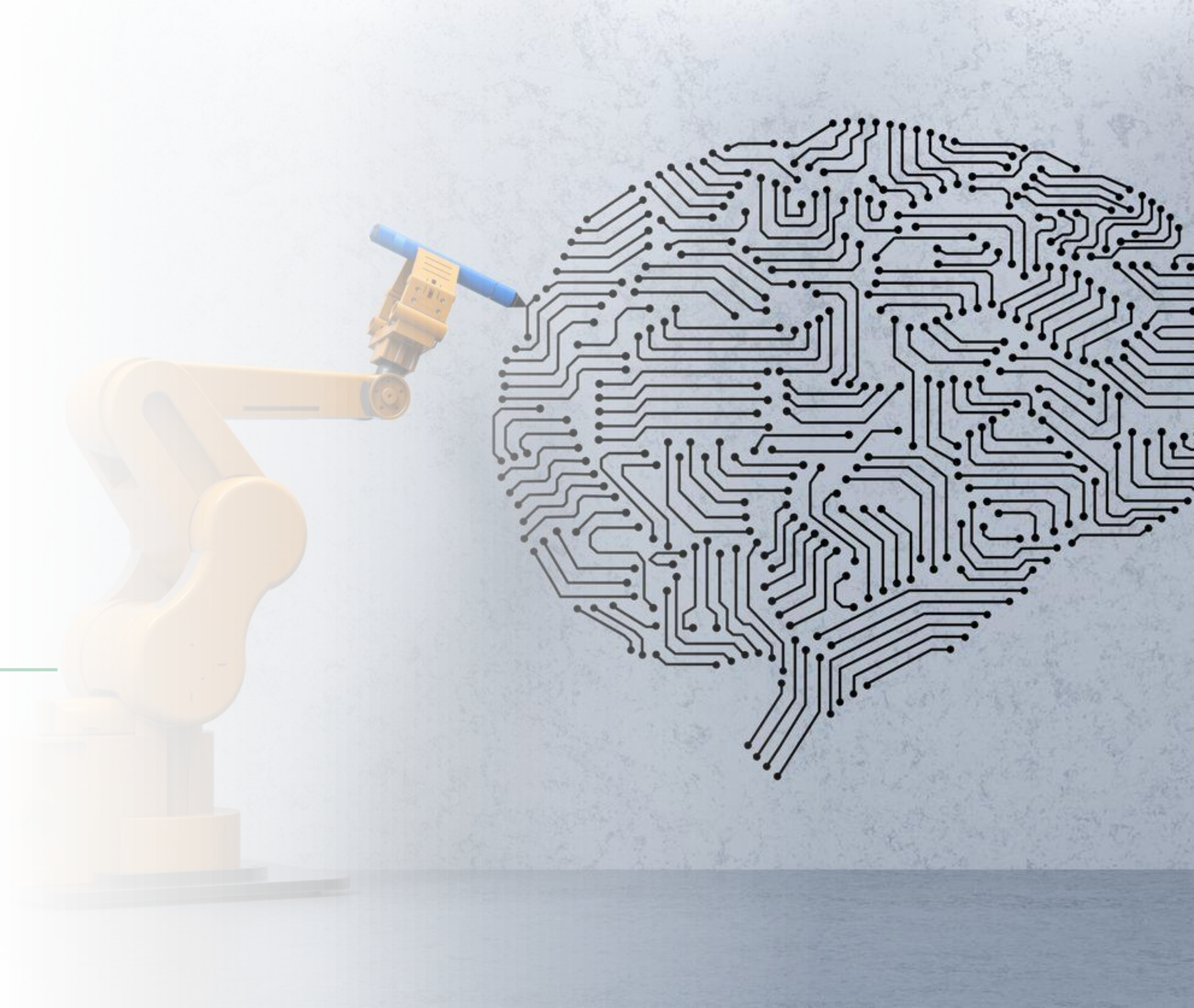


Aprendizaje por refuerzo

Clase 3: MDPs

Para el día de hoy…

- Procesos de decisión de Markov (MDPs)
- Políticas
- Ecuación de Bellman para MDPs

