INSTITUTO MAUÁ DE TECNOLOGIA

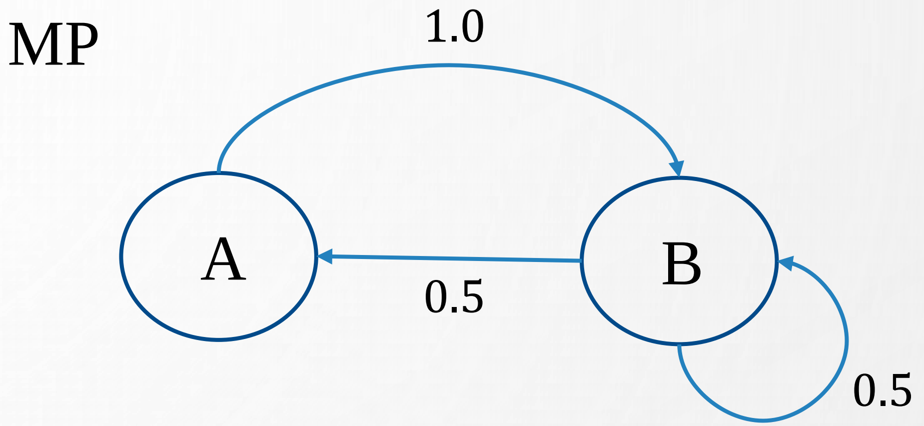
APRENDIZADO POR REFORÇO

Aula 2: Processos de Decisão de Markov (MDPs)

Lucas Pereira Cotrim lucas.cotrim@maua.br
Marcos Menon José marcos.jose@maua.br

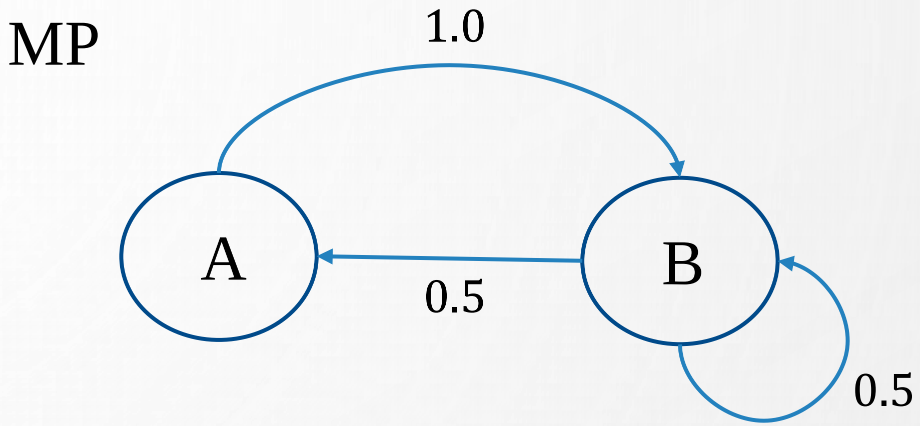
TÓPICOS DA AULA

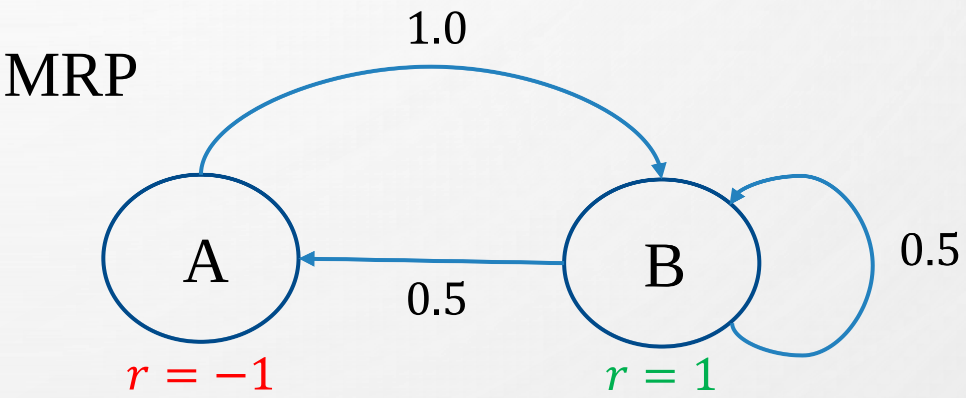
- Cadeias de Markov (MPs)



TÓPICOS DA AULA

- Cadeias de Markov (MPs)
- Processos de Recompensa de Markov (MRPs)

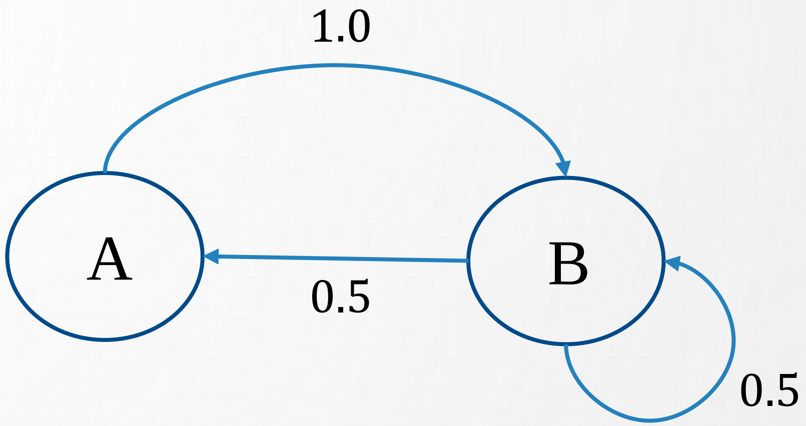




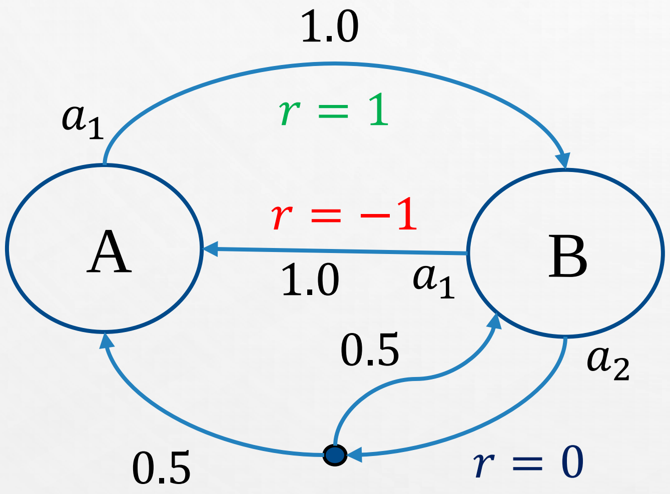
TÓPICOS DA AULA

- Cadeias de Markov (MPs)
- Processos de Recompensa de Markov (MRPs)
- Processos de Decisão de Markov (MDPs)
- Extensões de MDPs

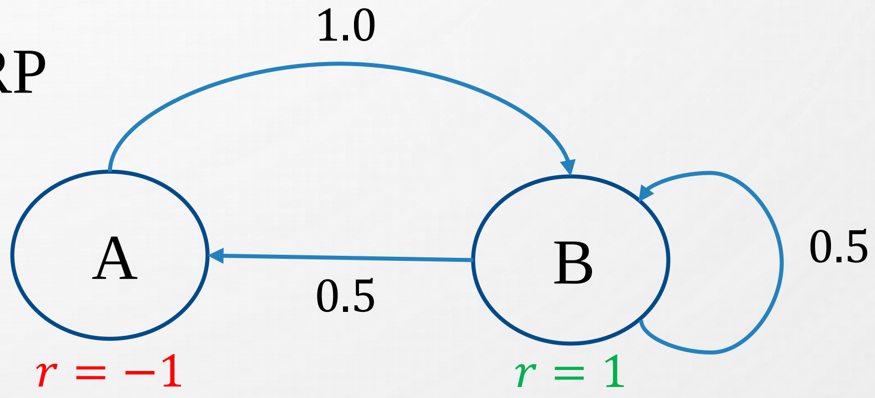
MP



MDP



MRP

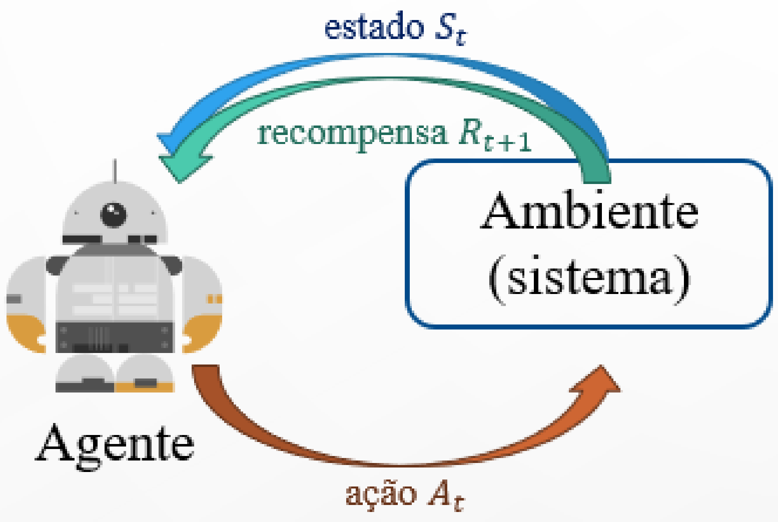


MDPs: Sutton & Barto Cp. 3

RELEMBRANDO AULA 1

TOMADA DE DECISÕES SEQUENCIAL

$$S_0 \xrightarrow[R_1]{A_0} S_1 \xrightarrow[R_2]{A_1} \dots$$



A cada instante de tempo o agente:

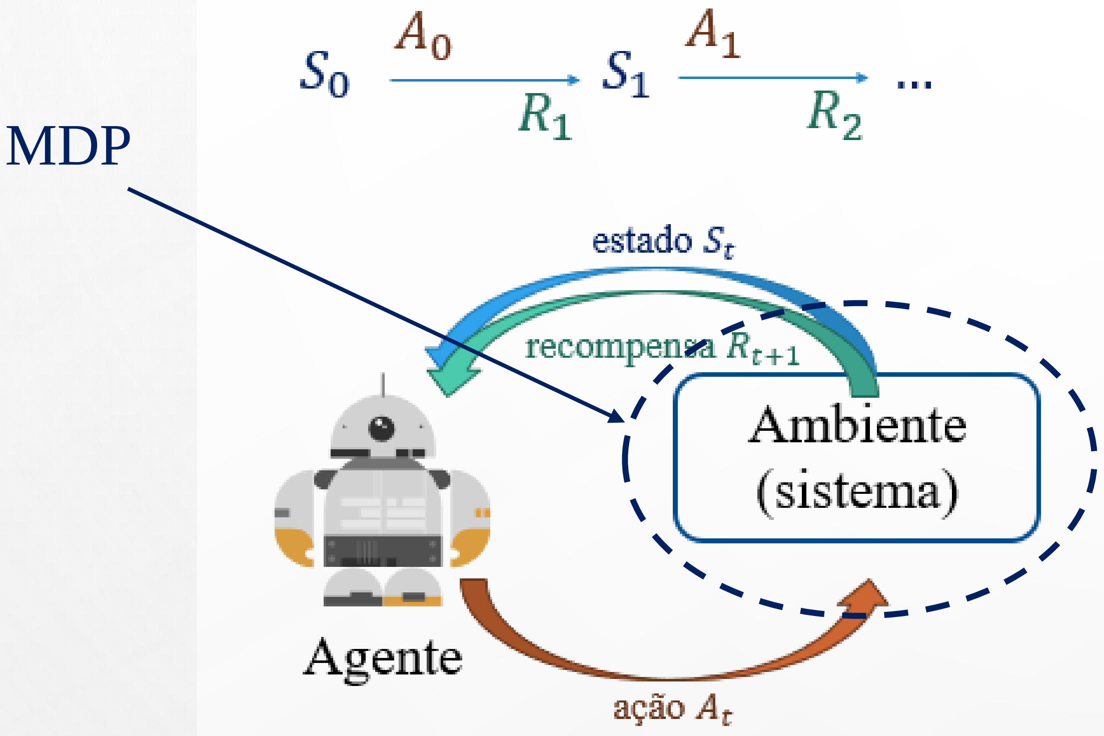
- Observa o estado do ambiente.
- Escolhe e executa uma ação.
- Recebe uma recompensa imediata.
- O sistema evoluí para um novo estado.

O processo é então repetido.

Tempo discreto: Decisões são tomadas somente em épocas de decisão $t \in \{0,1,\dots,N\}$

RELEMBRANDO AULA 1

TOMADA DE DECISÕES SEQUENCIAL



A cada instante de tempo o agente:

- Observa o estado do ambiente.
- Escolhe e executa uma ação.
- Recebe uma recompensa imediata.
- O sistema evoluí para um novo estado.

O processo é então repetido.

Tempo discreto: Decisões são tomadas somente em épocas de decisão

$t \in \{0,1,\dots,N\}$

INTRODUÇÃO A PROCESSOS DE DECISÃO DE MARKOV (MDP)

Um **Processo de Decisão de Markov (MDP)** é uma representação formal de um ambiente completamente observável para Aprendizado por Reforço.

- A maioria dos problemas de Aprendizado por Reforço pode ser formulada como um MDP.
- Treinar um agente de Aprendizado por Reforço busca resolver o MDP associado ao ambiente para obter a política ótima π^*.
- Por que estudar Cadeias de Markov (MPs) e Processos de Recompensa de Markov (MRPs)?

  - Dada uma política de ações, MDPs podem ser convertidos em MRPs.

  - A avaliação de políticas de ações é feita em MRPs.

PROCESSO DE MARKOV (MP)

Processo de Markov (MP)

PROPRIEDADE DE MARKOV

Um estado de **Markov** S_t contém toda a informação útil da história H_t:

- Um estado S_t é de Markov se, e somente se, satisfaz a **propriedade de Markov**:

$$\mathbb{P}(S_{t+1}|S_t) = \mathbb{P}(S_{t+1}|S_0, \dots, S_t)$$

- Ou seja, o estado futuro independe de estados passados dado o estado atual.
- O estado S_t é uma estatística suficiente do futuro.
- Um agente ótimo pode tomar decisões com base apenas em S_t, sem a necessidade de conhecer como o estado S_t foi alcançado.

Em um MDP completamente observável, o estado $S_t = S_t^a = S_t^e$ é de Markov.

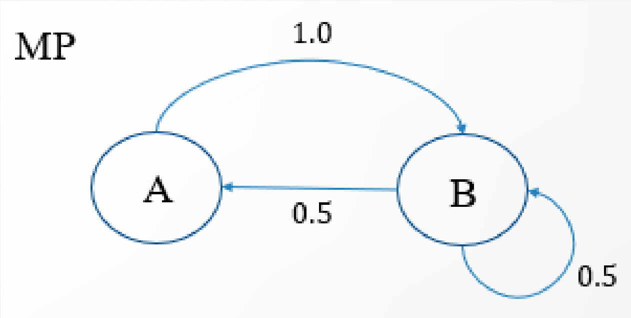
PROCESSO DE MARKOV (MP)

Uma **Cadeia de Markov** (ou Processo de Markov) é uma tupla $< \mathcal{S}, \mathcal{P} >$, onde:

- $\mathcal{S}$ é um conjunto finito de estados.
- $\mathcal{P}$ é uma função $\mathcal{P}$: $\mathcal{S} x \mathcal{S} \rightarrow [0,1] \subset \mathbb{R}$ de probabilidades de transições de estados.

A existência de uma função de probabilidades de transições de estados decorre da Propriedade de Markov: $\mathbb{P}(S_{t+1}|S_t) = \mathbb{P}(S_{t+1}|S_0, \dots, S_t)$

Dada uma Cadeia de Markov é possível amostrar sequências de estados $S_0, S_1, S_2, \dots$



MATRIZ DE TRANSIÇÃO DE ESTADOS

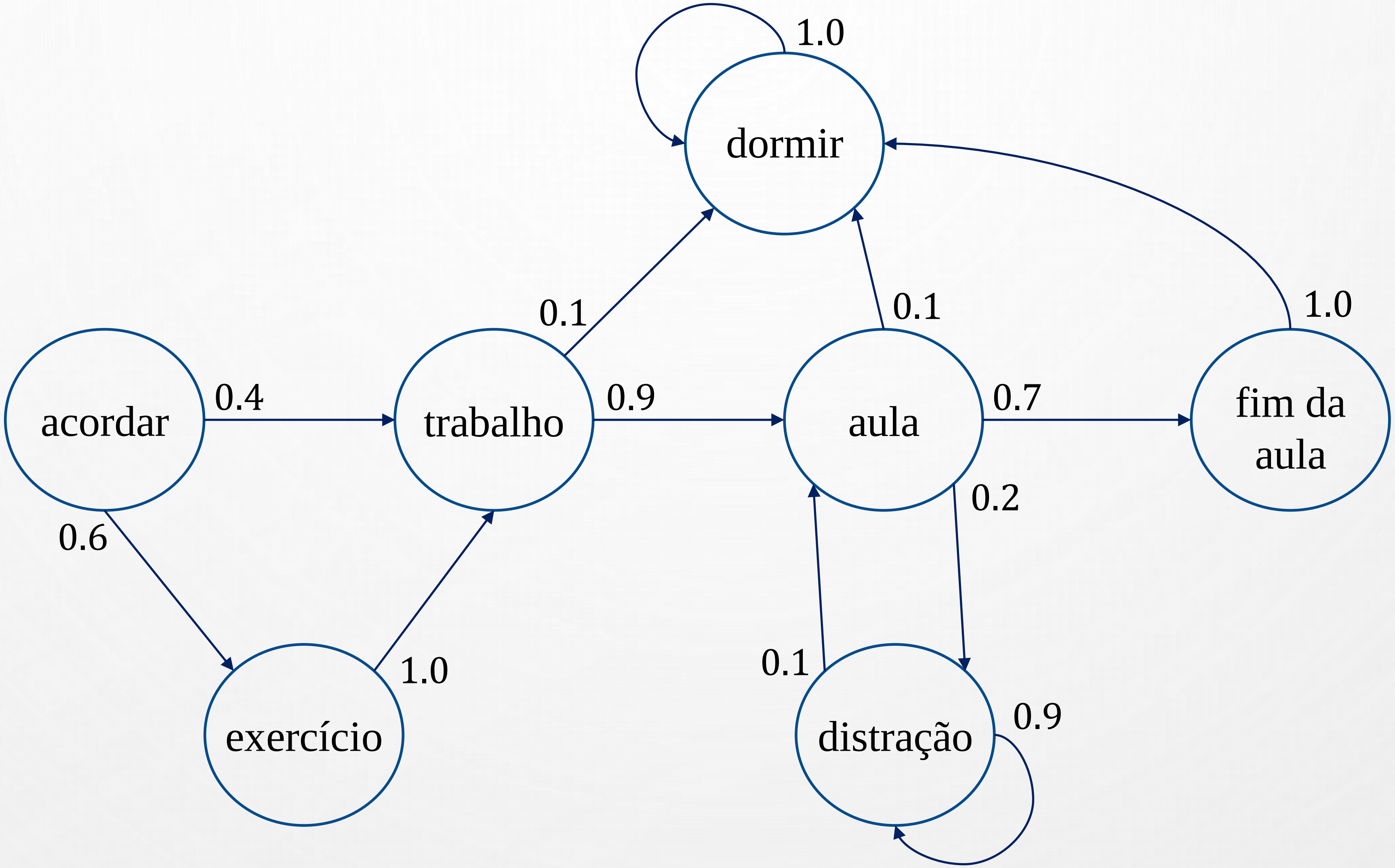
A função de transição de estados $\mathcal{P}\colon \mathcal{S} x \mathcal{S} \rightarrow [0,1]$ pode ser representada por uma **Matriz de Transição de Estados** $\mathcal{P} \in [0,1]^{n\,x\,n}$, onde n $= |\mathcal{S}|$ é o número de estados do MP e cada elemento é dado por:

$$\mathcal{P}_{ss'} = \mathbb{P}(S_{t+1} = s' | S_t = s)$$

De modo que a Matriz de Transição de Estados define as probabilidades de transição de todos estados s para todos estados sucessores s':

