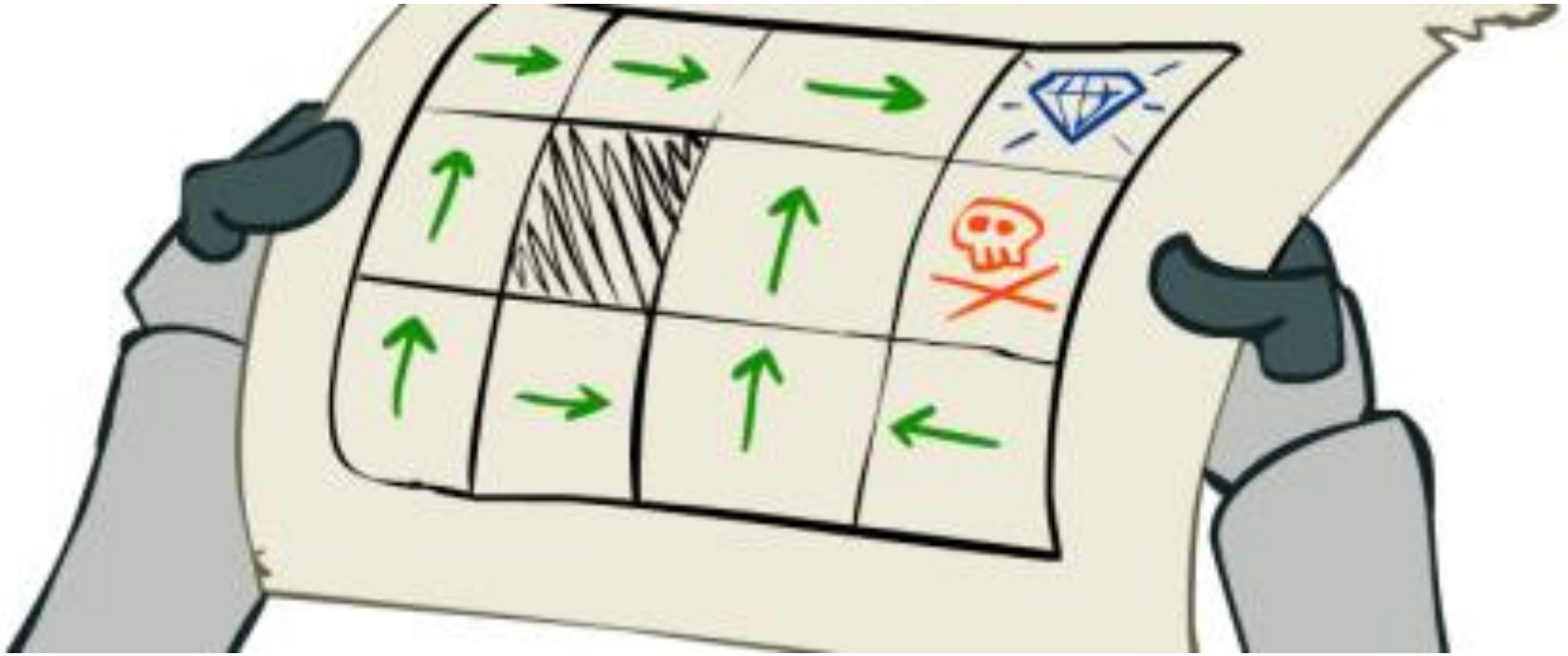
Tercera Parte:
Clase 2 – Métodos eficientes para MDPs con horizonte infinito

Optimización Dinámica - ICS



Mathias Klapp

¿Cómo vamos?

- Formalizamos el MDP con horizonte infinito

- Para el criterio retorno total descontado estudiamos:

    - Evaluar value-to-go de política estacionaria (forma exacta y numérica).

    - Optimizar value-to-go (forma exacta y numérica).

    - Existencia de política óptima estacionaria.