Introducción a MDPs

- Los procesos de decisión de Markov describen formalmente el ambiente para aprendizaje por refuerzo (cuando el ambiente es completamente observable)
- Casi todos los problemas de aprendizaje por refuerzo pueden ser descritos mediante MDPs
  - Control óptimo trata con MDPs continuos
  - Los procesos parcialmente observables pueden ser convertidos en MDPs

Propiedad de Markov

- Un estado S_t es Markov si y solo si

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \dots, S_t]$$

- El futuro es independiente del pasado dado el presente

Proceso de decisión de Markov

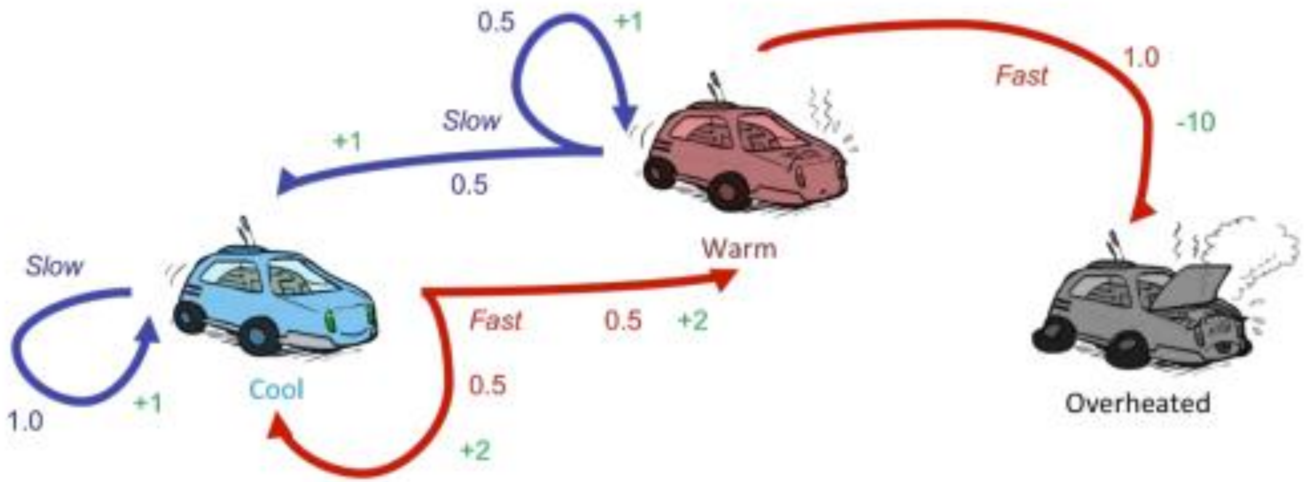
- Un proceso de decisión de Markov es una tupla $(\mathcal{S}, \mathcal{A}, p, \gamma)$
  - $\mathcal{S}$ es un conjunto de estados
  - $\mathcal{A}$ es un conjunto de acciones (pueden depender de $s \in \mathcal{S}$)
  - $p(r, s'|s, a)$ es la distribución de probabilidad conjunta de la recompensa $r \in \mathcal{R} \subseteq \mathbb{R}$ y el siguiente estado s', dado un estado s y una acción a. $\mathcal{S} \times \mathcal{R} \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$
  - $\gamma \in [0,1]$ es un factor de descuento
- p define la dinámica del problema
- Algunas veces es útil marginalizar el estado o la recompensa esperada
  - $p(s'|s, a) = \sum_r p(s', r|s, a)$, $\mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$
  - $\mathbb{E}[R|s, a] = \sum_r r \sum_{s'} p(r, s'|s, a)$, $\mathcal{R} \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$

Proceso de decisión de Markov (alternativa)