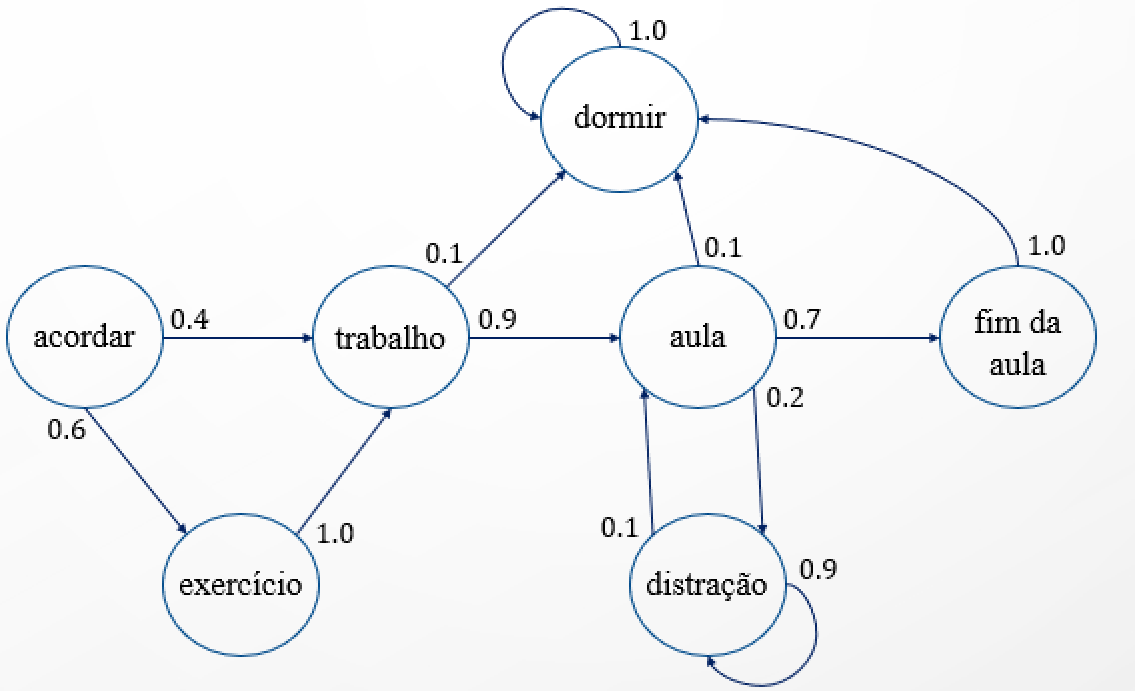
$$\mathcal{P} = \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix}$$

MP: EXEMPLO



dormir
1.0
0.1
0.1
1.0
acordar
0.4
trabalho
0.9
aula
0.7
fim da aula
0.6
0.2
exercício
1.0
0.1
distração
0.9

MP: EXEMPLO

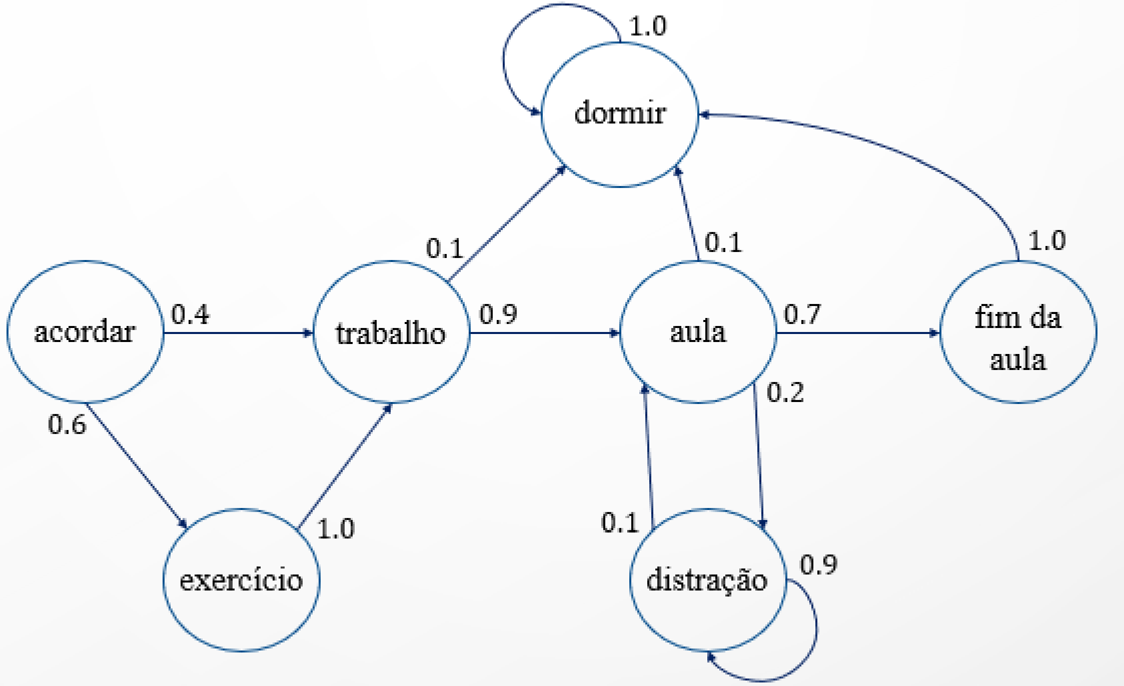


A Cadeia de Markov do exemplo é definida por um conjunto finito de estados $\mathcal{S}$ e por uma Matriz de Transição de Estados $\mathcal{P}$:

$$\mathcal{S} = \{acordar, exercício, trabalho, aula, \\ distração, fim\ da\ aula, dormir\}$$

$$\mathcal{P} = \begin{array}{r} \\ acordar \\ ex. \\ trab. \\ aula \\ dist. \\ fim \\ dormir \end{array} \begin{array}{c} \begin{array}{ccccccc} acordar & ex. & trab. & aula & dist. & fim & dormir \end{array} \\ \begin{bmatrix} 0 & 0.6 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.1 \\ 0 & 0 & 0 & 0.1 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

MP: EXEMPLO



Amostragem de episódios $(S_0, S_1, \ldots, S_T)$ a partir de Cadeia de Markov:

- Ep1: acordar, exercício, trabalho, aula, fim da aula, dormir
- Ep2: acordar, trabalho, aula, distração, aula, dormir
- Ep3: acordar, exercício, trabalho, dormir

É possível calcular a probabilidade de cada episódio:

- $\mathbb{P}(Ep_1) = \mathbb{P}(S_0) * \mathbb{P}(S_1|S_0) * \cdots * \mathbb{P}(S_T|S_{T-1}) = 1.0 * 0.6 * 1.0 * 0.9 * 0.7 = 0.378$
- $\mathbb{P}(Ep_2) = \mathbb{P}(S_0) * \mathbb{P}(S_1|S_0) * \cdots * \mathbb{P}(S_T|S_{T-1}) = 1.0 * 0.4 * 0.9 * 0.2 * 0.1 * 0.1 = 0.00072$
- $\mathbb{P}(Ep_3) = \mathbb{P}(S_0) * \mathbb{P}(S_1|S_0) * \cdots * \mathbb{P}(S_T|S_{T-1}) = 1.0 * 0.6 * 1.0 * 0.1 = 0.06$

PROCESSO DE RECOMPENSA DE MARKOV (MRP)

Processo de Recompensa de Markov (MRP)

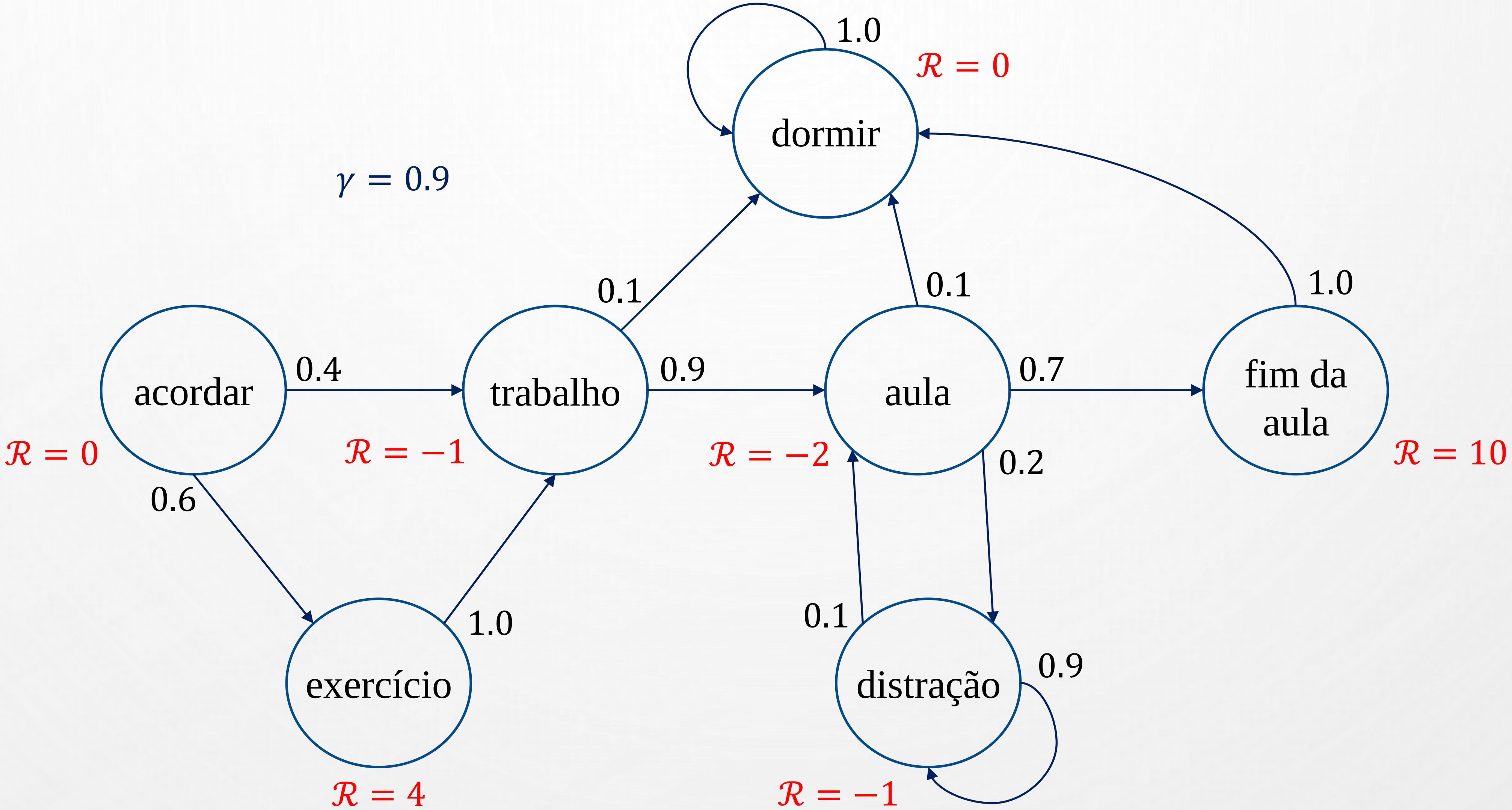
PROCESSO DE RECOMPENSA DE MARKOV (MRP)

Um **Processo de Recompensa de Markov** (ou MRP) é uma tupla $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$, onde:

- $\mathcal{S}$ é um conjunto finito de estados.
- $\mathcal{P}$ é uma função $\mathcal{P}$: $\mathcal{S}x\mathcal{S} \rightarrow [0,1] \subset \mathbb{R}$ de probabilidades de transições de estados.
- $\mathcal{R}$ é uma função de recompensa $\mathcal{R}: \mathcal{S} \rightarrow \mathbb{R}$ tal que $\mathcal{R}(s) = \mathbb{E}[R_{t+1} | S_t = s]$
- $\gamma \in [0,1] \subset \mathbb{R}$ é um fator de desconto.

Um MRP é a extensão de um MP onde estuda-se uma grandeza escalar associada aos estados (recompensa).

MRP: EXEMPLO



dormir
1.0
$\mathcal{R} = 0$
$\gamma = 0.9$
0.1
0.1
1.0
acordar
0.4
trabalho
0.9
aula
0.7
fim da aula
$\mathcal{R} = 0$
$\mathcal{R} = -1$
$\mathcal{R} = -2$
0.2
$\mathcal{R} = 10$
0.6
exercício
1.0
0.1
distração
0.9
$\mathcal{R} = 4$
$\mathcal{R} = -1$

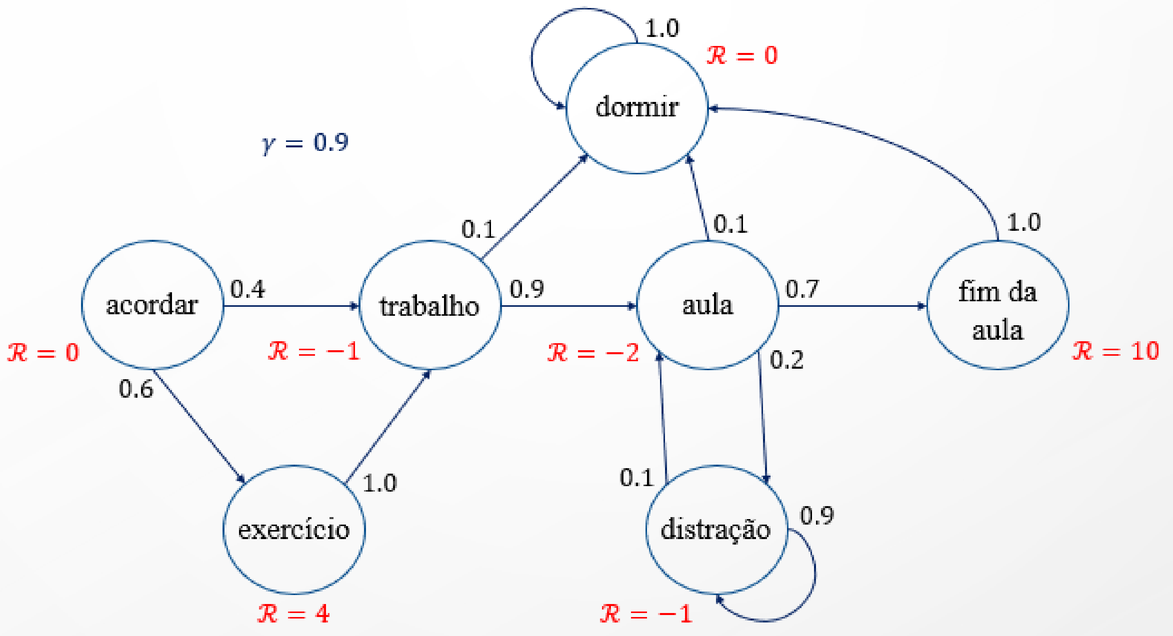
RETORNO G_t

O **Retorno** G_t a partir de determinado instante de tempo t é definido como a soma descontada de recompensas obtidas a partir deste instante:

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

- O Fator de Desconto γ representa o valor atual de recompensas futuras.
- O valor associado ao recebimento de R após $k+1$ instantes de tempo é $\gamma^k R$
- $\gamma \approx 0 \Rightarrow$ Avaliação "short-sighted" (apenas recompensas próximas são consideradas)
- $y \approx 1 \Rightarrow$ Avaliação "far-sighted" (recompensas distantes apresentam mesma importância que próximas)

MRP: EXEMPLO – RETORNOS G_0



Amostragem de episódios $(S_0, R_1, S_1, R_2 \dots, S_T, R_{T+1})$ a partir de Cadeia de Markov:

- Ep1: acordar, exercício, trabalho, aula, fim da aula, dormir
- Ep2: acordar, trabalho, aula, distração, aula, dormir
- Ep3: acordar, exercício, trabalho, dormir

É possível calcular o retorno G_0 do estado inicial para cada episódio:

- $Ep_1 \Rightarrow G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots + \gamma^T R_{T+1} = 0 + 0.9(4) + 0.9^2(-1) + 0.9^3(-2) + 0.9^4(10) + 0.9^5(0) = 7.89$
- $Ep_1 \Rightarrow G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots + \gamma^T R_{T+1} = 0 + 0.9(-1) + 0.9^2(-2) + 0.9^3(-1) + 0.9^4(-2) + 0.9^5(0) = -4.56$
- $Ep_1 \Rightarrow G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots + \gamma^T R_{T+1} = 0 + 0.9(4) + 0.9^2(-1) + 0.9^3(0) = 2.79$

