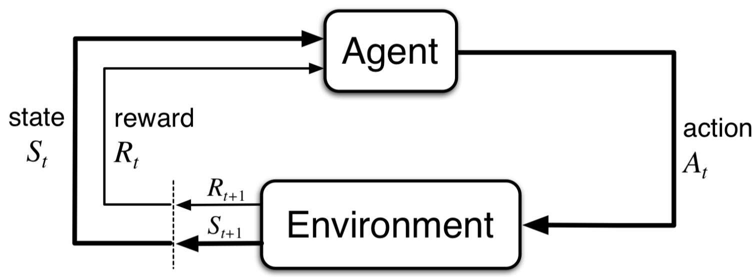
Universidade Federal de Viçosa
Centro de Ciências Exatas e Tecnológicas
Departamento de Informática

Processos de Decisão de Markov
Levi H. S. Lelis
levi.lelis@ufv.br

O Problema de Aprendizagem por Reforço

No problema de aprendizagem por reforço um agente interage com o ambiente, aprendendo a como maximizar o seu ganho de recompensa acumulada, no longo prazo.



A convenção que iremos adotar é a de que o agente escolhe uma ação A_t no estado S_t e o ambiente devolve uma recompensa e o próximo estado, R_{t+1} e S_{t+1}, respectivamente.

Modelo de Decisão de Markov (MDP)

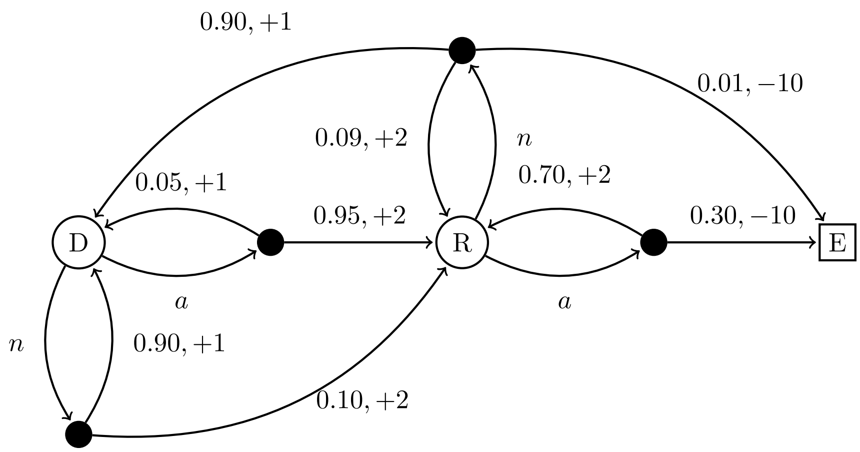
Considere o problema abaixo onde os círculos denotam estados de um carro: devagar (D), rápido (R) e estragado (E). O agente pode escolher acelerar (a) or viajar em uma velocidade normal (n). Os nós sólidos denotam o modelo de transição do ambiente. Por exemplo, se o agente escolhe viajar em velocidade normal no estado D, ele continua no estado D com probabilidade 0.90 e recebe recompensa de +1; com probabilidade 0.10 ele vai para o estado R e recebe recompensa de +2. O estado E, representado por um quadrado, denota um estado terminal. Estados terminais são equivalentes a estados não terminais em que todas as ações levam o agente para o mesmo estado terminal e retornam recompensa de 0. Um **episódio** do problema termina quando o agente atinge o estado E obtendo -10 na transição.

Um agente pode observar diferentes episódios do problema. Por exemplo, se o agente começa em D:

- $D, n, +1, D, n, +1, D, n, +2, R, a, +2, R, n, -10, E$

- $D, a, +2, R, a, -10, E$

- $D, a, +1, D, a, +2, R, a, -10, E$



Seja S o conjunto de estados, R o conjunto de recompensas, e $A(s)$ as ações disponíveis em s, nós temos que a dinâmica do modelo acima é governada pela probabilidade de transição dado um estado e uma ação.

$$p(s', r|s, a) = Pr(S_{t+1} = s', R_{t+1} = r|S_t = s, A_t = a)\,,$$

para todo $s, s' \in S$, $r \in R$ e $a \in A(s)$.