- Un proceso de decisión de Markov es una tupla $(\mathcal{S}, \mathcal{A}, p, r, \gamma)$
  - $\mathcal{S}$ es un conjunto de estados
  - $\mathcal{A}$ es un conjunto de acciones (pueden depender de $s \in \mathcal{S}$
  - $p(s'|s,a) = \sum_r p(s', r|s,a)$, $\mathcal{S} \times \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$ (T ó $\mathcal{P}$)
  - r: $\mathbb{E}[R|s,a] = \sum_r r \sum_{s'} p(r, s'|s,a)$, $\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ ($\mathcal{R}$)
  - $\gamma \in [0,1]$ es un factor de descuento
- Ambas definiciones son equivalentes

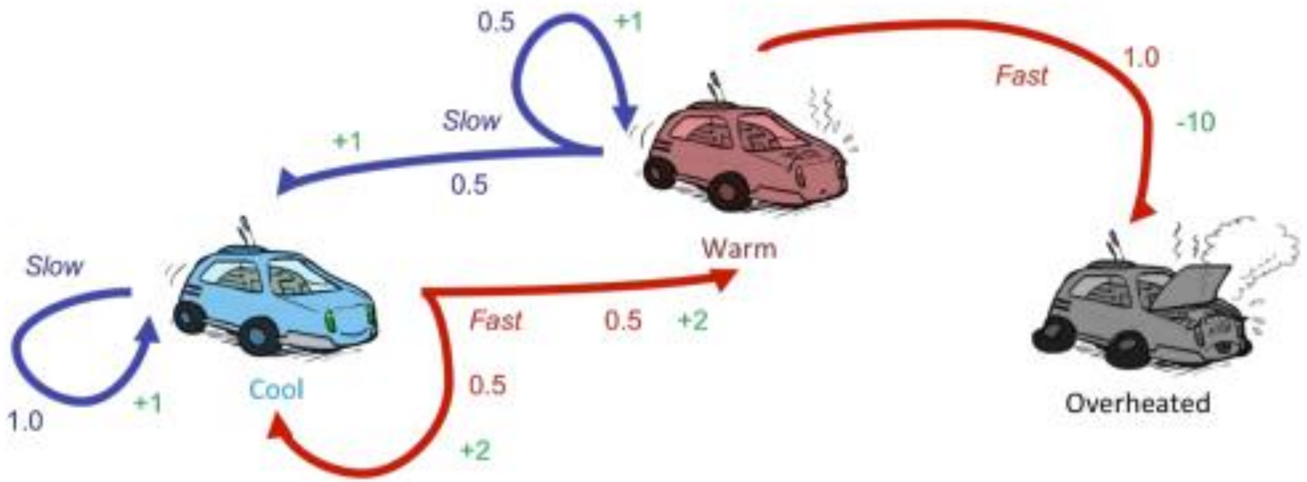
Un ejemplo

- $\mathcal{S} = \{Cool, Warm, Overheated\}$
- $\mathcal{A} = \{Slow, Fast\}$
- $p(s', r|s, a) =$
  - Cool 1 Cool Slow 1
  - Cool 2 Cool Fast 0.5
  - Warm 2 Cool Fast 0.5
  - Oh -10 Warm Fast 1
  - Cool 1 Warm Slow 0.5
  - Warm 1 Warm Slow 0.5
- $\gamma = 1$