FUNÇÃO VALOR DOS ESTADOS $V(s)$ PARA MRP

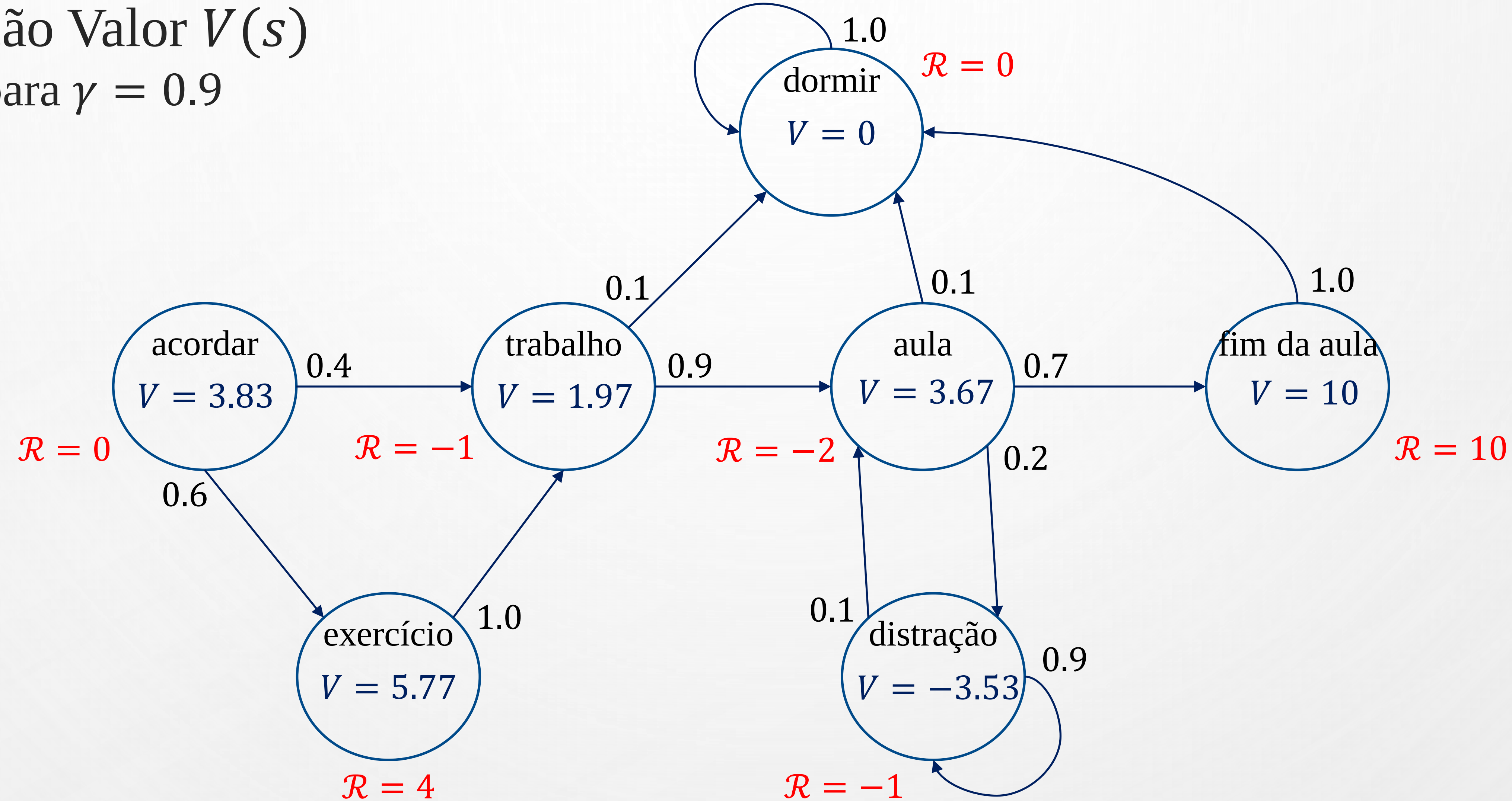
A **Função Valor dos Estados** $V(s)$ de um MRP é o valor esperado da soma descontada das recompensas obtidas a partir do estado s, sendo uma medida do valor a longo-prazo daquele estado:

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

- O retorno G_t é uma variável aleatória que depende do episódio.
- O valor de um estado $V(s)$ é o valor esperado dos retornos a partir daquele estado e é uma variável escalar fixa dado o MRP.

MRP: EXEMPLO – FUNÇÃO VALOR $V(s)$

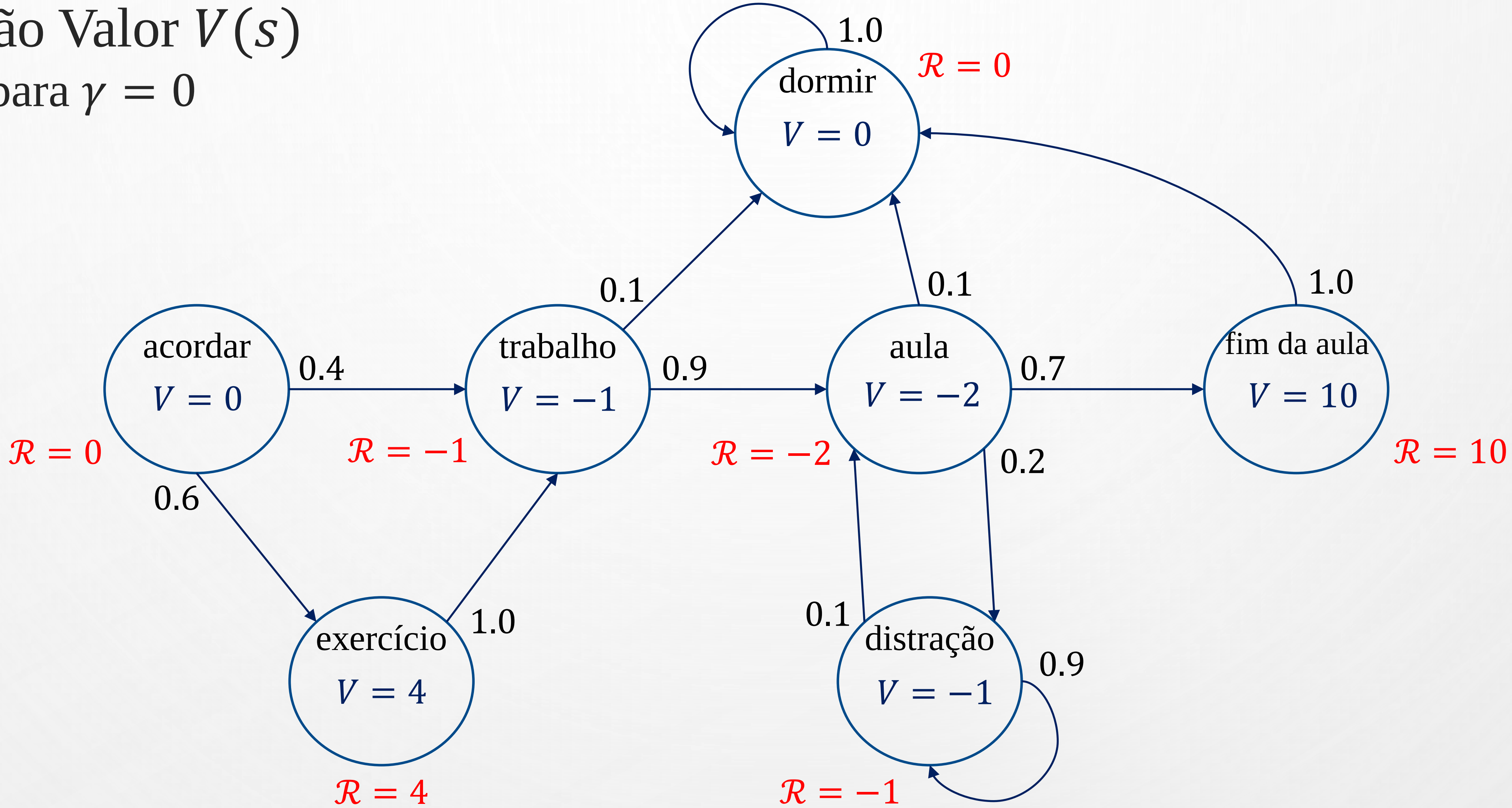
Função Valor $V(s)$
para $\gamma = 0.9$



MRP: EXEMPLO – FUNÇÃO VALOR $V(s)$

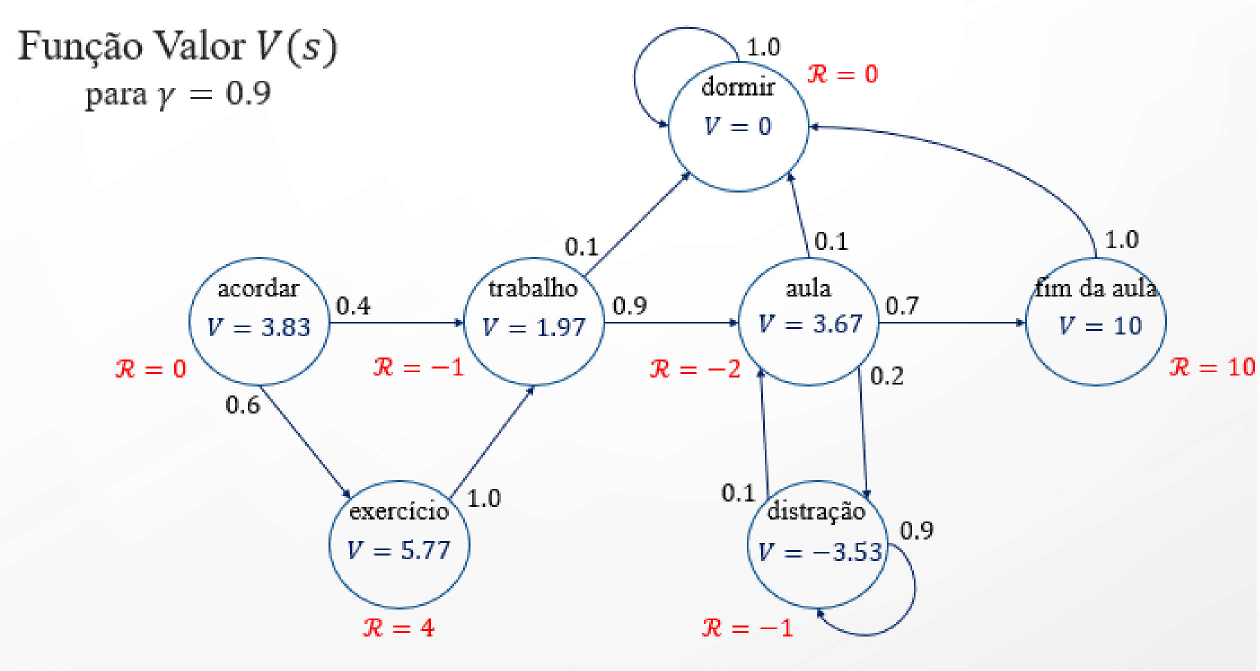
Função Valor $V(s)$

para $\gamma = 0$

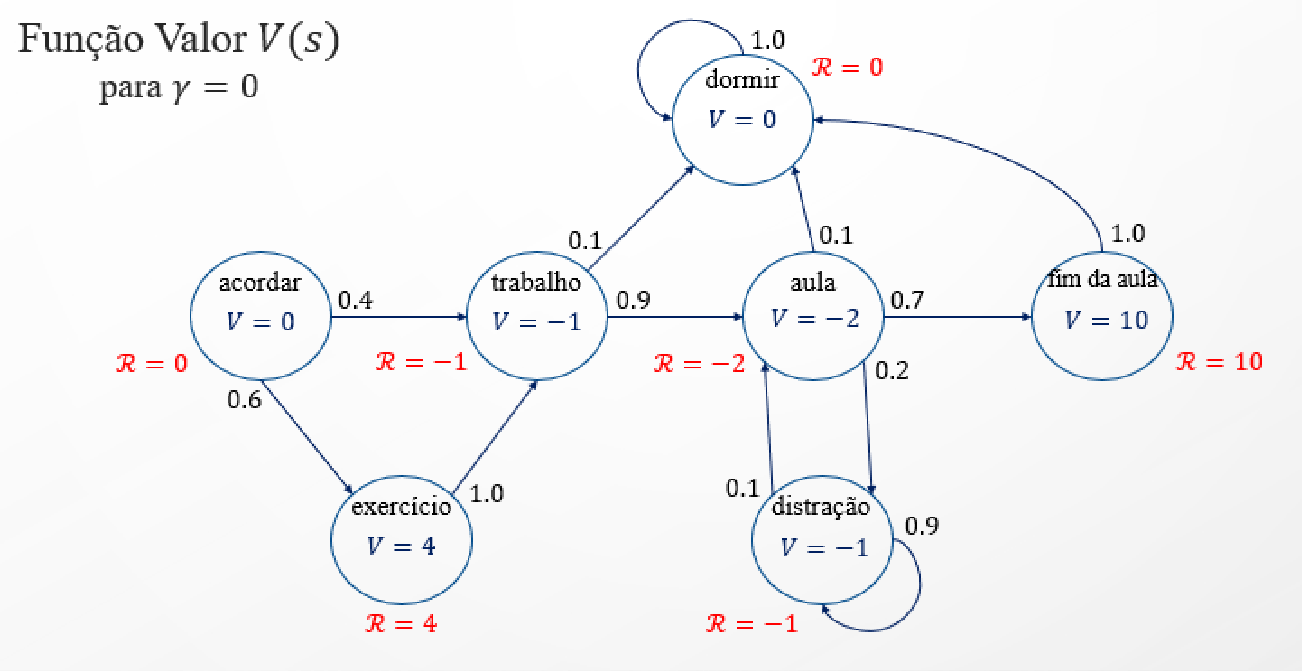


MRP: EXEMPLO – FUNÇÃO VALOR $V(s)$ PARA DIFERENTES FATORES DE DESCONTO γ

$\gamma = 0.9$



$\gamma = 0$



- Função Valor $V(s)$ depende do fator de desconto γ

EQUAÇÃO DE BELLMAN PARA MRP

Como relacionar os valores de diferentes estados?

Um retorno $G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$ pode ser decomposto em:

- Recompensa imediata R_{t+1}
- Retorno descontado de estado seguinte $\gamma G_{t+1} = \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

Substituindo na definição da função Valor V:

$$
\begin{aligned}
V(s) &\doteq \mathbb{E}[G_t | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma V(S_{t+1}) | S_t = s]
\end{aligned}
$$

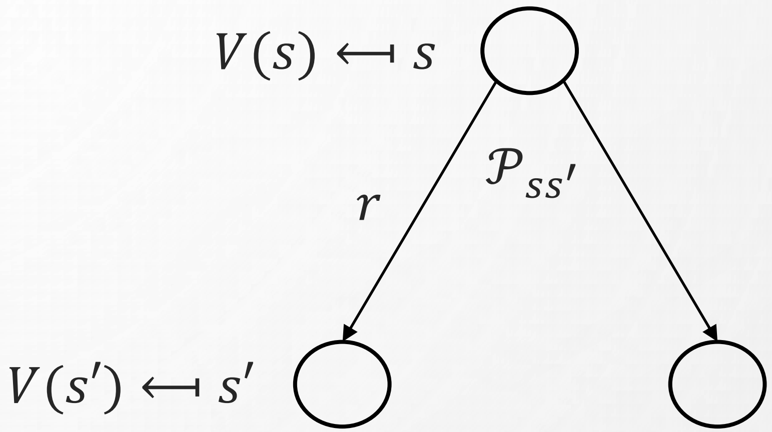
EQUAÇÃO DE BELLMAN PARA MRP

$$V(s) = \mathbb{E}[R_{t+1} + \gamma V(S_{t+1}) | S_t = s]$$

$$V(s) = \mathbb{E}[R_{t+1} | S_t = s] + \mathbb{E}[\gamma V(S_{t+1}) | S_t = s]$$

$$\boxed{V(s) = \mathcal{R}(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} V(s')}$$

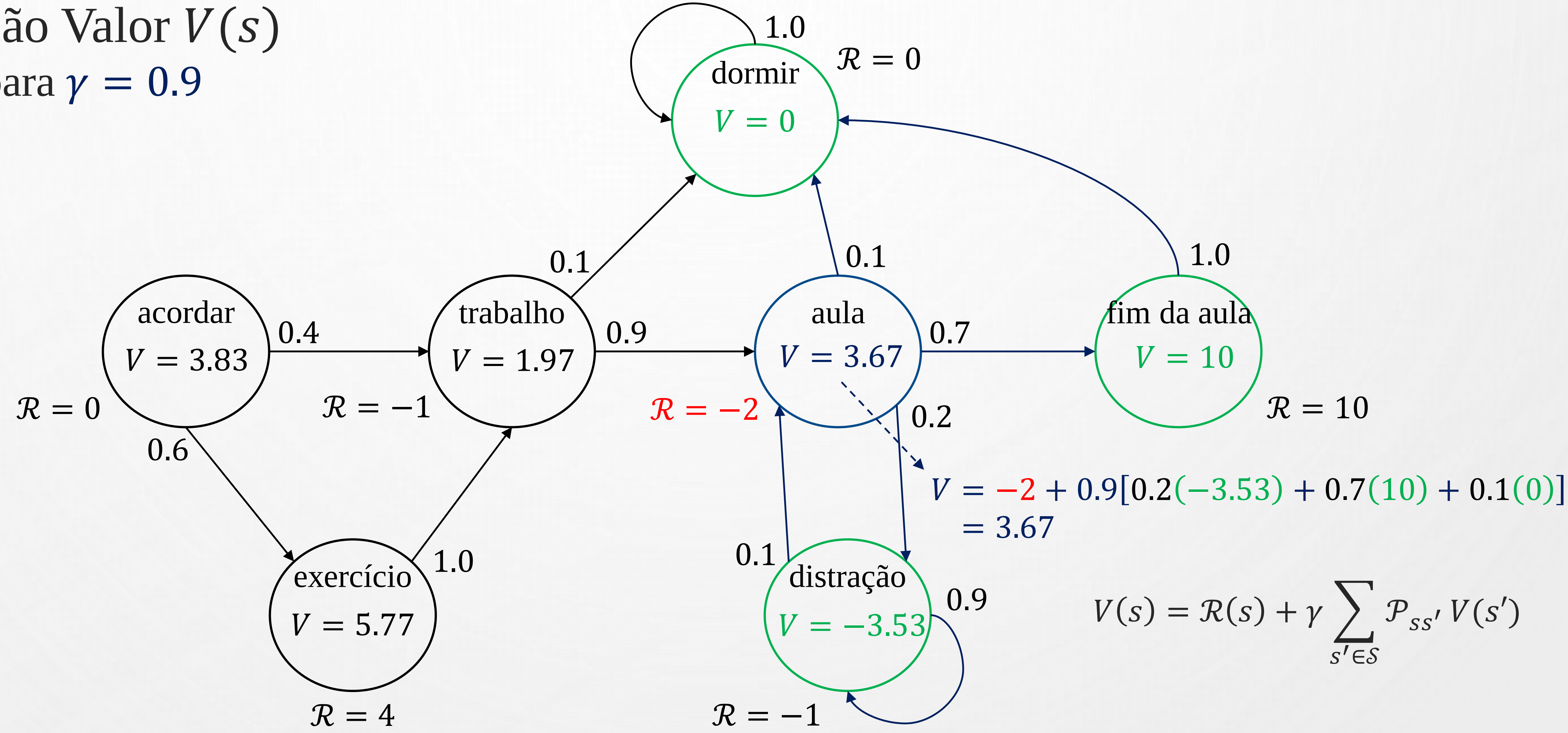
Bellman Expectation Equation for MRPs



MRP: EXEMPLO – FUNÇÃO VALOR $V(s)$

Função Valor $V(s)$

para $\gamma = 0.9$



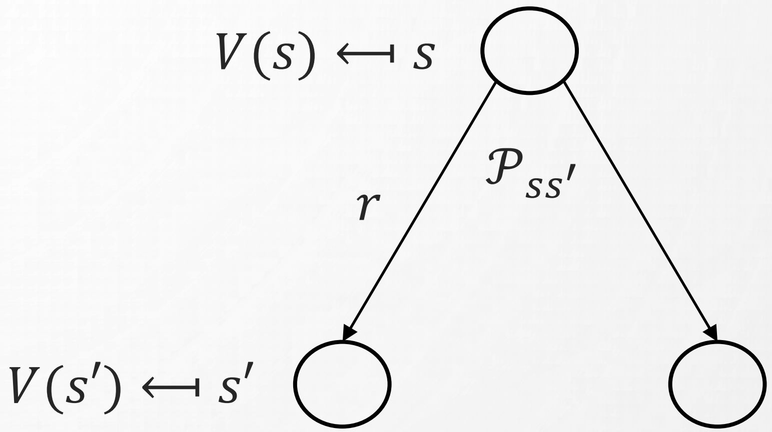
$$V(s) = \mathcal{R}(s) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'} V(s')$$