Métodos eficientes de solución:

- Iteración de valor
- Iteración de política
- Método de programación lineal

Iteración de valor (value iteration):

1. Input: $V^{(0)}$ (cualquier vector), ε (precisión)

2. Resolver para cada $s \in \mathbb{S}$:

$$V^{(n+1)}(s) = \max_{x \in \mathbb{X}(s)} \left\{ r(s,x) + \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V^{(n)}(j) \right\}$$

3. Si $\left\| V^{(n+1)} - V^{(n)} \right\| > \frac{\varepsilon \cdot (1-\lambda)}{2\lambda}$: ir a 2 (seguir iterando).

4. Retornar política ε-óptima para cada $s \in \mathbb{S}$:

$$d_\varepsilon(s) \in \operatorname{argmax}_{x \in \mathbb{X}(s)} \left\{ r(s,x) + \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V^{(n+1)}(j) \right\}$$

El algoritmo aplica directamente la contracción L^*. Garantías:

- $\left\| V^{d_\varepsilon} - V^{(n+1)} \right\| \leq \varepsilon/2$
- $\left\| V^{d_\varepsilon} - V^* \right\| \leq \varepsilon$

Métodos eficientes de solución:

- Iteración de valor
- Iteración de política
- Método de programación lineal

Iteración de política (*policy iteration*)

1. Input: d_0 (cualquier política de decisión)
2. Calcular $V^{(n)} = V^{d_n}$
3. Escoger nueva política:

$$d_{n+1}(s) = \underset{x \in \mathbb{X}(s)}{\operatorname{argmax}} \left\{ r(s,x) + \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V^{(n)}(j) \right\}, \forall s \in \mathbb{S}$$

   - Anticiclaje: Privilegiar *status quo* si existe solución múltiple, *i.e.* $d_{n+1}(s) = d_n(s)$.

4. Si $d_{n+1} \neq d_n$, ir a 2 (seguir iterando).
5. Retornar d_n

- Algorimo ``inteligente'' que evalúa cada política. Garantía:
  - $V^{(n+1)} \geq V^{(n)}$
  - Si no termina, entonces existe estado s: $V^{(n+1)}(s) > V^{(n)}(s)$
- Si $\mathbb{X}$ es finito: termina en un número finito de iteraciones.

Iteración de valor vs política

Iteración de valor:

- Simple y rápido por iteración.
- No necesariamente converge en tiempo finito.
- Resultados aproximados (método numérico).

Iteración de política:

- Más complejo por iteración.
- Converge en tiempo finito y tiende a hacer menos iteraciones.
- Va mejorando.
- En cada momento posee política y *value-to-go* factibles.

Ejemplo

Dos estados: s_1 y s_2

Acción $x \in [0,2]$ sólo se toma en s_1.

Retornos: $r(s_1, x) = -x^2$ $r(s_2) = -0{,}5$

Probabilidades:

$p(s_1|s_1, x) = 0{,}5x$ $p(s_2|s_1, x) = 1 - 0{,}5x$

$p(s_1|s_2) = 0$ $p(s_2|s_2) = 1$

Ejemplo

Ecuaciones de Bellman:

$$V_1 = \max_{x \in [0,2]} \{-x^2 + \lambda(0{,}5xV_1 + (1 - 0{,}5x)V_2)\}$$

$$V_2 = -0{,}5 + \lambda V_2$$

En este caso es trivial obtener $V_2 = -\frac{0{,}5}{1-\lambda}$.

Luego: $V_1 = \max_{x \in [0,2]} \left\{-x^2 + \lambda \left(0{,}5xV_1 - (1 - 0{,}5x)\frac{0{,}5}{1-\lambda}\right)\right\}$

Iteraciones ($\lambda = 0{,}9$)

Policy Iteration			
n	$d_n(s_1)$	$v^n(s_1)$	$v^*_\lambda(s_1) - v^n(s_1)$
0	0.000000000000000	−4.500000000000000	1.33×10^{-02}
1	0.112500000000000	−4.486668861092825	9.49×10^{-06}
2	0.115499506254114	−4.486659370799152	4.81×10^{-12}
3	0.115501641570191	−4.486659370794342	0[a]
4	0.115501641571273	−4.486659370794342	0