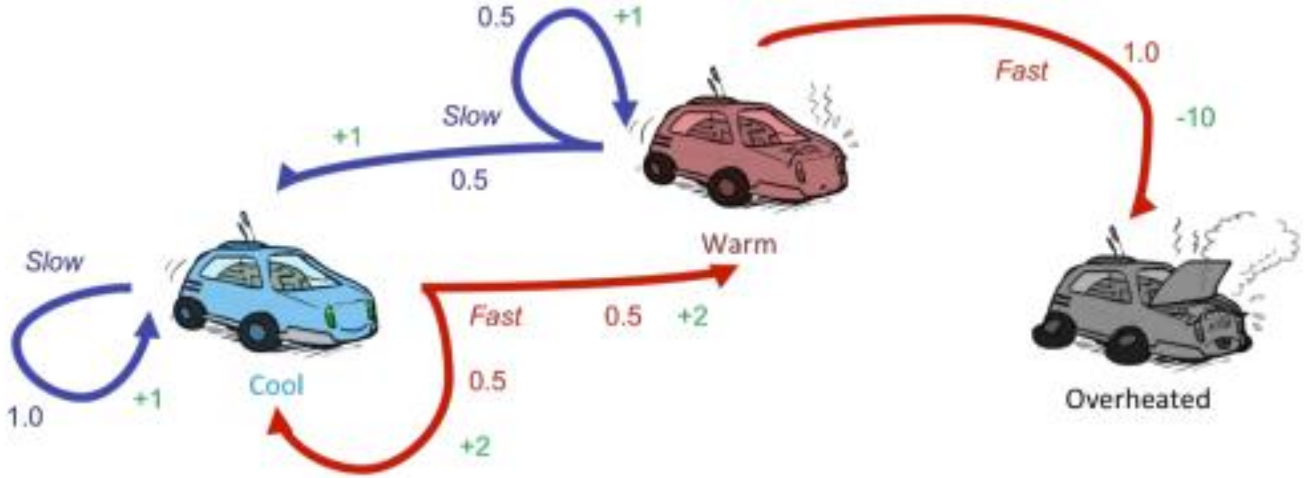
Un ejemplo

- $\mathcal{S} = \{Cool, Warm, Overheated\}$
- $\mathcal{A} = \{Slow, Fast\}$
- $p(s', r|s, a) =$
    - Cool 1 Cool Slow 1
    - Cool 2 Cool Fast 0.5
    - Warm 2 Cool Fast 0.5
    - Oh -10 Warm Fast 1
    - Cool 1 Warm Slow 0.5
    - Warm 1 Warm Slow 0.5
- $\gamma = 1$



Un ejemplo (alternativa)

- $\mathcal{S} = \{Cool, Warm, Overheated\}$
- $\mathcal{A} = \{Slow, Fast\}$
- $p(s'|s,a) = \{$
  - Cool Cool Slow 1
  - Cool Cool Fast 0.5
  - Warm Cool Fast 0.5
  - Oh Warm Fast 1
  - Cool Warm Slow 0.5
  - Warm Warm Slow 0.5}
- $r(s,a) = \{$
  - Cool Slow 1
  - Cool Fast 2
  - Warm Fast -10
  - Cool Slow 1
  - Warm Slow 1}
- $\gamma = 1$



Política

- Una política es un mapeo $\pi: \mathcal{S} \times \mathcal{A} \rightarrow [0,1]$. Para cada estado $s \in \mathcal{S}$ asigna a cada acción $a \in \mathcal{A}$ la probabilidad de tomar a estando en s (denotado $\pi(a|s)$)
- Para políticas deterministas se puede utilizar $a = \pi(s)$

El objetivo

- Encontrar la política que maximice el retorno G_t (esperado)
- El retorno G_t es la recompensa total descontada desde el tiempo t

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{(k=0)}^{\infty} \gamma^k R_{t+k+1}$$

- El descuento $\gamma \in [0, 1]$ es el valor presente de las recompensas futuras
- El valor de recibir la recompensa R después de $k + 1$ pasos es $\gamma^k R$
- Se prefieren recompensas inmediatas a recompensas con retraso
  - γ cercano a 0, es una evaluación miope
  - γ cercano a 1, valora a futuro

Razones para utilizar el descuento

- Matemáticamente conveniente
- Evita ciclos de retornos infinitos
- La incertidumbre del futuro puede no estar completamente representada
- En finanzas tiene relación a tasas de interés y valor del dinero en el tiempo
- Animales y humanos muestran preferencias a recompensas inmediatas
- Si todas las secuencias terminan, es posible utilizar procesos de recompensa de Markov sin descuento