EQUAÇÃO DE BELLMAN: FORMA MATRICIAL

A Equação de Bellman pode ser escrita em forma matricial:

$$\boldsymbol{v} = \boldsymbol{\mathcal{R}} + \gamma \boldsymbol{\mathcal{P}} \boldsymbol{v}$$

onde $\boldsymbol{v} \in \mathbb{R}^n$ é um vetor coluna com os valores de cada estado, $\boldsymbol{\mathcal{R}} \in \mathbb{R}^n$ é um vetor coluna com as recompensas imediatas e $\boldsymbol{\mathcal{P}} \in \mathbb{R}^{n\,x\,n}$ é a Matriz de Transição de Estados:



$$\begin{bmatrix} V(s_1) \\ \vdots \\ V(s_n) \end{bmatrix} = \begin{bmatrix} \mathcal{R}(s_1) \\ \vdots \\ \mathcal{R}(s_n) \end{bmatrix} + \gamma \begin{bmatrix} \mathcal{P}_{11} & \cdots & \mathcal{P}_{1n} \\ \vdots & \ddots & \vdots \\ \mathcal{P}_{n1} & \cdots & \mathcal{P}_{nn} \end{bmatrix} \begin{bmatrix} V(s_1) \\ \vdots \\ V(s_n) \end{bmatrix}$$

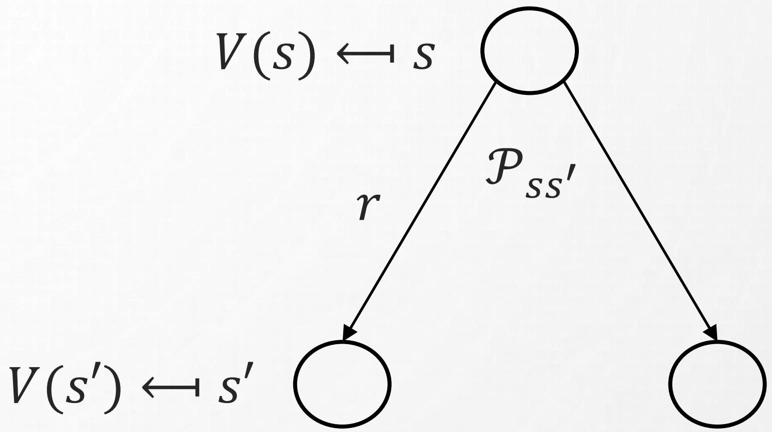
SOLUÇÃO DA EQUAÇÃO DE BELLMAN

A Equação de Bellman é linear e pode ser solucionada diretamente se o MRP $< \mathcal{S}, \mathcal{P}, \mathcal{R}, \gamma >$ é conhecido:

$$\boldsymbol{v} = \boldsymbol{\mathcal{R}} + \gamma \boldsymbol{\mathcal{P}} \boldsymbol{v}$$

$$(\boldsymbol{I} - \gamma \boldsymbol{\mathcal{P}})\boldsymbol{v} = \boldsymbol{\mathcal{R}}$$

$$\boxed{\boldsymbol{v} = (\boldsymbol{I} - \gamma \boldsymbol{\mathcal{P}})^{-1} \boldsymbol{\mathcal{R}}}$$



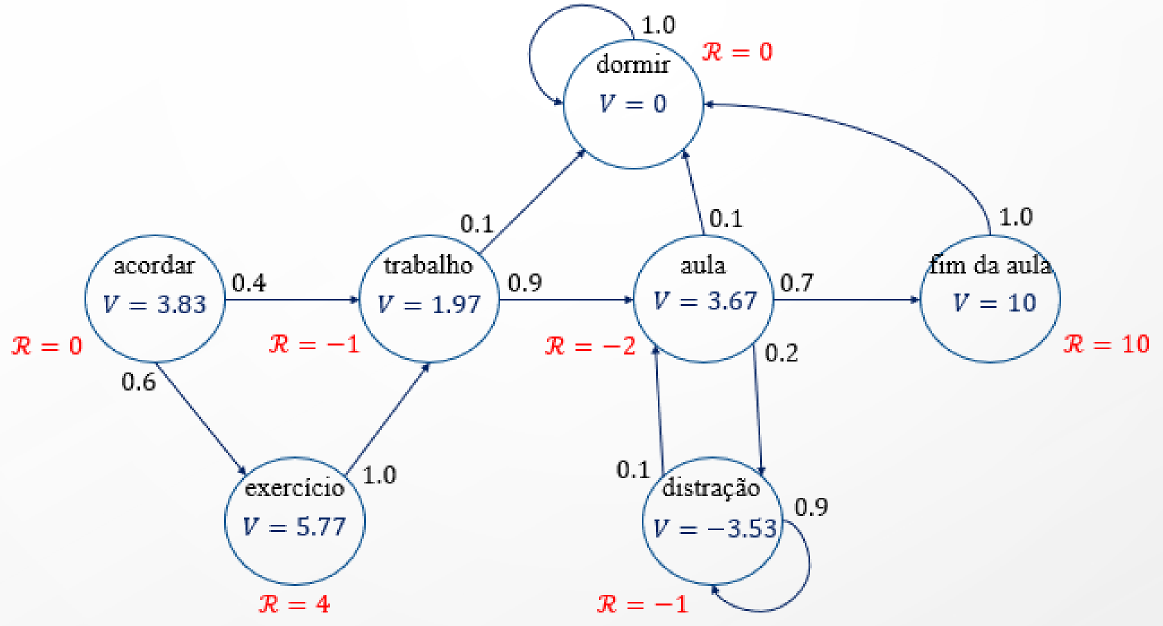
- Complexidade computacional $O(n^3)$ para n estados.
- Solução direta somente é possível em casos simples e para MRPs pequenos.
- Métodos Iterativos para MRPs maiores:

  - Programação Dinâmica

  - Monte-Carlo

  - Temporal Difference Learning (TD(0), TD(λ))

MRP: EXEMPLO – FUNÇÃO VALOR $V(s)$



$$\boldsymbol{v} = (\boldsymbol{I} - \gamma \boldsymbol{\mathcal{P}})^{-1} \boldsymbol{\mathcal{R}}$$

$$\boldsymbol{v} = \left[\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} - 0.9 \begin{pmatrix} 0 & 0.6 & 0.4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.9 & 0 & 0 & 0.1 \\ 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.1 \\ 0 & 0 & 0 & 0.1 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right]^{-1} \begin{bmatrix} 0 \\ 4 \\ -1 \\ -2 \\ -1 \\ 10 \\ 0 \end{bmatrix} = \begin{bmatrix} 3.83 \\ 5.77 \\ 1.97 \\ 3.67 \\ -3.53 \\ 10 \\ 0 \end{bmatrix}$$

PROCESSO DE DECISÃO DE MARKOV (MDP)

Processo de Decisão de Markov (MDP)

PROCESSO DE DECISÃO DE MARKOV (MDP)

Um **Processo de Decisão de Markov** (ou MDP) é uma tupla $< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$, onde:

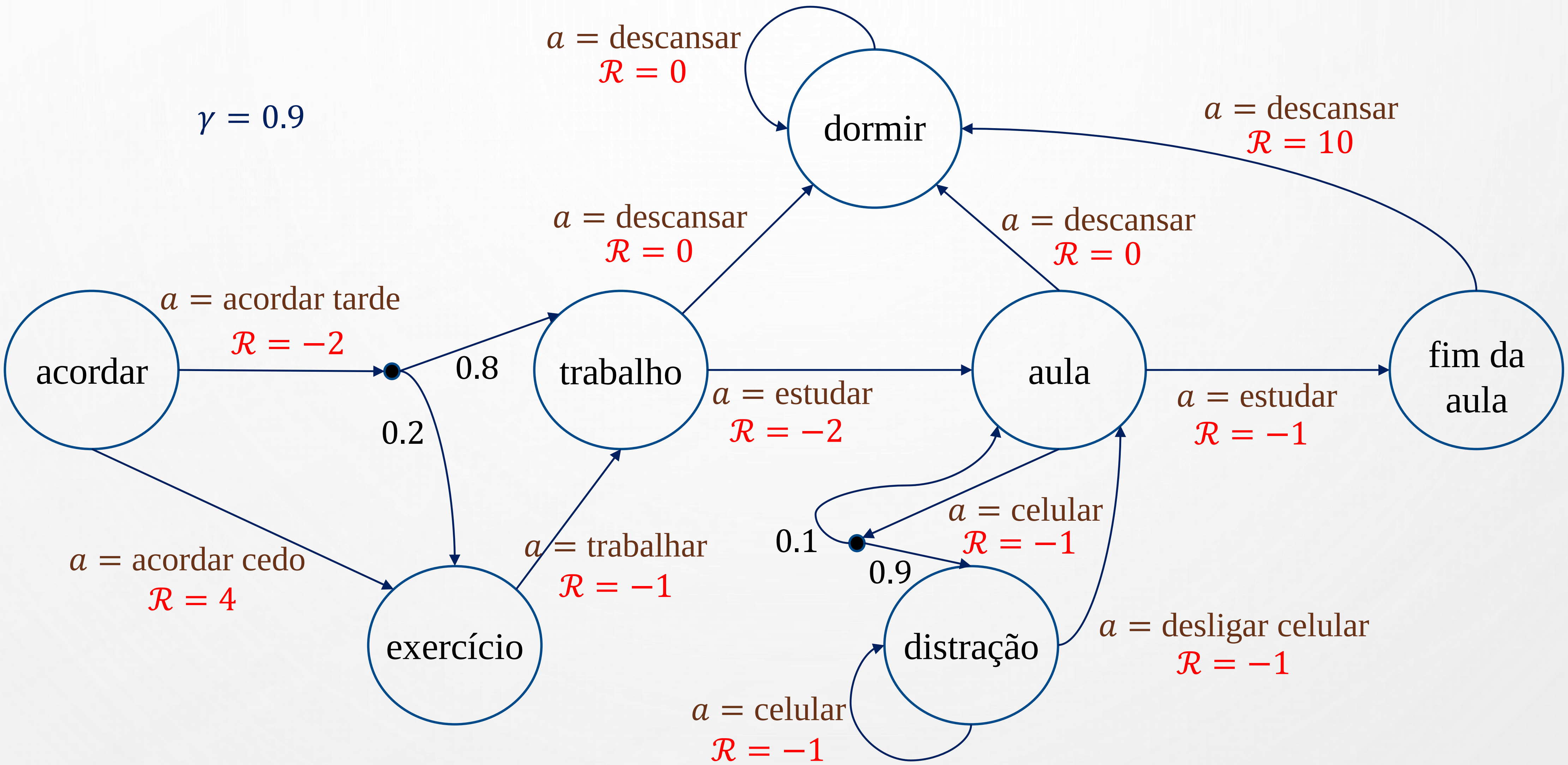
- $\mathcal{S}$ é um conjunto finito de estados.
- $\mathcal{A}$ é um conjunto finito de ações.
- $\mathcal{P}$ é uma função $\mathcal{P}$: $\mathcal{S}x\mathcal{A}x\mathcal{S} \rightarrow [0,1] \subset \mathbb{R}$ de probabilidades de transições de estados.

$$\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$$

- $\mathcal{R}$ é uma função de recompensa $\mathcal{R}: \mathcal{S}x\mathcal{A} \rightarrow \mathbb{R}$ tal que $\mathcal{R}(s,a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
- $\gamma \in [0,1] \subset \mathbb{R}$ é um fator de desconto.

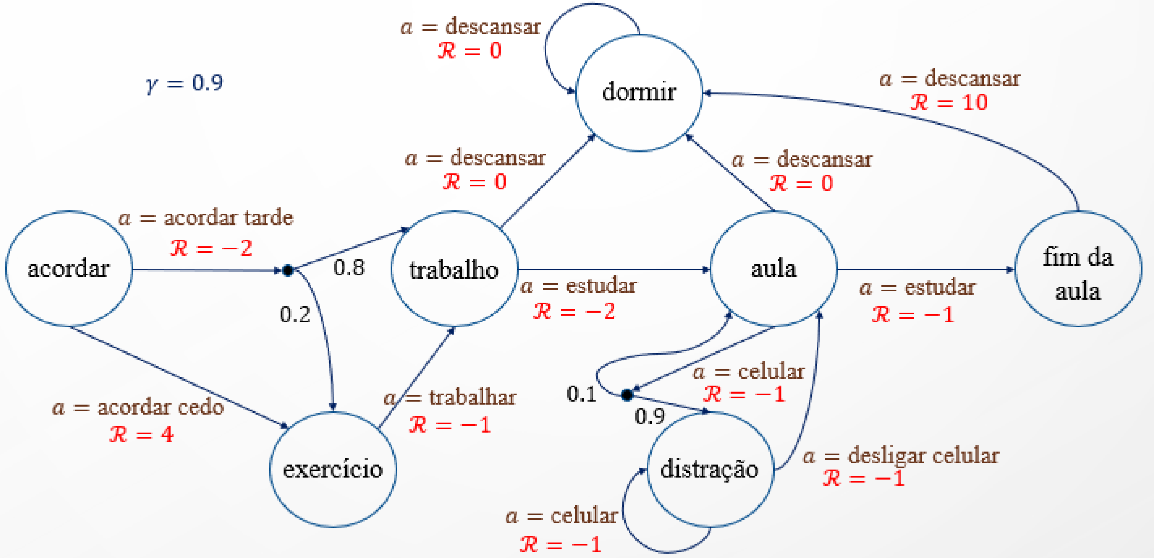
Um MDP é a extensão de um MRP onde estuda-se o processo de tomada de decisões de um agente que interage com o ambiente de modo a maximizar as recompensas obtidas.

MDP: EXEMPLO



$\gamma = 0.9$
a = descansar
$\mathcal{R} = 0$
dormir
a = descansar
$\mathcal{R} = 10$
a = descansar
$\mathcal{R} = 0$
a = descansar
$\mathcal{R} = 0$
a = acordar tarde
$\mathcal{R} = -2$
acordar
0.8
trabalho
aula
fim da aula
0.2
a = estudar
$\mathcal{R} = -2$
a = estudar
$\mathcal{R} = -1$
a = celular
$\mathcal{R} = -1$
0.1
0.9
a = acordar cedo
$\mathcal{R} = 4$
a = trabalhar
$\mathcal{R} = -1$
exercício
distração
a = desligar celular
$\mathcal{R} = -1$
a = celular
$\mathcal{R} = -1$

MDP: EXEMPLO



O MDP do exemplo é definido por:

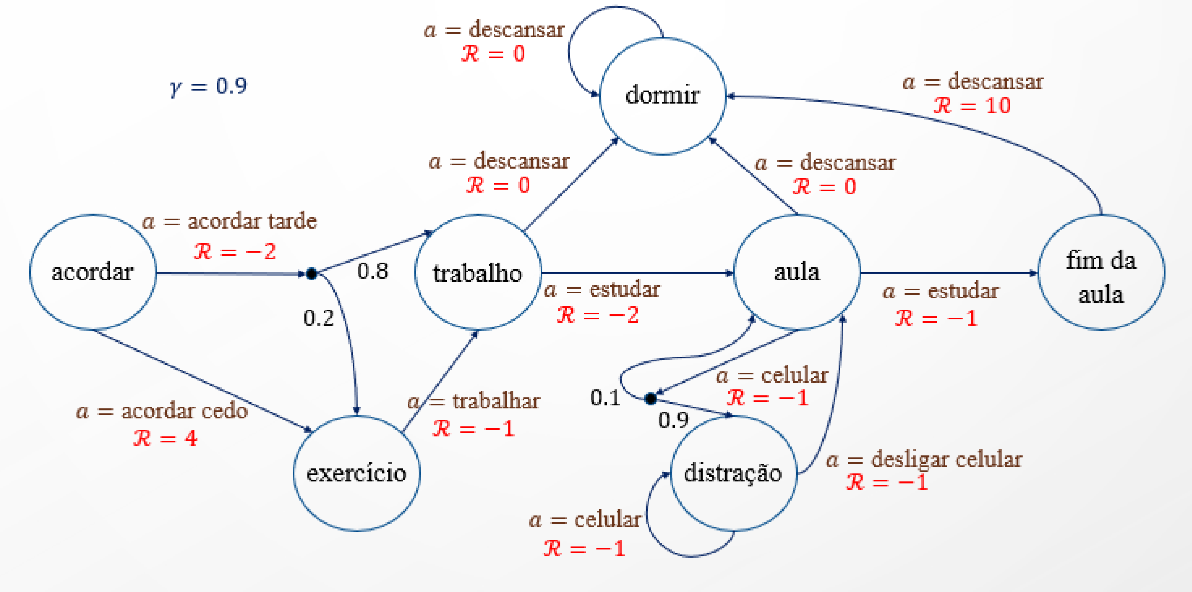
$$\mathcal{S} = \{acordar, exercício, trabalho, aula, distração, fim\ da\ aula, dormir\}$$

$$\mathcal{A} = \{acordar\ cedo, acordar\ tarde, trabalhar, estudar, celular, desligar\ celular, descansar\}$$

$$\mathcal{P}^{acordar\ tarde} = \begin{array}{r} \\ acordar \\ ex. \\ trab. \\ aula \\ dist. \\ fim \\ dormir \end{array} \begin{array}{c} \begin{array}{ccccccc} acordar & ex. & trab. & aula & dist. & fim & dormir \end{array} \\ \begin{bmatrix} 0 & 0.2 & 0.8 & 0 & 0 & 0 & 0 \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \\ & & & & & & \end{bmatrix} \end{array}$$

$$\mathcal{P}^{celular} = \begin{array}{r} \\ acordar \\ ex. \\ trab. \\ aula \\ dist. \\ fim \\ dormir \end{array} \begin{array}{c} \begin{array}{ccccccc} acordar & ex. & trab. & aula & dist. & fim & dormir \end{array} \\ \begin{bmatrix} & & & & & & \\ & & & & & & \\ & & & & & & \\ 0 & 0 & 0 & 0.1 & 0.9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ & & & & & & \\ & & & & & & \end{bmatrix} \end{array}, \ldots$$

MDP: EXEMPLO



O MDP do exemplo é definido por:

$$\mathcal{R} = \begin{array}{r} \\ acordar \\ ex. \\ trab. \\ aula \\ dist. \\ fim \\ dormir \end{array} \begin{array}{c} \begin{array}{ccccccc} ac.cedo & ac.tarde & trab. & est. & cel. & desl.cel. & desc. \end{array} \\ \begin{bmatrix} 4 & -2 & & & & & \\ & & -1 & & & & \\ & & & -2 & & & 0 \\ & & & -1 & -1 & & 0 \\ & & & & -1 & -1 & \\ & & & & & & 10 \\ & & & & & & 0 \end{bmatrix} \end{array}$$

- A Função de Transição de Estados $\mathcal{P}$ pode ser descrita por m matrizes de transição $\{\mathcal{P}_{ss'}^{a_i}\}_{i=1}^{m}$, uma para cada ação.
- A função de Recompensa $\mathcal{R}$ pode ser descrita por uma matriz $\mathcal{R} \in \mathbb{R}^{n\,x\,m}$, na qual cada linha corresponde a um estado e cada coluna corresponde a uma ação.