O valor de $p(s', r|s, a)$ denota uma distribuição válida de probabilidade. Isto é, para um dado estado e ação, a soma das probabilidades para os possíveis estados e recompensas retornadas tem que ser igual a 1.

$$\sum_{s' \in S} \sum_{r \in R} p(s', r|s, a) = 1$$

No problema acima, a probabilidade de atingirmos um estado s' e uma recompensa r depende apenas do estado atual s e a ação a tomada em s. Essa probabilidade de transição não depende do histórico do agente no mundo, mas apenas do estado atual. Essa propriedade é conhecida como a propriedade de Markov.

Definição 1 (Propriedade de Markov) *Um problema satisfaz a propriedade de Markov se cada um de seus estados armazena todas as informações relevantes para decidir as interações futuras do agente com o ambiente.*

Os algoritmos estudados nesse disciplina assumem a propriedade de Markov.

Objetivos e Recompensas

No início da disciplina vimos algoritmos de busca heurística, como o A*. Tais algoritmos usam uma função heurística para alcançar o seu **objetivo**. Em algoritmos de aprendizagem por reforço nós não temos a definição um objetivo, mas de uma função de recompensa. Assim assumimos a hipótese da recompensa.

Definição 2 *Objetivos e propósitos dos agentes podem ser expressos em termos do acúmulo da recompensa em longo prazo.*

Exemplos:

1. O problema do caminho mais curto que resolvemos com o A* pode ser modelado através de recompensas onde o agente recebe a recompensa de -1 para cada passo e $+1$ quando atinge o objetivo. Para maximizar sua recompensa o agente deve resolver o problema do caminho mais curto.

2. Em Xadrez, Go e Damas a recompensa pode ser 0 para todos os movimentos durante o jogo e $+1, 0, -1$ para estados terminais do jogo representando vitória, empate e derrota do jogador.

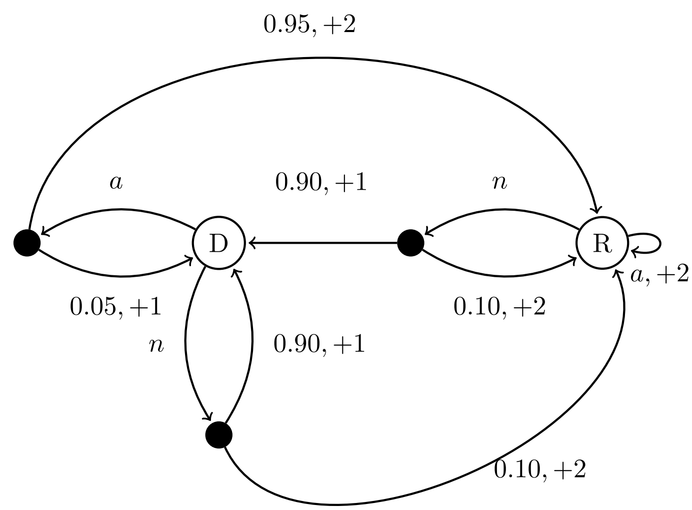
Retornos e Recompensas

No problema do carro acima o retorno de um episódio é dado pela soma das recompensas ao longo do episódio. De uma forma geral, o retorno após o passo de tempo t é dado por:

$$G_t = R_{t+1} + R_{t+2} + \cdots + R_T\,,$$

onde T é o passo final.

Essa formulação matemática funciona bem quando o problema é episódico (eventualmente termina). Mas passamos a ter dificuldades quando o problema é contínuo. Considere, por exemplo, a versão modificada abaixo do problema dos carros.