Funciones de valor

- La función de (estado) valor $v_\pi(s)$ asigna el valor a largo plazo del estado s siguiendo la política π

$$v_\pi(s) = \mathbb{E}[G_t | S_t = s, \pi]$$

- La función de (estado) acción

$$q_\pi(s,a) = \mathbb{E}[G_t | S_t = s, A_t = a, \pi]$$

- Nótese que

$$v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a) = \mathbb{E}[q_\pi(S_t, A_t) | S_t = s, \pi], \forall s$$

Funciones de valor óptimas

- La función de valor óptima $v^*(s)$ es la función de valor máxima sobre todas las políticas

$$v^*(s) = \max_{\pi} v_{\pi}(s)$$

- La función de acción óptima $q^*(s)$ es la función de acción máxima sobre todas las políticas

$$q^*(s,a) = \max_{\pi} q_{\pi}(s,a)$$

Política óptima

- Define un orden parcial entre políticas
- $\pi \geq \pi' \leftrightarrow v_{\pi}(s) \geq v_{\pi'}(s), \forall s$
- Para cualquier MDP
  - Existe una política optima π^* tal que $\pi^* \geq \pi, \forall \pi$
  - $v^{\pi *}(s) = v^*(s)$
  - $q^{\pi *}(s, a) = q^*(s, a)$

Encontrando una política óptima

- Es posible encontrar una política óptima maximizando sobre $q^*(s, a)$

$$\pi^*(s, a) = \begin{cases} 1 & si\ a = argmax_{a \in \mathcal{A}} q^*(s, a) \\ 0 & de\ lo\ contrario \end{cases}$$

- Notas
  - Siempre existe una política óptima determinista para cualquier MDP
  - Si conocemos $q^*(s, a)$, conocemos la política óptima
  - Puede haber múltiples políticas óptimas
  - Si existen múltiples acciones que maximizan $q^*(s, \cdot)$, elegimos cualquiera

Función de valor

$$v_\pi(s) = \mathbb{E}[G_t | S_t = s, \pi]$$

- Puede ser definida recursivamente
- $v_\pi(s)$
  - $= \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s, \pi]$
  - $= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t \sim \pi(S_t)]$
  - $= \sum_a \pi(a|s) \sum_r \sum_{s'} p(r, s'|s, a)\big(r + \gamma v_\pi(s')\big)$

Función de acción

$$q_\pi(s,a) = \mathbb{E}[G_t | S_t = s, A_t = a,\ \pi]$$

- Puede ser definida recursivamente
- $q_\pi(s,a)$
  - $= \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1}) | S_t = s, A_t = a]$
  - $= \mathbb{E}[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a]$
  - $\sum_r \sum_{s'} p(r, s' | s, a)(r + \gamma \sum_{a'} \pi(a' | s') q_\pi(s', a'))$
- Nótese que
  - $v_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a) = \mathbb{E}[q_\pi(S_t, A_t) | S_t = s, \pi], \forall s$

Ecuaciones de Bellman (esperadas)

- Dado un MDP, $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, para cualquier política π, las funciones de valor obedecen las siguientes ecuaciones

$$v_\pi(s) = \sum_a \pi(a|s)[r(s,a) + \gamma \sum_{s'} p(s'|a,s) v_\pi(s')]$$

$$q_\pi(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|a,s) \sum_{a'} \pi(a|s) q_\pi(s',a')$$

Ecuaciones de Bellman (óptimas)

- Dado un MDP, $M = (\mathcal{S}, \mathcal{A}, p, r, \gamma)$, para la política óptima π, las funciones de valor obedecen las siguientes ecuaciones

$$v^*(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|a,s) v^*(s') \right]$$

$$q^*(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|a,s) \max_{a' \in \mathcal{A}} q^*(s',a')$$