POLÍTICA DE AÇÕES π

Um **Política de Ações** $\pi: \mathcal{S} x \mathcal{A} \rightarrow [0,1] \subset \mathbb{R}$ é uma função que mapeia estados s em distribuições de probabilidades sobre o espaço de ações $\mathcal{A}$:

$$\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$$

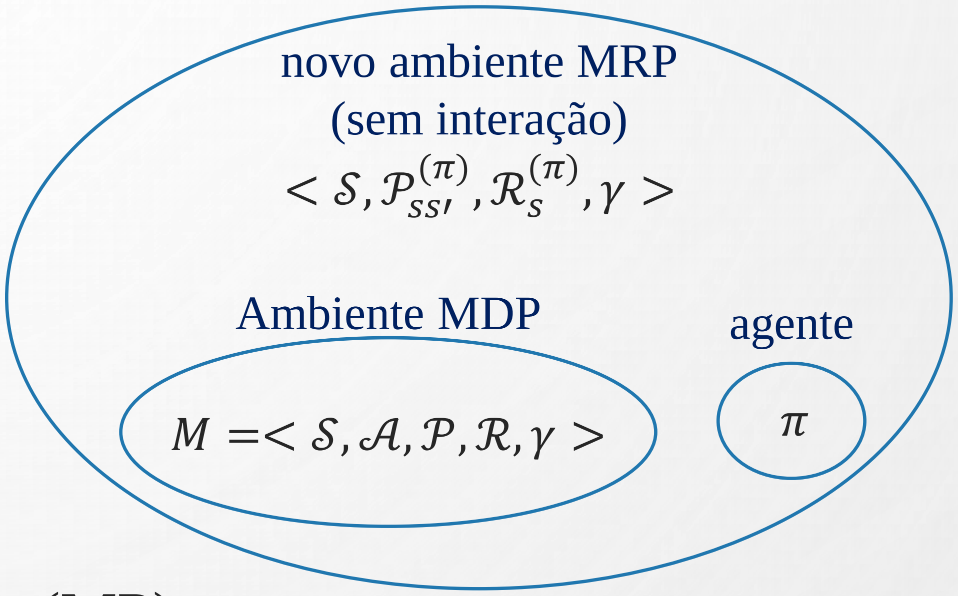
- A política de ações define completamente o comportamento do agente.
- A política é função apenas do estado, não da história (Propriedade de Markov).
- Políticas são estacionárias (independentes do tempo)

CONVERSÃO DE MDP PARA MP E MRP DADA UMA POLÍTICA

Dado um MDP $M = < \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma >$ e uma Política de Ações π, podemos definir:

$$\mathcal{P}_{ss'}^{(\pi)} = \sum_{a \in \mathcal{A}} \pi(a|s)\, \mathcal{P}_{ss'}^{a}$$

$$\mathcal{R}_{s}^{(\pi)} = \sum_{a \in \mathcal{A}} \pi(a|s)\, \mathcal{R}(s,a)$$



- O conjunto $< \mathcal{S}, \mathcal{P}_{ss'}^{(\pi)} >$ é um Processo de Markov (MP).
- O conjunto $< \mathcal{S}, \mathcal{P}_{ss'}^{(\pi)}, \mathcal{R}_{s}^{(\pi)}, \gamma >$ é um Processo de Recompensa de Markov (MRP).

FUNÇÕES VALOR DE ESTADOS E DE PARES ESTADO-AÇÃO PARA UM MDP

A **Função Valor dos Estados** $V_\pi(s)$ é o valor esperado do retorno a partir de um estado dado que o agente segue a política π:

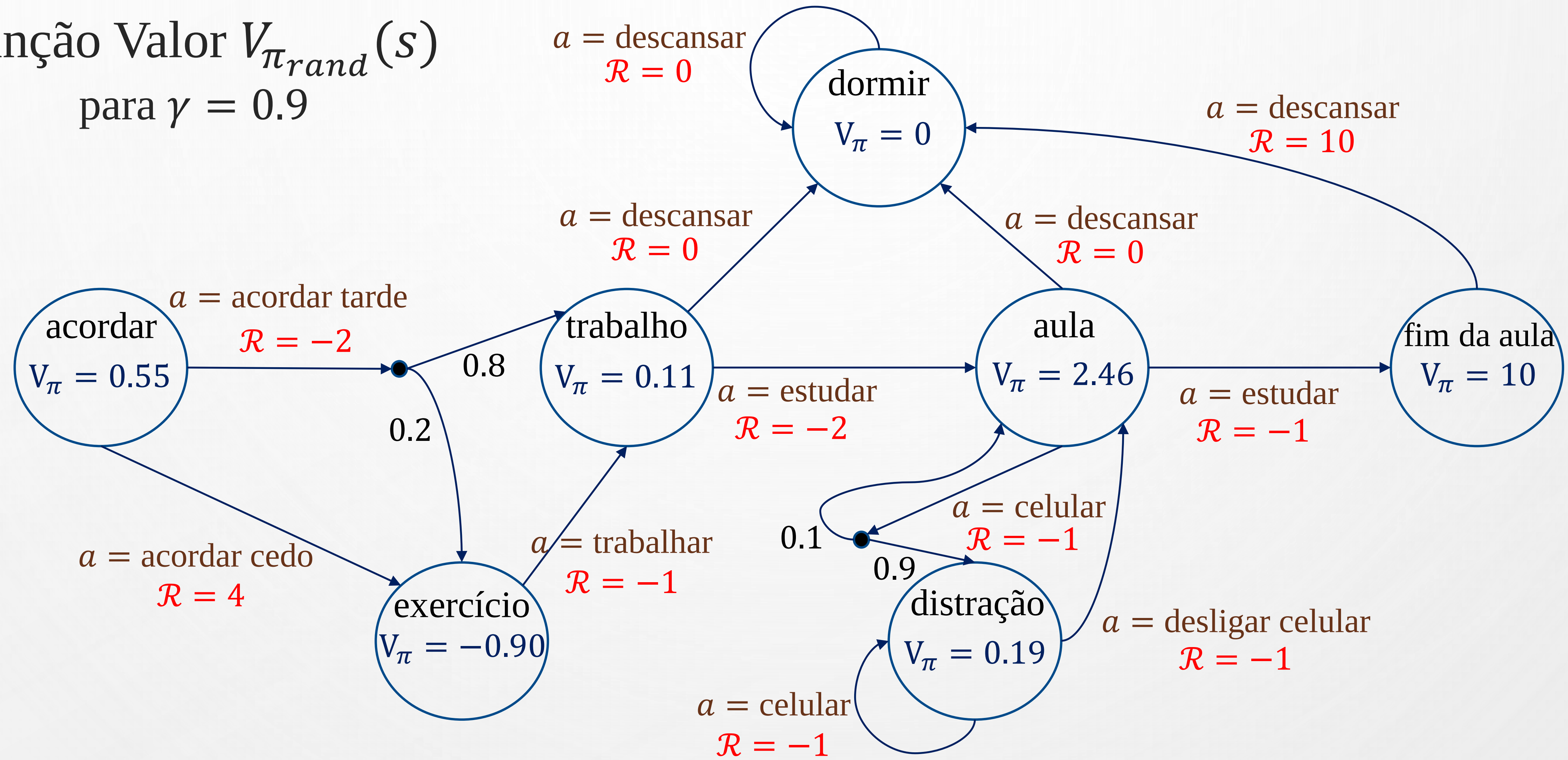
$$V_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s]$$

A **Função Valor das Ações** $Q_\pi(s, a)$ é o valor esperado do retorno a partir de um estado após a ação a ser tomada e dado que o agente segue a política π:

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

MDP: EXEMPLO – FUNÇÃO VALOR $V_\pi(s)$ PARA POLÍTICA ALEATÓRIA

Função Valor $V_{\pi_{rand}}(s)$
para $\gamma = 0.9$



EQUAÇÃO DE BELLMAN PARA MDP

A função Valor dos Estados $V(s)$ pode ser decomposta em recompensa imediata e valor descontado do estado seguinte:

$$V_\pi(s) = \mathbb{E}_\pi[R_{t+1} + \gamma V_\pi(S_{t+1}) | S_t = s] \quad (1)$$

Analogamente, a Função Valor das Ações $Q_\pi(s, a)$ pode ser escrita como:

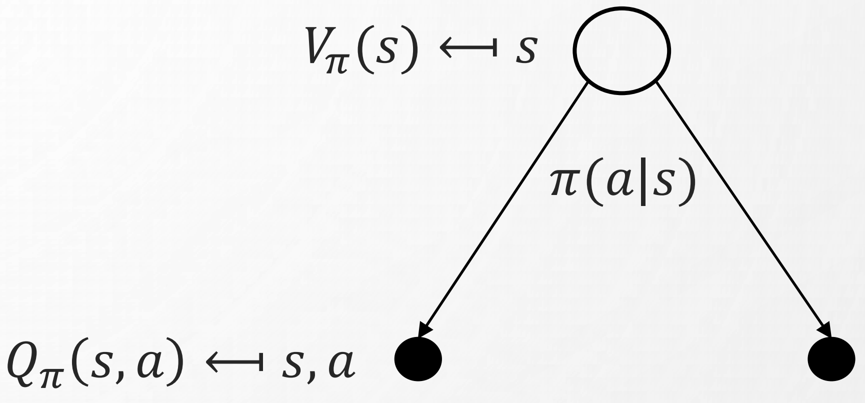
$$Q_\pi(s, a) = \mathbb{E}_\pi[R_{t+1} + \gamma Q_\pi(S_{t+1}, A_{t+1}) | S_t = s, A_t = a] \quad (2)$$

EQUAÇÃO DE BELLMAN PARA MDP (V_π)

Podemos combinar as equações (1) e (2) para escrever V_π em função de Q_π:

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\, Q_\pi(s,a) \qquad (3)$$

- O valor de um estado é dado pela média ponderada dos valores de ações naquele estado, onde a ponderação é dada pela política (probabilidade da ação ser tomada).

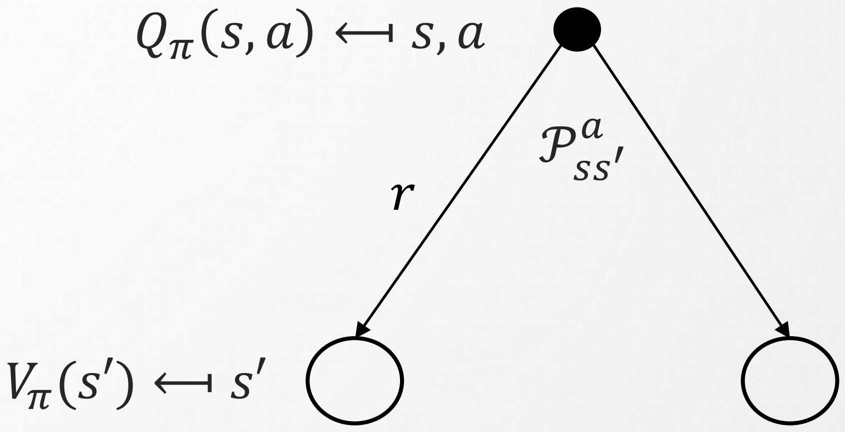


EQUAÇÃO DE BELLMAN PARA MDP (Q_π)

Analogamente, podemos combinar as equações (1) e (2) para escrever Q_π em função de V_π:

$$Q_\pi(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} V_\pi(s') \qquad (4)$$

- O valor de uma ação é dado pela recompensa imediata somada à média ponderada dos valores de estados sucessores, onde a ponderação é dada pela probabilidade de transição de estados.

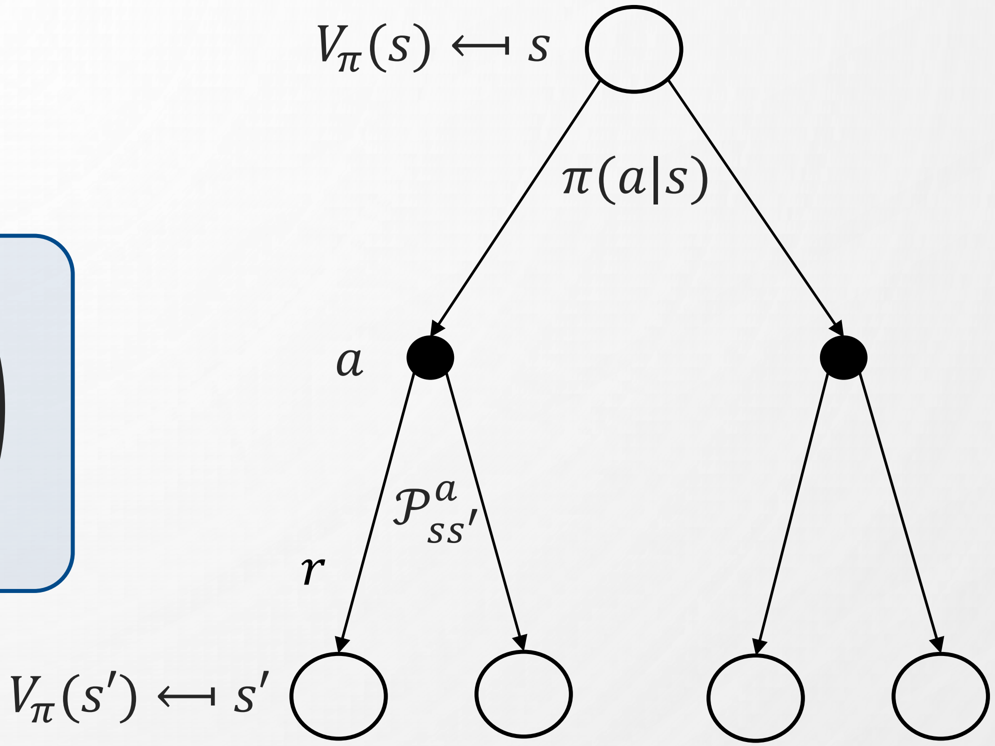


EQUAÇÃO DE BELLMAN PARA MDP (V_π)

Substituindo (4) em (3), podemos escrever $V_\pi(s)$ em função de $V_\pi(s')$:

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} V_\pi(s') \right)$$

Bellman Expectation Equation for V_π

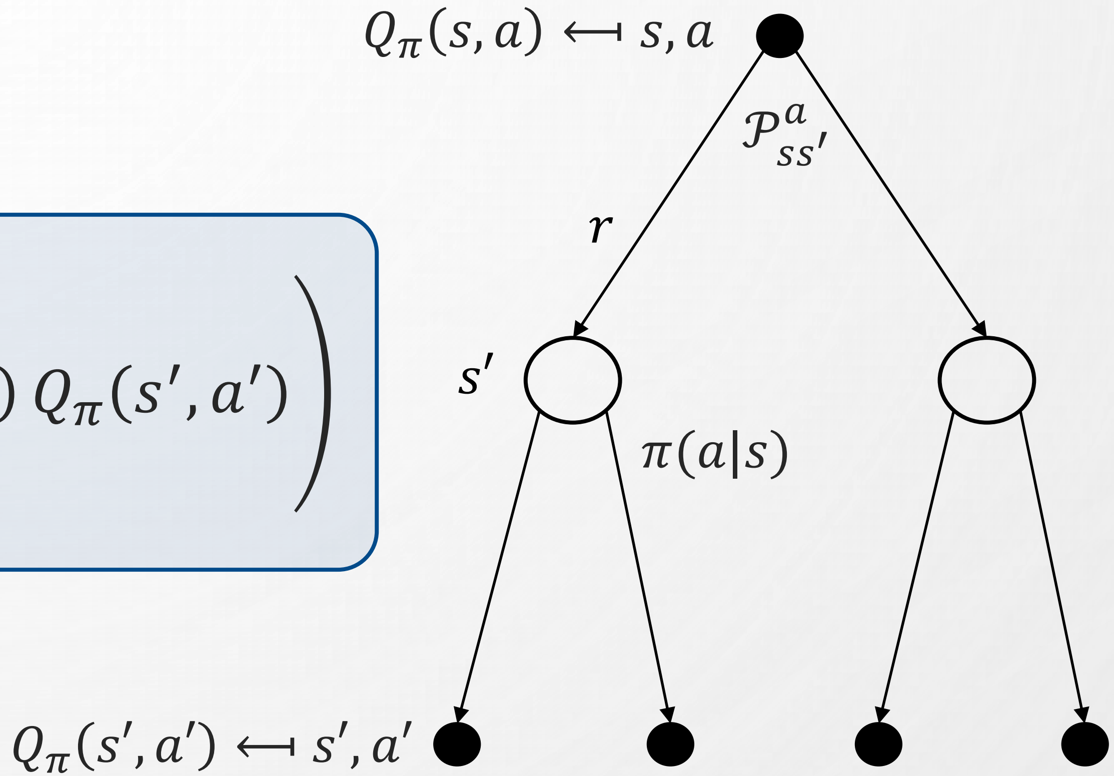


EQUAÇÃO DE BELLMAN PARA MDP (Q_π)