Nessa versão do problema do carro o agente interage com o ambiente indefinidamente. Se o nosso objetivo é derivar formas de agir no ambiente de forma a maximizar o retorno no longo prazo, o agente que só escolhe n terá o mesmo retorno que o agente que só escolhe a, que é ∞. Mas claramente a ação a dá um retorno maior para o agente.

Para resolver o problema de conseguirmos diferenciar formas de comportar diferente em problemas não-episódicos, utilizamos um fator de desconto $0 \leq \gamma \leq 1$ para as recompensas de longo prazo.

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}\,,$$

Quando o valor de $\gamma < 1$ a série não vai para ∞ se os valores de R são limitados.

- Quanto maior o valor de γ, mais importância daremos para retornos futuros.

- Quanto menor o valor de γ, mais importância daremos para os retornos imediatos.

O fator de desconto γ pode ser aplicado em problemas episódicos também. No inicial exemplo dos carros teríamos os seguintes valores de G_t com $\gamma = 0.5$ para cada amostra.

- Amostra: $D, n, +1, D, n, +1, D, n, +2, R, a, +2, R, n, -10, E$.
  - Recompensa: $1 + 0.5 \cdot 1 + 0.5^2 \cdot 2 + 0.5^3 \cdot 2 + 0.5^4 \cdot (-10) = 1.625$
- $D, a, +2, R, a, -10, E$
  - Recompensa: $2 + 0.5 \cdot (-10) = -3$
- $D, a, +1, D, a, +2, R, a, -10, E$
  - Recompensa: $1 + 0.5 \cdot 2 + 0.5^2 \cdot (-10) = -0.5$

Políticas e Funções de Valor

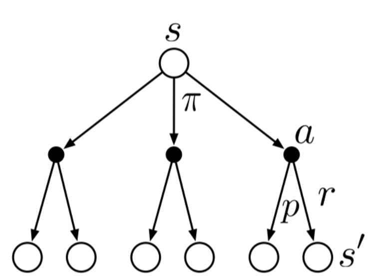
O que define o comportamento do agente é o que chamamos de **política**. Uma política $\pi(a|s)$ denota a probabilidade na qual um agente toma a ação a dado que ele está em um estado s. Em aprendizagem por reforço nós estamos interessados em encontrar políticas que maximizam o ganho de recompensa do agente.

Definição 3 *Uma política $\pi(a|s)$ denota a probabilidade com que o agente toma a decisão a no estado s.*

A função de valor de estado $v_\pi(s)$ para uma política π retorna o valor esperado de retorno G começando a partir do estado s.

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t | S_t = s] \\ &= \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a)[r + \gamma v_\pi(s')]\,, \text{ para todo } s \in S\,. \end{aligned}$$

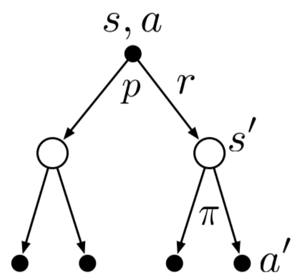
A segunda equação acima é a equação de Bellman para $v_\pi(s)$. Graficamente podemos representar $v_\pi(s)$ com o diagrama de *backup*.



A função de valor de ação $q_\pi(s, a)$ para uma política π retorna o valor esperado G começando a partir de s e escolhendo ação a.

$$\begin{aligned} q_\pi(s, a) &= \mathbb{E}[G_t | S_t = s, A_t = a] \\ &= \sum_{s',r} p(s', r | s, a)[r + \gamma \sum_{a \in A(s')} \pi(a'|s') q_\pi(s', a')]\,, \text{ para todo } s \in S \text{ e } a \in A(s)\,. \end{aligned}$$

A segunda equação acima é a equação de Bellman para a função de valor de ação, que é representada graficamente através do seguinte diagrama.