Los problemas en aprendizaje por refuerzo

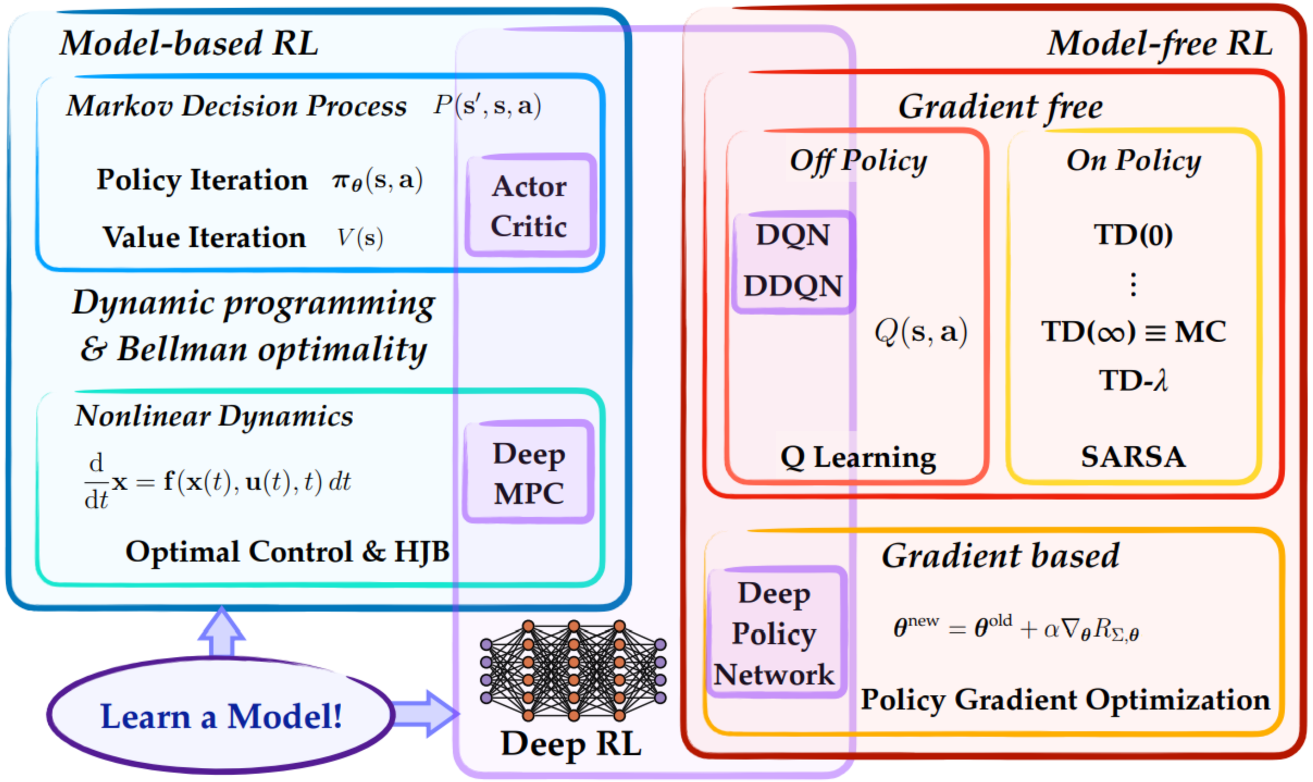
Predicción	Control
• Calcular v_π o q_π (o estimar) • Dada una política, ¿qué tan buena es?	• Calcular v^* o q^* (o estimar) • ¿Cuál es la política óptima?