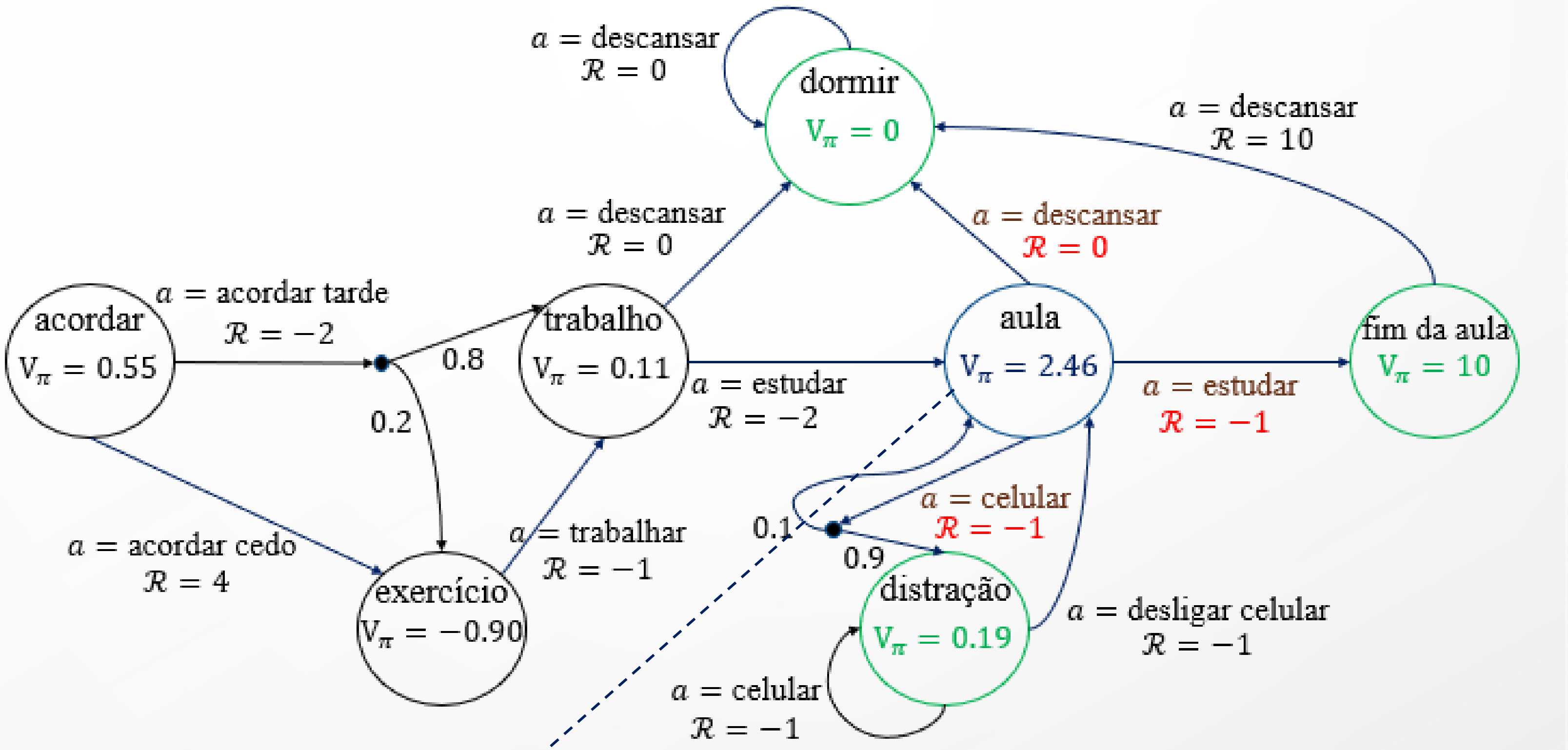
Analogamente, substituindo (3) em (4), podemos escrever $Q_\pi(s, a)$ em função de $Q_\pi(s', a')$:

$$Q_\pi(s, a) = \mathcal{R}(s, a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} \left(\sum_{a' \in \mathcal{A}} \pi(a'|s')\, Q_\pi(s', a') \right)$$

Bellman Expectation Equation for Q_π



MDP: EXEMPLO – FUNÇÃO VALOR $V_\pi(s)$ PARA POLÍTICA ALEATÓRIA



$$V = \frac{1}{3}[-1 + 0.9(0.1(2.46) + 0.9(0.19))] + \frac{1}{3}[-1 + 0.9(10)] + \frac{1}{3}[0 + 0.9(0)] = 2.46$$

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)\left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V_\pi(s')\right)$$

EQUAÇÃO DE BELLMAN PARA MDP EM FORMA MATRICIAL (V_π)

$$V_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} V_\pi(s') \right)$$

Utilizando o MRP induzido pela política π podemos escrever a Equação de Bellman em forma matricial:

$$\boldsymbol{v}_{\boldsymbol{\pi}} = \boldsymbol{\mathcal{R}}^{(\boldsymbol{\pi})} + \gamma \boldsymbol{\mathcal{P}}^{(\boldsymbol{\pi})} \boldsymbol{v}, \quad \text{onde} \quad \begin{cases} \mathcal{P}^{(\pi)}_{ss'} = \sum_{a \in \mathcal{A}} \pi(a|s)\, \mathcal{P}^a_{ss'} \\ \mathcal{R}^{(\pi)}_s = \sum_{a \in \mathcal{A}} \pi(a|s)\, \mathcal{R}(s,a) \end{cases}$$

com solução:

$$\boxed{\boldsymbol{v}_{\boldsymbol{\pi}} = \left(\boldsymbol{I} - \gamma \boldsymbol{\mathcal{P}}^{(\boldsymbol{\pi})}\right)^{-1} \boldsymbol{\mathcal{R}}^{(\boldsymbol{\pi})}}$$

FUNÇÃO VALOR ÓTIMA

A Função Valor dos Estados Ótima é definida como a maior Função Valor dos Estados sobre todas as políticas possíveis:

$$V^*(s) = \max_{\pi} V_{\pi}(s)$$

A Função Valor das Ações Ótima é definida como a maior Função Valor das Ações sobre todas as políticas possíveis:

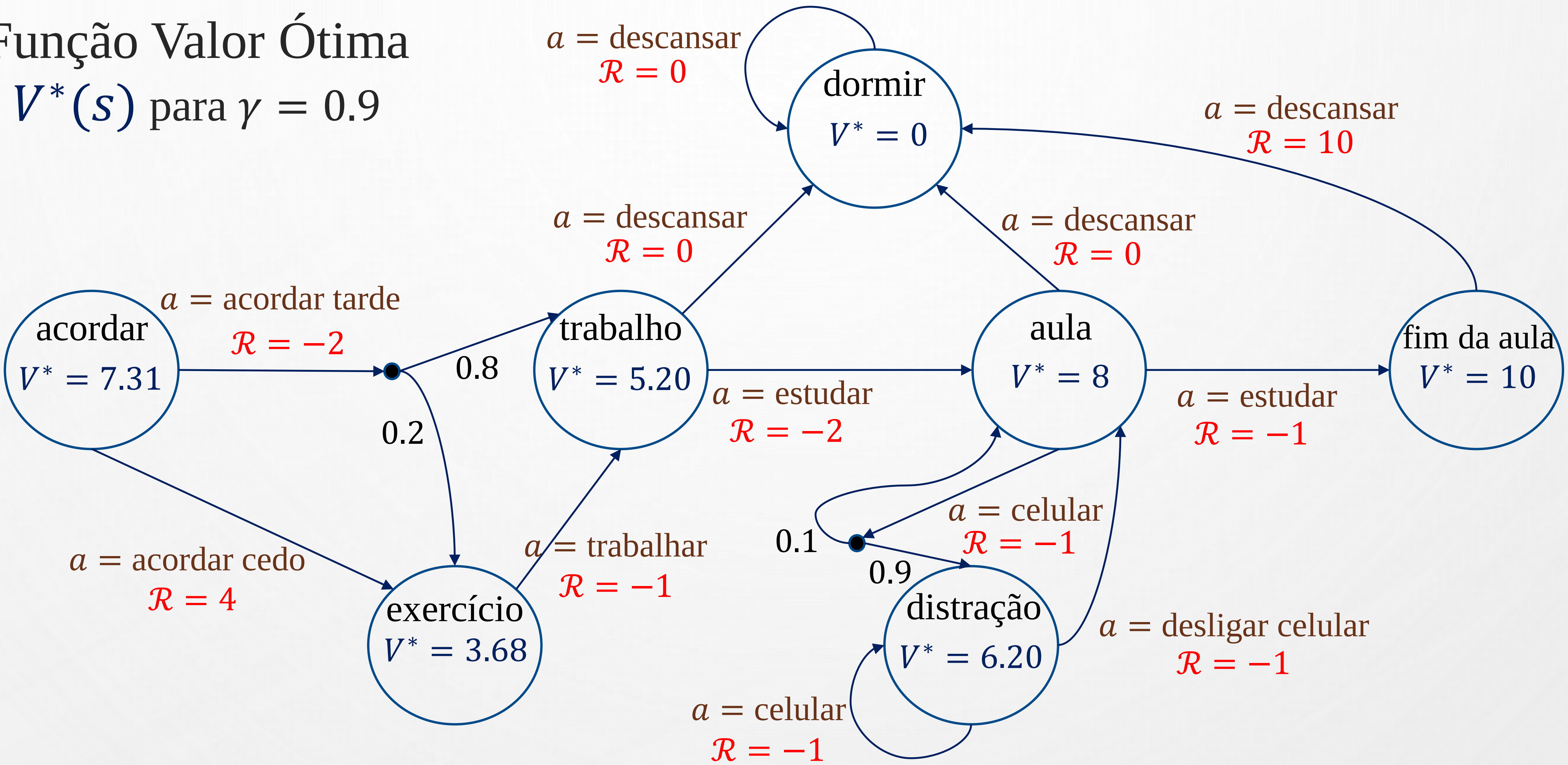
$$Q^*(s,a) = \max_{\pi} Q_{\pi}(s,a)$$

- A Função Valor Ótima especifica a melhor performance em um MDP.
- O MDP é solucionado quando a Função Valor Ótima é encontrada.

MDP: EXEMPLO – FUNÇÃO VALOR ÓTIMA $V^*(s)$

Função Valor Ótima

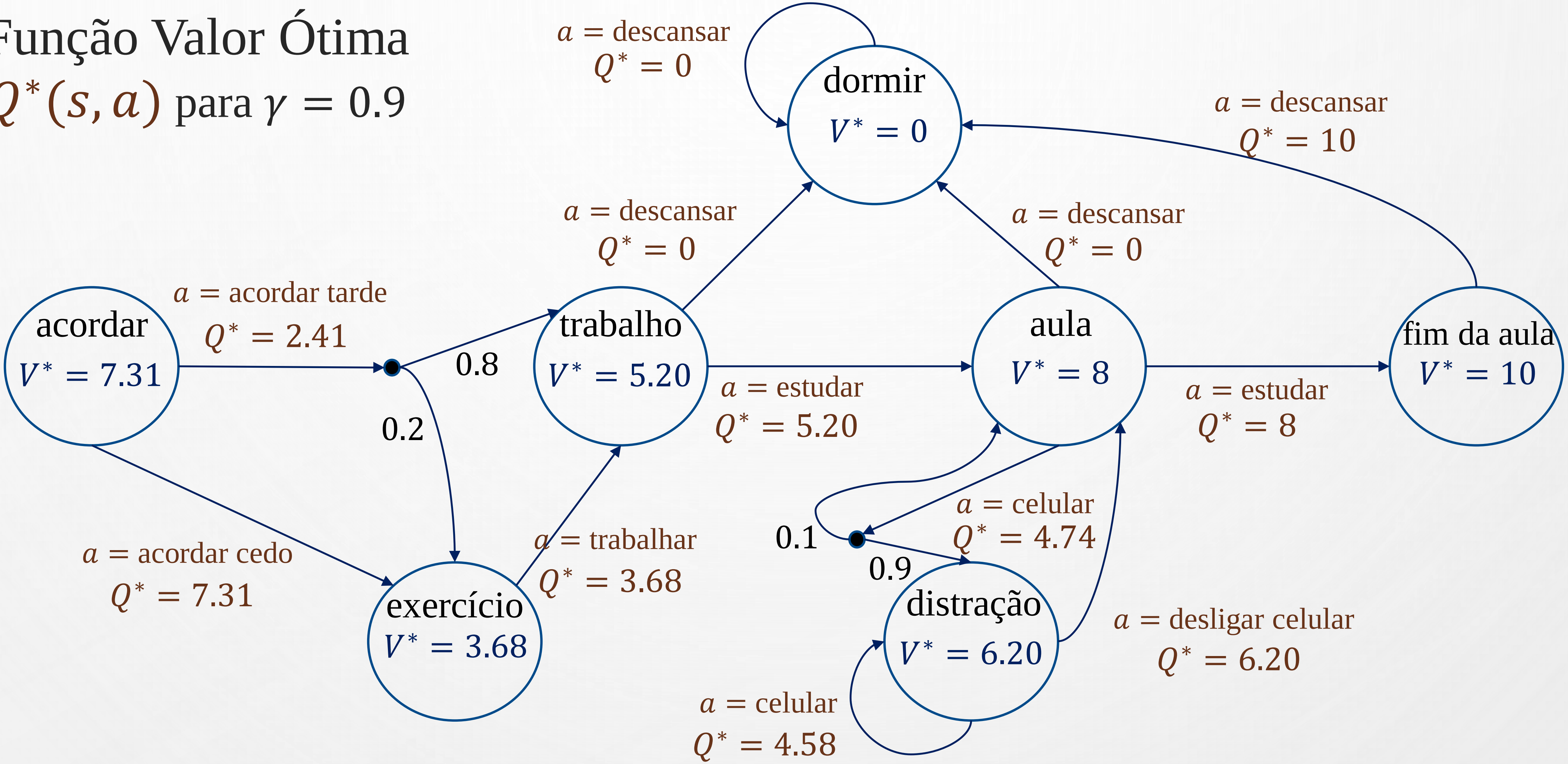
$V^*(s)$ para $\gamma = 0.9$



MDP: EXEMPLO – FUNÇÃO VALOR ÓTIMA $Q^*(s,a)$

Função Valor Ótima

$Q^*(s,a)$ para $\gamma = 0.9$

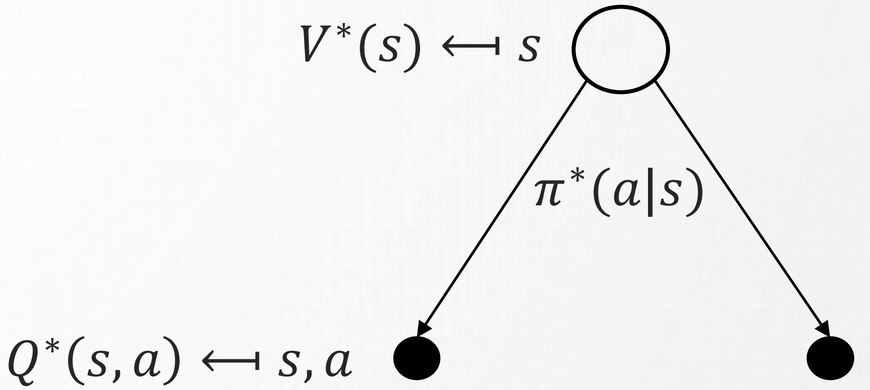


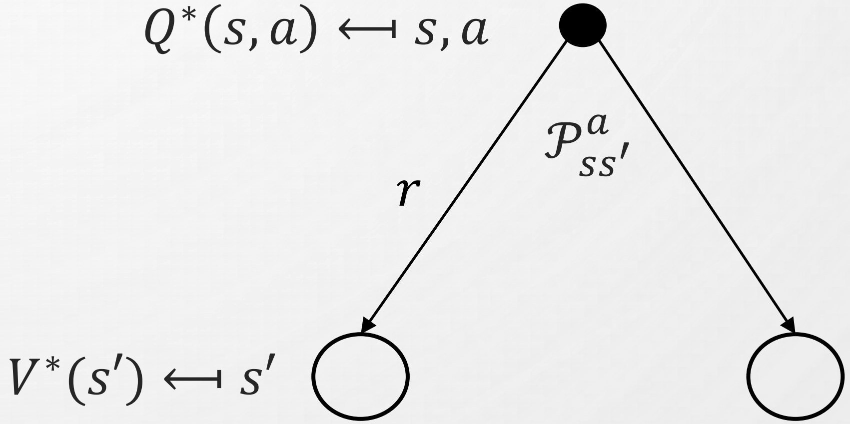
EQUAÇÃO DE BELLMAN DA OTIMALIDADE PARA MDP

Substituindo π por π^* nas equações (3) e (4), temos que as Funções de Valor Ótimas de Estados e Ações estão relacionadas por:

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s,a) \tag{5}$$

$$Q^*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} V^*(s') \tag{6}$$



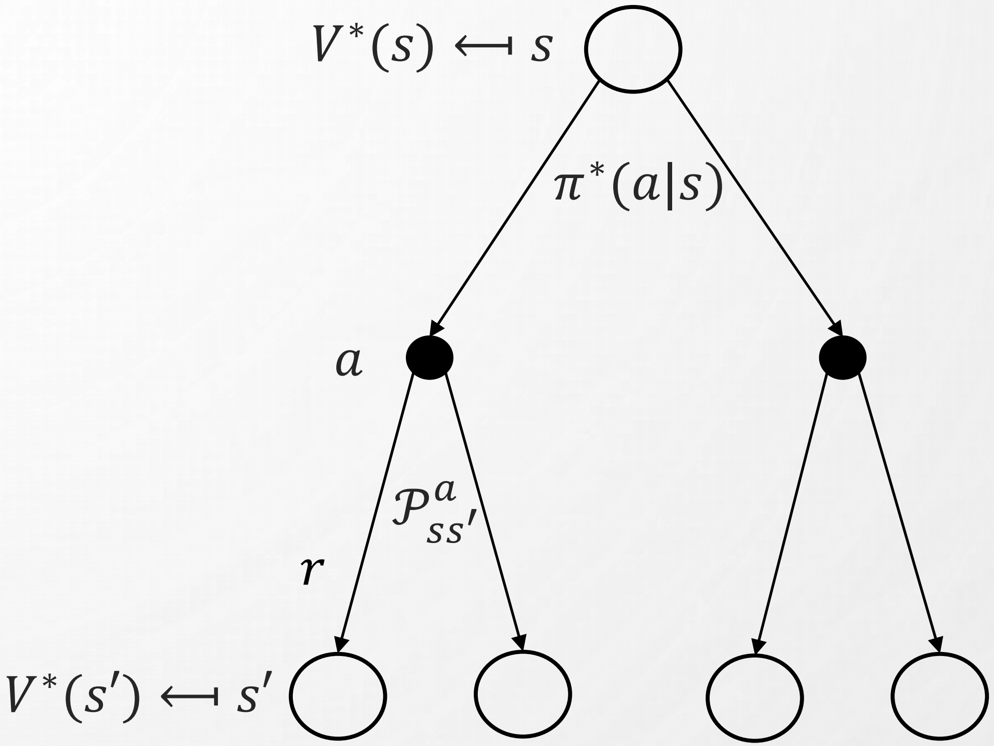


EQUAÇÃO DE BELLMAN DA OTIMALIDADE PARA V^*

Substituindo (6) em (5) podemos escrever $V^*(s)$ em função de $V^*(s')$:

$$V^*(s) = \max_{a \in \mathcal{A}} \left[\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} V^*(s') \right]$$