Value Iteration			
n	$d_n(s_1)$	$v^n(s_1)$	$v^*_\lambda(s_1) - v^n(s_1)$
0	[b]	−4.500000000000000	1.33×10^{-02}
1	0.112500000000000	−4.487343750000001	6.84×10^{-04}
2	0.115347656250000	−4.486694918197633	3.55×10^{-05}
3	0.115493643405533	−4.486661218332917	1.85×10^{-06}
4	0.115501225875094	−4.486659466821352	9.60×10^{-08}
5	0.115501619965196	−4.486659375785417	4.99×10^{-09}
6	0.115501640448281	−4.486659371053757	2.59×10^{-10}
7	0.115501641512905	−4.486659370807826	1.35×10^{-11}
8	0.115501641568239	−4.486659370795043	7.01×10^{-13}
9	0.115501641571116	−4.486659370794379	3.64×10^{-14}
10	0.115501641571265	−4.486659370794344	1.78×10^{-15}
11	0.115501641571273	−4.486659370794342	0
12	0.115501641571273	−4.486659370794342	0

[a] Zero represents less than 10^{-16}.

[b] Value iteration was initiated with $v^0(s_1) = -4.5$, the value obtained from the first iteration of policy iteration algorithm.

Métodos eficientes de solución:

- Iteración de valor
- Iteración de política
- Método de programación lineal

Método de programación lineal (LP)

- Adapta lo conocido para resolver un MDP. Explota maquinaria existente para LP.
- En alza debido al avance de programación lineal y métodos de descomposición (generación de columna y Benders).

- Idea: Resolver para cada $s \in \mathbb{S}$:

$$V(s) = \max_{x \in \mathbb{X}(s)} \left\{ r(s,x) + \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V(j) \right\}$$

equivale a buscar el menor valor de cada $V(s)$ tal que:

$$V(s) \geq r(s,x) + \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V(j)\,, \qquad \forall x \in \mathbb{X}(s)$$

Método primal de LP

1. Escoger $\alpha_s > 0$ para cada $s \in \mathbb{S}$ (vector de 1's por ejemplo)
2. Resolver:

$$\min_V \alpha^T V \qquad \text{s.t.}$$
$$V(s) - \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V(j) \geq r(s,x), \quad \forall x \in \mathbb{X}(s), \forall s \in \mathbb{S}$$

Garantía:

- La solución óptima del LP es V^*
- Para cada $s \in \mathbb{S}$, hay al menos un x con la restricción activa.

Método dual de LP

Primal:

$$\min_{V} \alpha^T V \qquad \text{s.t.}$$
$$V(s) - \lambda \sum_{j \in \mathbb{S}} p(j|s,x) V(j) \geq r(s,x), \quad \forall x \in \mathbb{X}(s), \forall s \in \mathbb{S}$$

Dual:

$$\max_{f \geq 0} \sum_{s \in \mathbb{S}} \sum_{x \in \mathbb{X}(s)} r(s,x)\, f_{s,x} \qquad \text{s.t.}$$
$$\sum_{x \in \mathbb{X}(j)} f_{j,x} - \sum_{s \in \mathbb{S}} \sum_{y \in \mathbb{X}(s)} \lambda p(j|s,y) f_{s,y} = \alpha_j, \quad \forall j \in \mathbb{S}$$

- Se puede usar generación dinámica de filas / columnas.

Relación con problema de rutas

$$\max_{f \geq 0} \sum_{s \in \mathbb{S}} \sum_{x \in \mathbb{X}(s)} r(s,x)\, f_{s,x} \qquad \text{s.t.}$$

$$\sum_{x \in \mathbb{X}(j)} f_{j,x} - \sum_{s \in \mathbb{S}} \sum_{y \in \mathbb{X}(s)} \lambda p(j|s,y) f_{s,y} = \alpha_j, \quad \forall j \in \mathbb{S}$$

El LP dual se puede interpretar como un problema generalizado de flujo.

En cada nodo (estado) j se generan α_j unidad de flujo.

Flujo entrante se diluye por λ al pasar por j.

Interpretación de las decisiones

$$\max_{f \geq 0} \sum_{s \in \mathbb{S}} \sum_{x \in \mathbb{X}(s)} r(s,x)\, f_{s,x} \qquad \text{s.t.}$$

$$\sum_{x \in \mathbb{X}(j)} f_{j,x} - \sum_{s \in \mathbb{S}} \sum_{y \in \mathbb{X}(s)} \lambda p(j|s,y) f_{s,y} = \alpha_j, \quad \forall j \in \mathbb{S}$$

Teorema:

Considere una regla de decisión d randomizada y defina:

$$f_{s,x}^{d} = \sum_{j \in \mathbb{S}} \alpha_j \left(\sum_{t=1}^{\infty} \lambda^{t-1} \mathbb{P}^{d}(s_n = s, x_n = x | s_1 = j) \right), \forall s \in \mathbb{S}, \forall x \in \mathbb{X}(s)$$

, entonces el vector f^d es factible para el LP

Demostración: Reemplazar y ver…..

pag 225 libro de Puterman

Relación del LP con decisiones

$$\max_{f \geq 0} \sum_{s \in \mathbb{S}} \sum_{x \in \mathbb{X}(s)} r(s,x)\, f_{s,x} \qquad \text{s.t.}$$

$$\sum_{x \in \mathbb{X}(j)} f_{j,x} - \sum_{s \in \mathbb{S}} \sum_{y \in \mathbb{X}(s)} \lambda p(j|s,y) f_{s,y} = \alpha_j, \quad \forall j \in \mathbb{S}$$

Teorema 2:

Suponga que f es una solución factible del LP, entonces para cada estado $s \in \mathbb{S}$: $\sum_{x \in \mathbb{X}(s)} f_{s,x} > 0$ defina una política d markoviana randomizada dada por:

$$\mathbb{P}\left(d^f(s) = x\right) = \frac{f_{x,s}}{\sum_{x \in \mathbb{X}(s)} f_{s,x}}$$

Se cumple que $f^{d^f} = f$.

Relación 1 a 1 entre soluciones factibles y reglas de decisión randomizadas.

Relación del LP con decisiones

$$\max_{f \geq 0} \sum_{s \in \mathbb{S}} \sum_{x \in \mathbb{X}(s)} r(s,x)\, f_{s,x} \qquad \text{s.t.}$$

$$\sum_{x \in \mathbb{X}(j)} f_{j,x} - \sum_{s \in \mathbb{S}} \sum_{y \in \mathbb{X}(s)} \lambda p(j|s,y) f_{s,y} = \alpha_j, \quad \forall j \in \mathbb{S}$$

Teorema 3: Relación 1 a 1 entre reglas determinísticas y vértices

- Si f es una solución básica factible del LP, entonces d^f es una política estacionaria determinítica.
- Si d es una regla determinística, entonces f^d es una solución básica factible del LP

- Implica que el LP posee una solución óptima f^* en un vértice que define una política d^*estacionaria determinística.

Relación del LP con decisiones

$$\max_{f \geq 0} \sum_{s \in \mathbb{S}} \sum_{x \in \mathbb{X}(s)} r(s,x)\, f_{s,x} \qquad \text{s.t.}$$

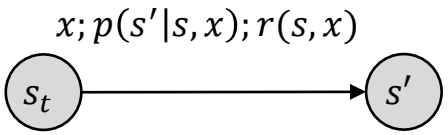
$$\sum_{x \in \mathbb{X}(j)} f_{j,x} - \sum_{s \in \mathbb{S}} \sum_{y \in \mathbb{X}(s)} \lambda p(j|s,y) f_{s,y} = \alpha_j, \quad \forall j \in \mathbb{S}$$

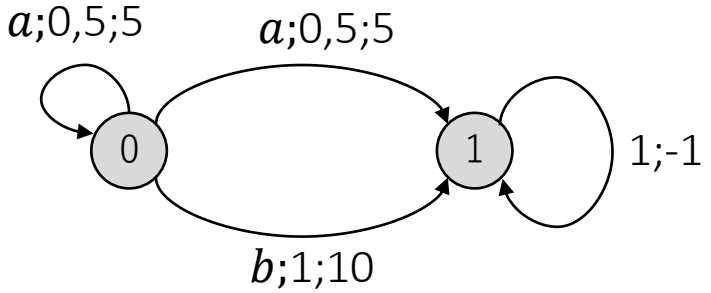
Teorema 4: Escoger α es indiferente

- Para cualquier $\alpha > 0$ el LP tiene la misma solución óptima.