Ecuación de Bellman (esperado) en forma de matriz

- Dada una política π el problema se reduce a un MRP, entonces
$$\boldsymbol{v} = \boldsymbol{r}^{\pi} + \gamma \boldsymbol{P}^{\pi} \boldsymbol{v}$$
- Donde
  - $v_i = v(s_i)$
  - $r_i^{\pi} = \mathbb{E}[R_{t+1} | S_t = s_i, A_t \sim \pi(S_t)]$
  - $P_{ij}^{\pi} = p(s_j | s_i) = \sum_a \pi(a|s_i) p(s_j|s_i, a)$
- Por tanto, se puede resolver el sistema lineal de ecuaciones

Resolviendo la ecuación de Bellman (optimalidad)

- Es una ecuación no lineal
- En general, no se puede utilizar la solución directa



Programación dinámica

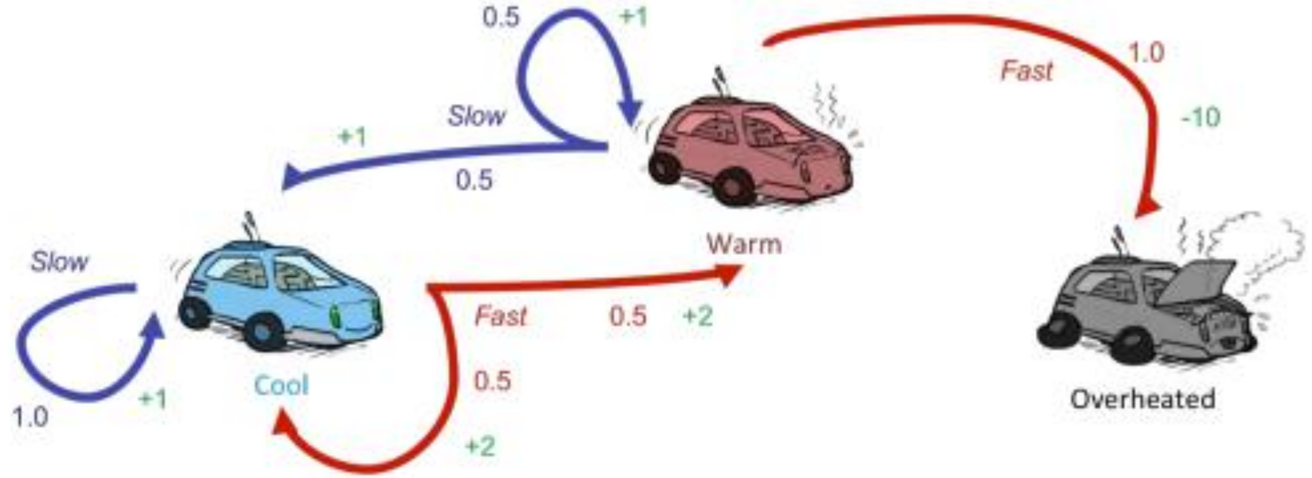
- Se refiere a un conjunto de algoritmos que pueden ser utilizados para encontrar políticas óptimas dado un MDP
- Los métodos consisten en dos partes
  - Evaluación de una política
  - Mejora de una política

Evaluación de una política

- $v_\pi(s) = \mathbb{E}[R_{t+1} + \gamma v_\pi(S_{t+1})|s, \pi]$
- Algoritmo
  - Inicializar v_0
  - Repetir
    - $\forall s \in \mathcal{S}: v_{k+1}(s) \leftarrow \mathbb{E}[R_{t+1} + \gamma v_k(S_{t+1})|s, \pi]$
  - Mientras $v_{k+1}(s) <> v_k(s)$