Bellman Optimality Equation for V^*

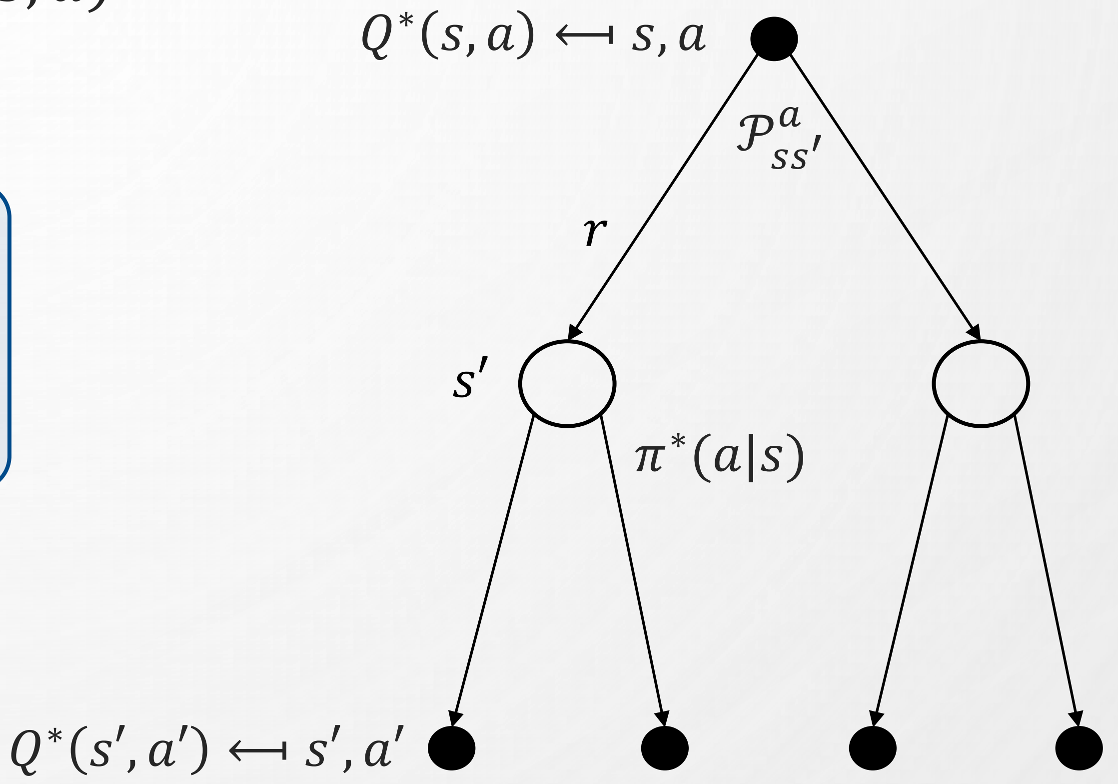


EQUAÇÃO DE BELLMAN DA OTIMALIDADE PARA Q^*

Substituindo (5) em (6) podemos escrever $Q^*(s,a)$ em função de $Q^*(s',a')$:

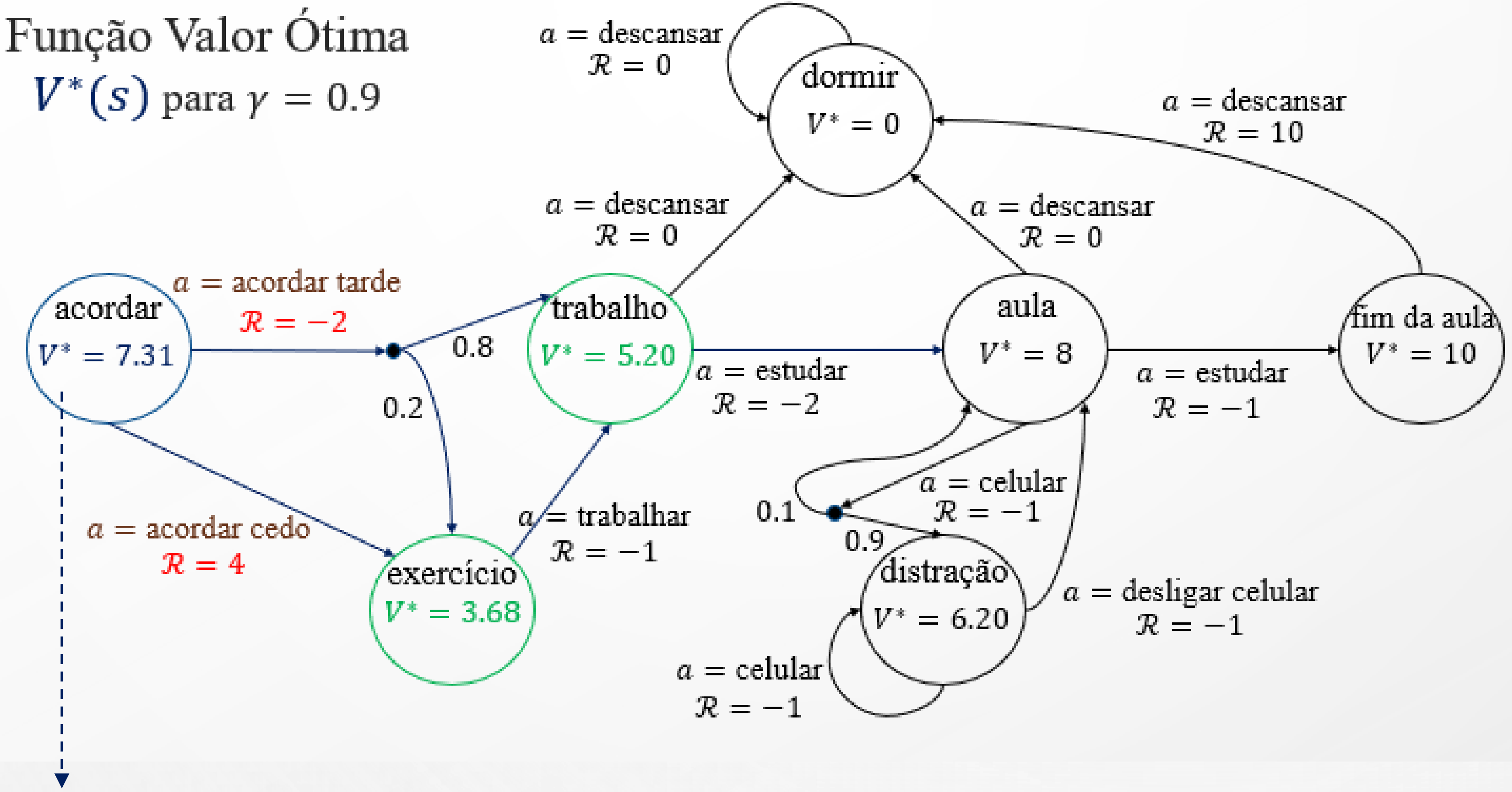
$$Q^*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} \max_{a' \in \mathcal{A}} Q^*(s',a')$$

Bellman Optimality Equation for Q^*



MDP: EXEMPLO – FUNÇÃO VALOR ÓTIMA $V^*(s)$

Função Valor Ótima
$V^*(s)$ para $\gamma = 0.9$



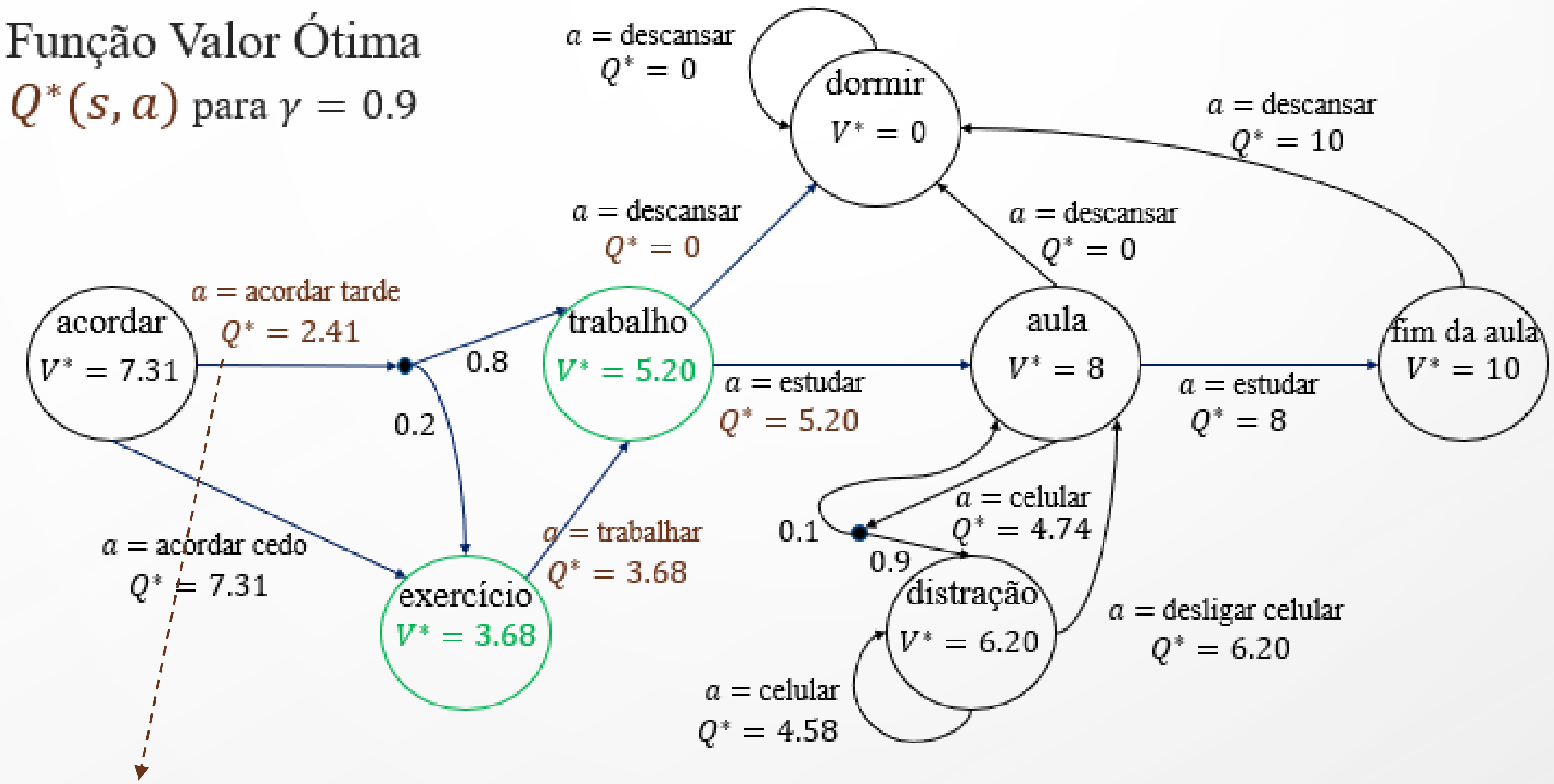
$$
\begin{aligned}
V^* &= \max\{4 + 0.9(3.68), \quad -2 + 0.9(0.2(3.68) + 0.8(5.20))\} \\
&= \max\{7.31, \quad 2.41\} \\
&= 7.31
\end{aligned}
$$

$$V^*(s) = \max_{a \in \mathcal{A}} \left[\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a V^*(s') \right]$$

MDP: EXEMPLO – FUNÇÃO VALOR ÓTIMA $Q^*(s,a)$

Função Valor Ótima

$Q^*(s,a)$ para $\gamma = 0.9$



$$
\begin{aligned}
Q^* &= -2 + 0.9[0.8 * \max\{0, 5.20\} + 0.2 * \max\{3.68\}] \\
&= -2 + 0.9[0.8 * 5.20 + 0.2 * 3.68] \\
&= 2.41
\end{aligned}
$$

$$Q^*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}^a_{ss'} \max_{a' \in \mathcal{A}} Q^*(s',a')$$

MDP: POLÍTICA ÓTIMA π^*

- Podemos comparar duas políticas da seguinte forma:

$$\pi \geq \pi' \Leftrightarrow V_\pi(s) \geq V_{\pi'}(s), \forall s \in \mathcal{S}$$

- A política ótima π^* é tal que $\pi^* \geq \pi, \forall \pi$.
- Dada a Função Valor das Ações Ótima $Q^*(s,a)$, podemos obter uma política ótima escolhendo, para cada estado, a ação de maior valor Q:

$$\pi^*(a|s) = \begin{cases} 1, & se\ a = \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, Q^*(s,a) \\ 0, & caso\ contrário \end{cases}$$

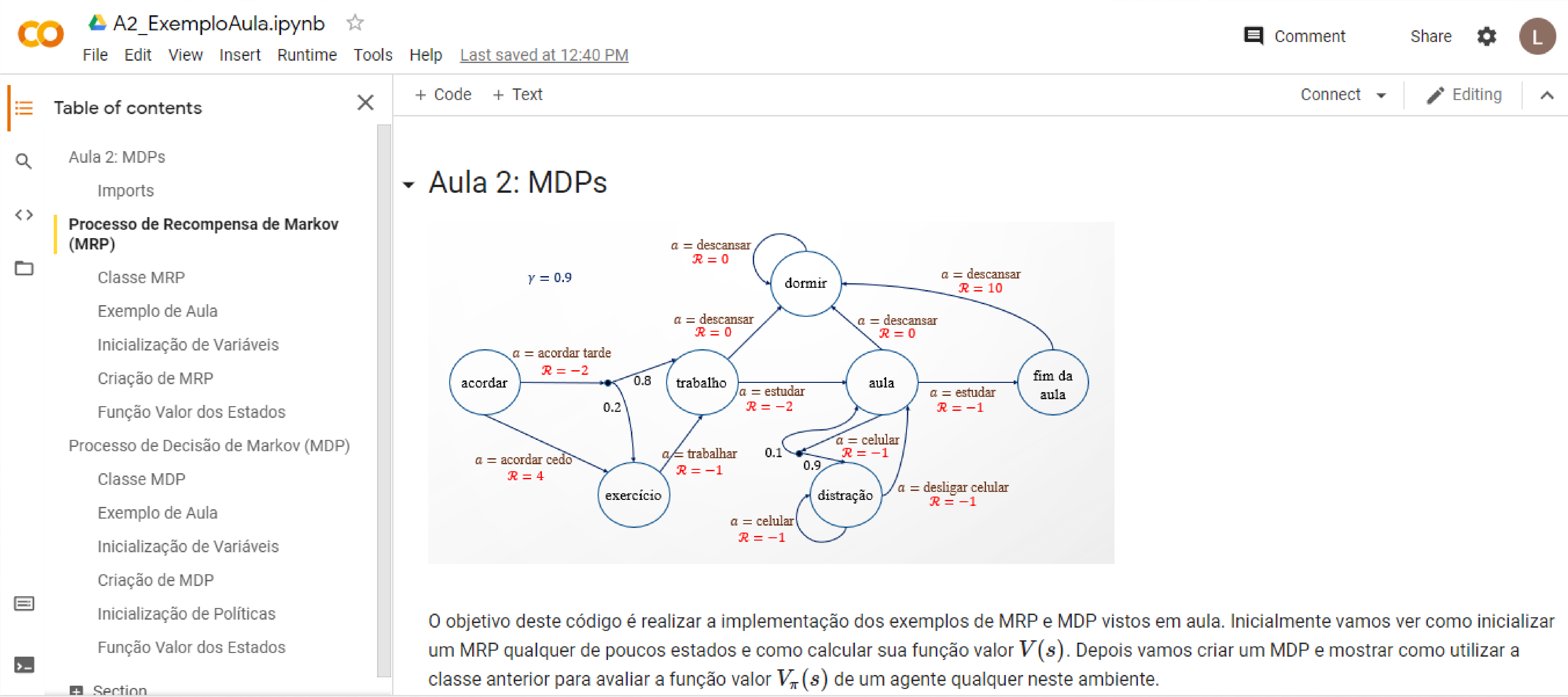
EQUAÇÃO DE BELLMAN DA OTIMALIDADE PARA MDP

- A Equação de Bellman da Otimalidade é não linear e não possui solução no caso geral.
- Existem diversos métodos iterativos para resolvê-la (próxima aula):
  - Value Iteration
  - Policy Iteration
  - SARSA
  - Q-Learning

$$V^*(s) = \max_{a \in \mathcal{A}} \left[\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} V^*(s') \right]$$

$$Q^*(s,a) = \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} \max_{a' \in \mathcal{A}} Q^*(s',a')$$

MDP EXAMPLE: PYTHON NOTEBOOK



MDP EXAMPLE: SUPPLY CHAIN

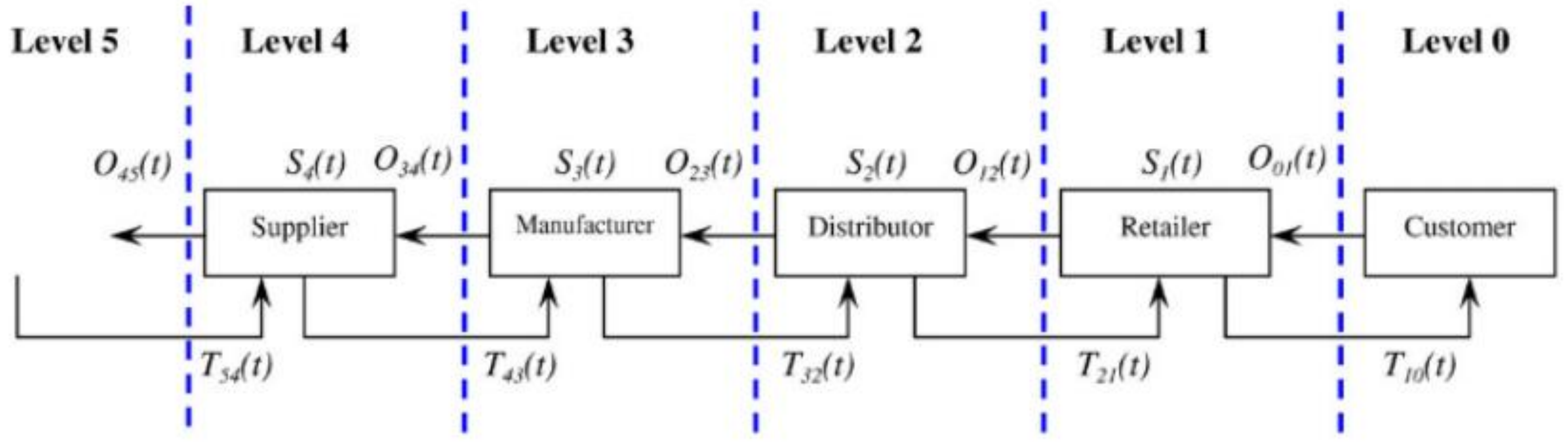
Exemplo: Supply Chain

MDP EXAMPLE: SUPPLY CHAIN

- Gestão de Cadeia Logística (Supply Chain Management):

Considere o problema de manutenção de inventário de determinado produto, em que cada nível da cadeia mantém uma quantidade do produto e fornece para o nível inferior (Beer Distribution Game)

- $S_i(t)$: Estoque do nível i no instante de tempo t.
- $O_{ij}(t)$: Tamanho de pedido do nível i para nível j no instante de tempo t.
- $T_{ij}(t)$: Distribuição do nível i para o nível j no instante de tempo t.