Exemplo: Considere a política $\pi(a|s) = 0.5$ para todos as ações e estados no problema episódico dos carros e $\gamma = 0.99$. O valor de $v_\pi(D) = 8.43$, $v_\pi(R) = 5.65$ e $v_\pi(E) = 0$. Vamos verificar esses valores?

$$\begin{aligned}
v_\pi(D) &= 0.5[0.9(1 + 0.99 \cdot 8.43) + 0.1(2 + 0.99 \cdot 5.65)] \\
&\quad + 0.5[0.95(2 + 0.99 \cdot 5.65) + 0.05(1 + 0.99 \cdot 8.43] = 8.4 \\
v_\pi(R) &= 0.5[0.9(1 + 0.99 \cdot 8.43) + 0.09(2 + 0.99 \cdot 5.65) + 0.01(-10 + 0.99 \cdot 0)] \\
&\quad + 0.5[0.7(2 + 0.99 \cdot 8.43) + 0.3(-10 + 0.99 \cdot 0)] = 5.6
\end{aligned}$$

Políticas e Funções de Valor Ótimas

Solucionar um problema de aprendizagem por reforço é encontrar um política ótima – que atinge o maior valor esperado de recompensa. A função de valor ótima é denotada por,

$$v_*(s) = \max_\pi v_\pi(s), \text{ para todo } s \in S.$$

O valor de ação de uma política ótima é denotado por,

$$q_*(s, a) = \max_\pi q_\pi(s, a)\,.$$

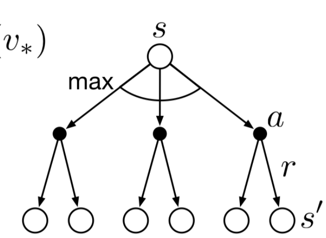
A formulação da equação de Bellman para as funções $v_*(s)$ e $q_*(s, a)$ são as seguintes:

$$v_*(s) = \max_{a \in A(s)} \sum_{s', r} p(s', r|s, a)[r + \gamma v_*(s')]$$

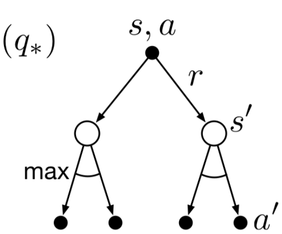
$$q_*(s, a) = \sum_{s', r} p(s', r|s, a)[r + \gamma \max_{a' \in A(s')} q_*(s', a')]$$

O diagrama de *backup* de v_* e q_* são denotados como exibido na figura abaixo.

(v_*)



(q_*)



As curvas cortando as ações nos diagramas indicam onde está o operador max nas equações.

Exemplos. Para o problema episódico do carro nós temos os seguintes valores de v_* e q_* para $\gamma = 0.99$:

$$\begin{aligned} v_*(D) &= 99.196 \\ v_*(R) &= 98.105 \\ v_*(E) &= 0 \end{aligned}$$

$$\begin{aligned} \pi(n|D) &= 1.0 \\ \pi(a|D) &= 0.0 \\ \pi(n|R) &= 1.0 \\ \pi(a|R) &= 0.0 \end{aligned}$$

Se alteramos o valor de γ para 0.9 os valores ótimos passam a ser:

$$\begin{aligned} v_*(D) &= 14.208 \\ v_*(R) &= 13.589 \\ v_*(E) &= 0 \end{aligned}$$

$$\begin{aligned} \pi(n|D) &= 0.0 \\ \pi(a|D) &= 1.0 \\ \pi(n|R) &= 1.0 \\ \pi(a|R) &= 0.0 \end{aligned}$$

Para o problema não-episódico do carro e $\gamma = 0.99$ os valores ótimos são:

$$\begin{aligned} v_*(D) &= 199.945 \\ v_*(R) &= 199.997 \\ v_*(E) &= 0 \end{aligned}$$

$$\begin{aligned} \pi(n|D) &= 0.0 \\ \pi(a|D) &= 1.0 \\ \pi(n|R) &= 0.0 \\ \pi(a|R) &= 1.0 \end{aligned}$$