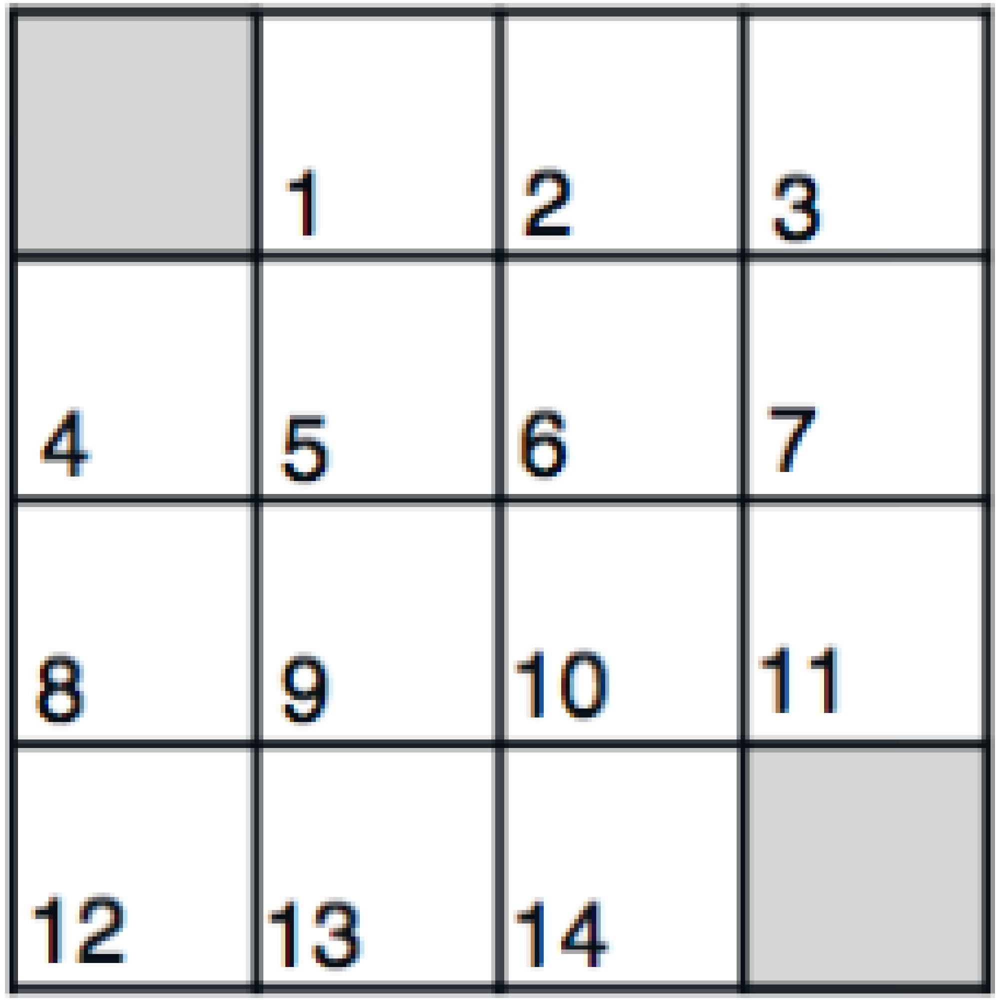
Regresemos al ejemplo

- $\mathcal{S} = \{Cool, Warm, Overheated\}$
- $\mathcal{A} = \{Slow, Fast\}$
- $p(s'|s,a) = \{$
  - Cool Cool Slow 1
  - Cool Cool Fast 0.5
  - Warm Cool Fast 0.5
  - Oh Warm Fast 1
  - Cool Warm Slow 0.5
  - Warm Warm Slow 0.5}
- $r(s,a) = \{$
  - Cool Slow 1
  - Cool Fast 2
  - Warm Fast -10
  - Cool Slow 1
  - Warm Slow 1}
- $\gamma = 0.1$



Tarea 1

- MDPs
  - MDP del gridworld
    - Estados: coordenadas cartesianas
    - Acciones: up, down,left, right
    - Recompensa: -1 por movimiento
    - $\gamma = 1$
  - MDP del juego de gato con rival aleatorio
- Algoritmos
  - Iteración de valor
  - Iteración de política
- Pruebas
  - Aplicar ambos algoritmos a los dos MDPs
  - Para cada algoritmo mostrar $v^*(s)$ y $\pi^*(a|s)$
- Fecha de entrega: 4/03/2022



Para la otra vez…

- Implementación de la evaluación de una politica
- Iteración de valor
- Iteración de política