Fonte: Geevers, K. **Deep Reinforcement Learning in Inventory Management**, 2020

MDP EXAMPLE: SUPPLY CHAIN

- Espaço de Estados $\mathcal{S}$:

Neste problema, o espaço de estados $\mathcal{S}$ é o conjunto de todas as combinações de estoques entre todos os níveis. Se cada estoque pode possuir as seguintes quantidades do produto $\mathrm{S_i} \in \{0,1,\dots,9\}$, temos que o número total de estados deste MDP é:

$$|\mathcal{S}| = 10^4 = 10\,000$$

Um estado qualquer é dado por:
$s = [S_4, S_3, S_2, S_1]$

- Espaço de Ações $\mathcal{A}$:

Cada ação é dada pelo conjunto de pedidos de cada nível. Se, por questões logísticas, limitamos o tamanho de um pedido para $O_i \in \{0,1,2,3\}$, temos que o número total de ações deste MDP é:

$$|\mathcal{A}| = 4^4 = 256$$

Uma ação qualquer é dada por:
$\mathrm{a} = [O_{45}, O_{34}, O_{23}, O_{12}]$

MDP EXAMPLE: SUPPLY CHAIN

- Matriz de Transição de Estados:

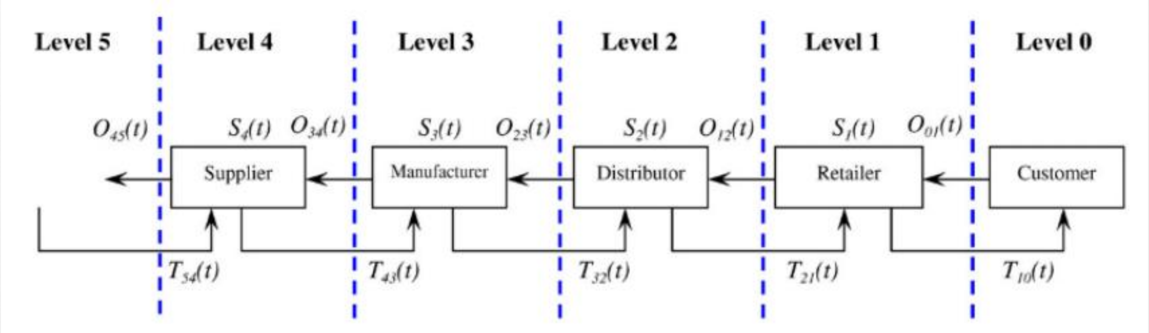
A dinâmica deste MDP é dada pela atualização dos estoques de cada nível em função do balanço entre pedidos obtidos pelo nível superior e distribuições fornecidas ao nível inferior:

$$S_i(t+1) = S_i(t) + O_{i,i+1}(t) - T_{i,i-1}(t)$$

Onde a distribuição $T_{i,i-1}(t)$ só pode ser satisfeita se houver disponibilidade no estoque:

$$T_{i,i-1}(t) = \min\{S_i(t); \quad O_{i-1,i}(t)\}$$

A partir desta formulação é possível escrever um conjunto de matrizes de transições de estados $\{\mathcal{P}_{ss'}^a\}_{a\in\mathcal{A}}$



MDP EXAMPLE: SUPPLY CHAIN

- Função de Recompensa:

A função de recompensa deve considerar o custo de manutenção dos estoques e penalizar falhas de fornecimento de um pedido (backorders). Uma possível formulação é:

$$\mathcal{R}(s = [S_4, S_3, S_2, S_1], a = [O_{45}, O_{34}, O_{23}, O_{12}\ ]) = -\sum_{i=1}^{4} \alpha_i \underbrace{S_i}_{\text{Custo de manutenção de estoque}} + \beta_i \underbrace{\left|T_{i,i-1} - O_{i-1,i}\,(t)\right|}_{\text{Falha de Fornecimento}}$$

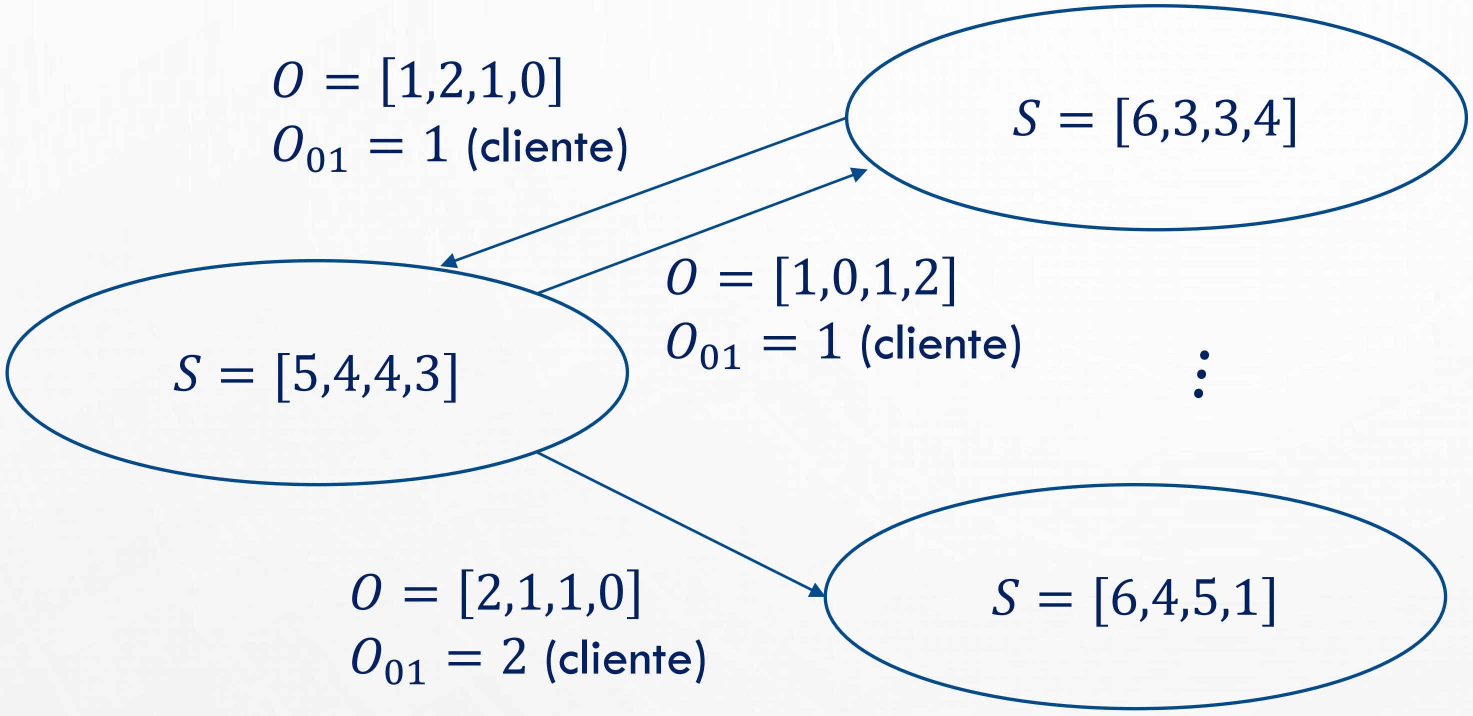
Os pesos α_i e β_i podem ser dados por:

$$\alpha = [1,1,1,1]$$

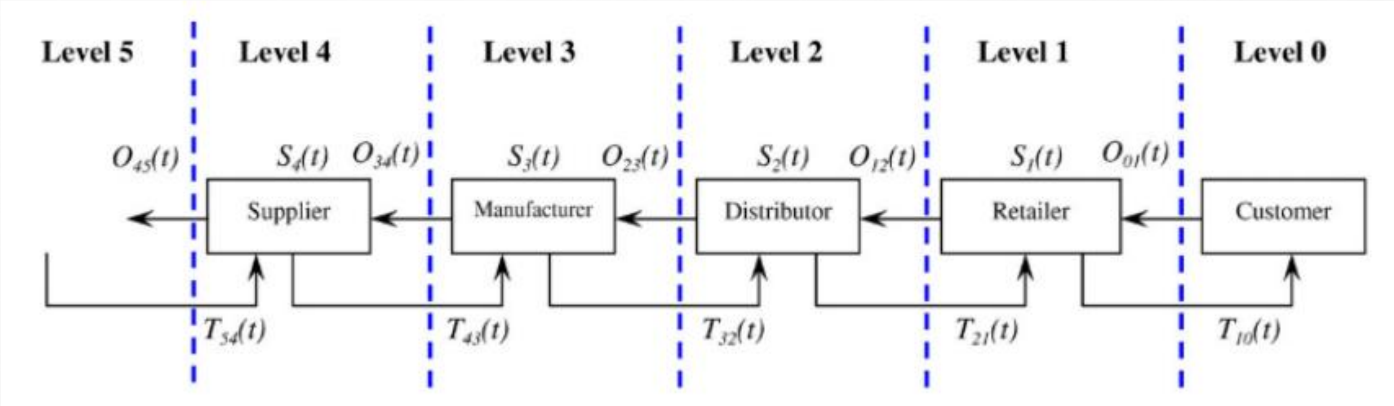
$$\beta = [2,2,2,3]$$

Penalizando principalmente a falta de disponibilidade ao cliente.

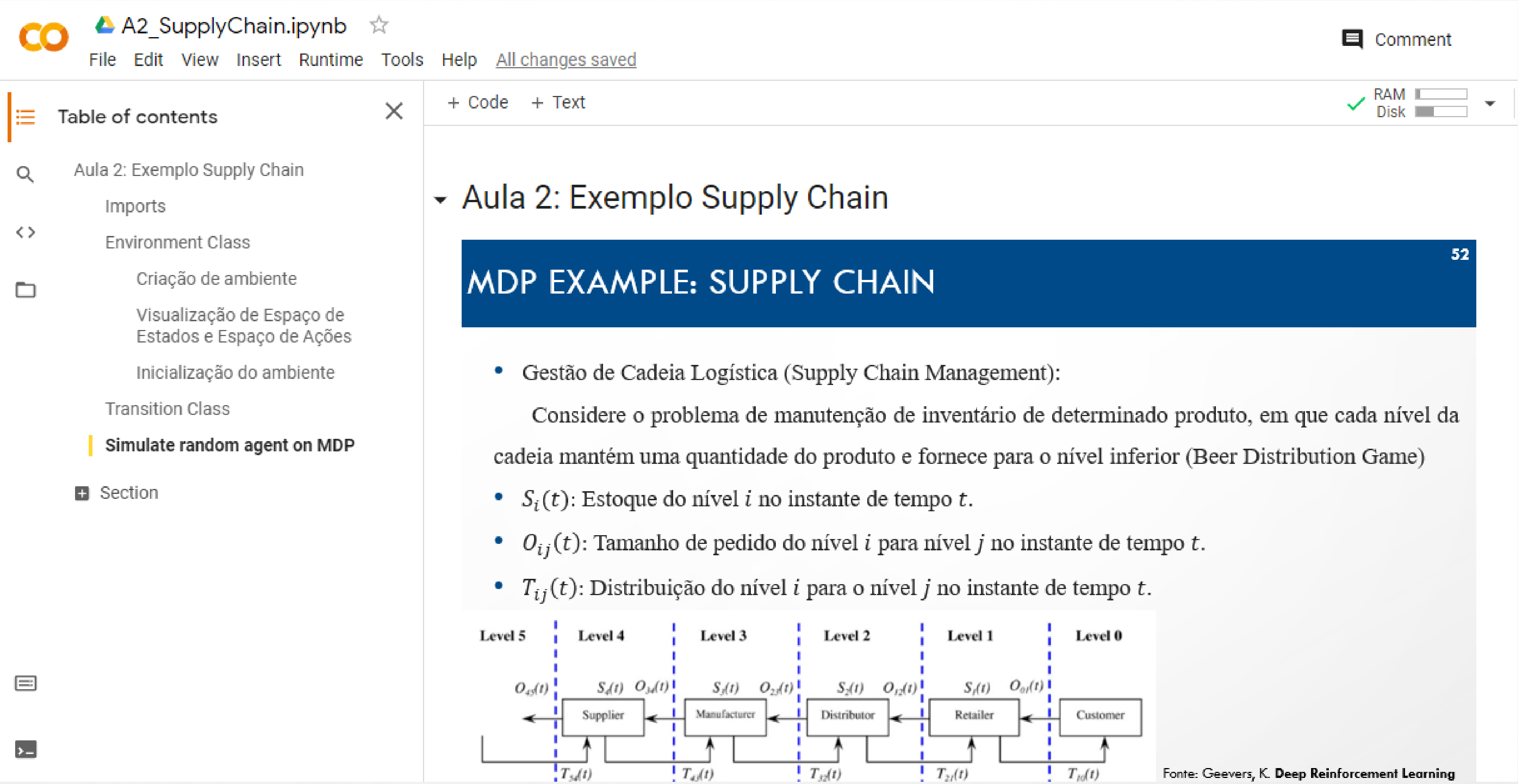
MDP EXAMPLE: SUPPLY CHAIN



- MDP é extenso demais para ser representado por completo
- $|\mathcal{S}| = 10\ 000$
- $|\mathcal{A}| = 256$



MDP EXAMPLE: SUPPLY CHAIN – PYTHON NOTEBOOK



EXTENSÕES DE MDPS

Extensões de MDPs

EXTENSÕES DE MDPS

Existem problemas que só podem ser modelados pelas seguintes extensões de MDPs:

- MDPs infinitos/contínuos.
- MDPs parcialmente observáveis.

Alguns exemplos de aplicações de extensões de MDPs são:

- Controle de articulações em manipulador robótico (Ações, dadas pelas tensões em cada motor, são contínuas. Espaço de estados, posições e velocidades de cada articulação, é contínuo).
- Agente que joga poker (cartas de outros jogadores são desconhecidas).

MDPS INFINITOS

MDPs infinitos possuem três categorias:

- Espaços de estados/ações com número infinito, mas contável, de elementos.
  - Análogo a MDP finito, mesmos métodos de aprendizado são aplicáveis.
- Espaços de estados/ações contínuos.
  - Solução fechada para LQR (Linear-Quadratic Regulator).
- Tempo Contínuo
  - Agente interage em tempo contínuo, não somente em épocas de decisão.
  - Equação de Hamilton-Jacobi-Bellman (Equação de Bellman para $\Delta t \rightarrow 0$).
  - Solução requer equações diferenciais parciais.

MDP PARCIALMENTE OBSERVÁVEL (POMDP)

Um Processo de Decisão de Markov Parcialmente Observável (POMDP) é uma tupla $< \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{P}, \mathcal{R}, \mathcal{Z}, \gamma >$, onde:

- $\mathcal{S}$ é um conjunto finito de estados.
- $\mathcal{A}$ é um conjunto finito de ações.
- $\mathcal{O}$ é um conjunto finito de observações.
- $\mathcal{P}$ é uma função $\mathcal{P}$: $\mathcal{S}x\mathcal{A}x\mathcal{S} \rightarrow [0,1] \subset \mathbb{R}$ de probabilidades de transições de estados.

  $\mathcal{P}_{ss'}^{a} = \mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a]$
- $\mathcal{R}$ é uma função de recompensa $\mathcal{R}$: $\mathcal{S}x\mathcal{A} \rightarrow \mathbb{R}$ tal que $\mathcal{R}(s,a) = \mathbb{E}[R_{t+1} | S_t = s, A_t = a]$
- $\mathcal{Z}$ é uma função de observação $\mathcal{Z}_{s'o}^{a} = \mathbb{P}[O_{t+1} = o | S_{t+1} = s', A_t = a]$
- $\gamma \in [0,1] \subset \mathbb{R}$ é um fator de desconto.

MDP PARCIALMENTE OBSERVÁVEL (POMDP) – BELIEF STATES

Em um POMDP o estado atual não é conhecido. Em vez disso, temos uma distribuição de probabilidades sobre todo espaço de estados condicionada na história, essa distribuição é denominada *Belief State* $b(h)$:

$$b(h) = \begin{bmatrix} \mathbb{P}[S_t = s_1 | H_t = h] \\ \vdots \\ \mathbb{P}[S_t = s_n | H_t = h] \end{bmatrix}$$

onde $H_t = O_0, A_0, R_1, \ldots, O_{t-1}, A_{t-1}, R_t$.

- Assim, não é mais possível afirmar com certeza em qual estado o ambiente se encontra e é preciso tomar decisões com base em probabilidades do ambiente se encontrar em cada estado.

EXERCÍCIO E_1

Exercício E_1

EXERCÍCIO E_1: PROCESSOS DE DECISÃO DE MARKOV

Exercício E1 – Processos de Decisão de Markov (MDPs)

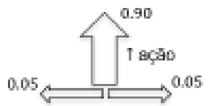
1) Dado um ambiente do tipo *Grid World* composto por 6 casas em duas fileiras, conforme a figura abaixo:

s_0		

Considere um robô móvel na posição s_0 do ambiente com capacidade de se movimentar nas 4 direções: $\mathcal{A} = \{N = \uparrow, E = \rightarrow, S = \downarrow, W = \leftarrow\}$. A posição de destino do robô é representada pela casa verde, enquanto a casa cinza representa um obstáculo (ambos são estados terminais, qualquer ação tomada mantém o agente no próprio estado). Uma Função de Recompensa $\mathcal{R}$ que representa a tarefa de posicionamento do robô é ilustrada abaixo, indicando a recompensa obtida por um agente ao tomar qualquer ação na casa correspondente ($\mathcal{R}(s,.)$):

-0.05	-1	$+1$
-0.05	-0.05	-0.05

Quando o robô executa uma ação, ele se move para a casa vizinha na direção escolhida em 90% das vezes, com 5% de chance de ocorrer um escorregamento em cada direção perpendicular à ação tomada. Se o robô fosse colidir com uma parede ele em vez disso permanece na mesma posição.



a) Faça um diagrama parcial do MDP associado a esse problema, representando todas as transições de estados, ações e recompensas a partir do estado inicial

- Exercício E_1 já disponível no OpenLMS.
- Tópico: Processo de Decisão de Markov
  - a) Diagrama de MDP.
  - b) Cálculo de Retorno G_0.
  - c) Equação de Bellman e Função Valor $V(s)$.
- Entrega: Até 04/04 até 23:59

REFERÊNCIAS BIBLIOGRÁFICAS

[1] Sutton, R. and Barto, A. *Reinforcement Learning: An Introduction*, The MIT Press (2020). [Cp. 3]

[2] https://towardsdatascience.com/understanding-the-markov-decision-process-mdp-8f838510f150

[3] https://pymdptoolbox.readthedocs.io/en/latest/api/mdp.html

[4] https://github.com/MeepMoop/MDPy

[5] Geevers, K. *Deep Reinforcement Learning in Inventory Management*, 2020

DÚVIDAS

Muito obrigado a todos!

Dúvidas