Ejemplo ($\lambda = 0{,}8$)

Resolvamos:





$$\max_{f \geq 0} \sum_{a \in \mathbb{S}} \sum_{x \in \mathbb{X}(s)} r(s,x)\, f_{s,x} \qquad \text{s.t.}$$

$$\sum_{x \in \mathbb{X}(j)} f_{j,x} - \sum_{s \in \mathbb{S}} \sum_{y \in \mathbb{X}(s)} \lambda p(j|s,y) f_{s,y} = 1,\ \forall j \in \mathbb{S}$$

$$\max_{f \geq 0} \quad 5f_{0,a} + 10f_{0,b} - f_1 \qquad \text{s.t.}$$

$$f_{0,a} + f_{0,b} - \lambda\left(0{,}5 f_{0,a}\right) = 1$$

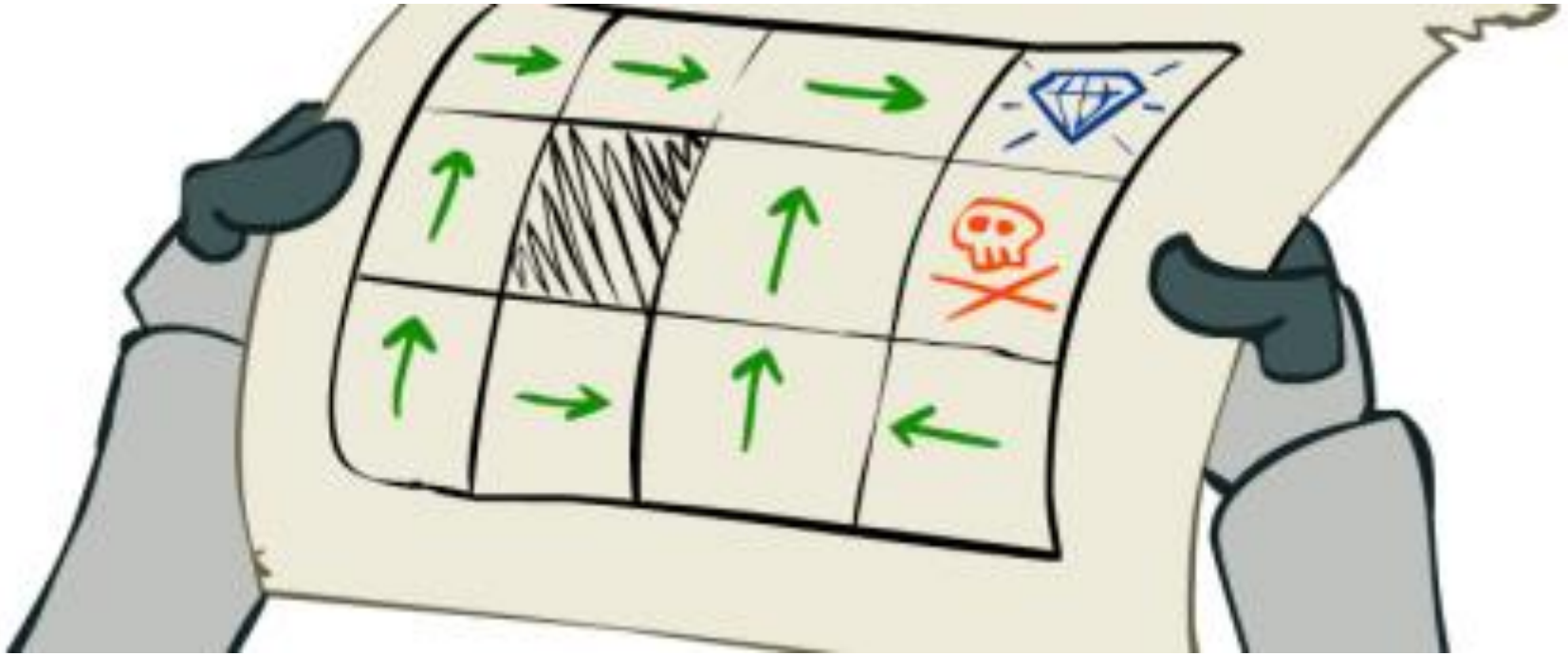
$$f_1 - \lambda\left(0{,}5 f_{0,a} + f_{0,b} + f_1\right) = 1$$

Políticas óptimas monotónicas

- Aplican los mismos resultados derivados para horizonte finito.
- Basta demostrar superaditividad del value-to-go y del retorno.
- Iteración de política y LP se pueden modificar para explotarlo

Tercera Parte:
Clase 2 –Políticas para MDP de Horizonte Infinito

Optimización Dinámica - ICS



Mathias Klapp