Inteligência Artificial para Robótica Móvel

CT-213

Instituto Tecnológico de Aeronáutica

Relatório do Laboratório 10 - Programação Dinâmica

Leonardo Peres Dias

1 de junho de 2025

Sumário

1 Breve Explicação em Alto Nível da Implementação

1.1 Avaliação de Política

A avaliação de política tem como objetivo estimar a função de valor $V^\pi(s)$ para uma política fixa π, ou seja, o valor esperado do retorno acumulado ao seguir π a partir de cada estado s. A equação de Bellman para avaliação de política é dada por:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s, a) V^\pi(s') \right], \tag{1}$$

onde:

- $\pi(a \mid s)$ é a probabilidade de executar a ação a no estado s sob a política π;
- $R(s,a)$ é a recompensa esperada ao executar a ação a no estado s;
- $P(s' \mid s, a)$ é a probabilidade de transição para o estado s' ao executar a ação a no estado s;
- γ é o fator de desconto ($0 \leq \gamma \leq 1$).

Na implementação, o algoritmo itera sobre todos os estados válidos do ambiente e atualiza seus valores de acordo com a equação acima. A política é representada como uma distribuição de probabilidade sobre as ações, e o valor de cada estado é atualizado com base na expectativa sobre as transições e as recompensas associadas.

A atualização dos valores é feita de forma iterativa até que a variação máxima entre as iterações seja menor que uma tolerância ε, garantindo a convergência para V^π.

1.2 Iteracão de Valor

O método de *iteração de valor* busca encontrar a política ótima π^* iterando sobre a equação de Bellman de otimalidade até a convergência para a função de valor ótima $V^*(s)$. A equação de Bellman de otimalidade é dada por:

$$V^*(s) = \max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s' \mid s,a) V^*(s') \right], \tag{2}$$

onde:

- $V^*(s)$ é o valor ótimo do estado s;
- $R(s,a)$ é a recompensa esperada ao executar a ação a no estado s;
- $P(s' \mid s,a)$ é a probabilidade de transição para o estado s' dado o par (s,a);
- γ é o fator de desconto.

O algoritmo começa com uma estimativa inicial arbitrária para $V(s)$ (por exemplo, $V(s) = 0$) e atualiza os valores iterativamente conforme a equação acima. A cada iteração, calcula-se o valor de cada estado assumindo que se tomará sempre a melhor ação a partir daquele ponto.

A iteração é interrompida quando a variação máxima entre os valores antigos e os novos for menor que uma tolerância ε, indicando convergência numérica para V^*. Após a convergência, uma política ótima π^* pode ser obtida extraindo a ação que maximiza a equação de Bellman para cada estado:

$$\pi^*(s) = \arg\max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^*(s') \right]. \tag{3}$$

1.3 Iteração de Política

A *iteração de política* é um método iterativo que alterna entre: a **avaliação de política** e a **melhoria de política por busca gulosa**. O objetivo é encontrar a política ótima π^* para um problema de decisão de Markov (MDP).

1. Avaliação de Política Dada uma política fixa π, estima-se a função de valor $V^\pi(s)$ resolvendo iterativamente a equação de Bellman para avaliação de política:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^\pi(s') \right]. \tag{4}$$

2. Melhoria de Política A partir da função de valor $V^\pi(s)$ obtida, a política é atualizada para uma nova política π' escolhendo, para cada estado, as ações que maximizam o valor esperado do retorno. Essa etapa utiliza a equação de Bellman de otimalidade:

$$\pi'(s) = \arg\max_{a \in \mathcal{A}} \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V^\pi(s') \right]. \tag{5}$$

2 Tabelas Comprovando Funcionamento do Código

2.1 Caso $p_c = 1.0$ e $\gamma = 1.0$

2.1.1 Avaliação de Política

-384.09	-382.73	-381.19	*	-339.93	-339.93
-380.45	-377.91	-374.65	*	-334.92	-334.93
-374.34	-368.82	-359.85	-344.88	-324.92	-324.93
-368.76	-358.18	-346.03	*	-289.95	-309.94
*	-344.12	-315.05	-250.02	-229.99	*
-359.12	-354.12	*	-200.01	-145.00	0.00

Tabela 1: Função valor obtida por avaliação de política.

SURDL	SURDL	SURDL	*	SURDL	SURDL
SURDL	SURDL	SURDL	*	SURDL	SURDL
SURDL	SURDL	SURDL	SURDL	SURDL	SURDL
SURDL	SURDL	SURDL	*	SURDL	SURDL
*	SURDL	SURDL	SURDL	SURDL	*
SURDL	SURDL	*	SURDL	SURDL	S

Tabela 2: Política utilizada na avaliação de política.

2.1.2 Iteracão de Valor

-10.00	-9.00	-8.00	*	-6.00	-7.00
-9.00	-8.00	-7.00	*	-5.00	-6.00
-8.00	-7.00	-6.00	-5.00	-4.00	-5.00
-7.00	-6.00	-5.00	*	-3.00	-4.00
*	-5.00	-4.00	-3.00	-2.00	*
-7.00	-6.00	*	-2.00	-1.00	0.00

Tabela 3: Função valor obtida por iteração de valor.

RD	RD	D	*	D	DL
RD	RD	D	*	D	DL
RD	RD	RD	R	D	DL
R	RD	D	*	D	L
*	R	R	RD	D	*
R	U	*	R	R	SURD

Tabela 4: Política derivada da iteração de valor.

2.1.3 Iteração de Política

-10.00	-9.00	-8.00	*	-6.00	-7.00
-9.00	-8.00	-7.00	*	-5.00	-6.00
-8.00	-7.00	-6.00	-5.00	-4.00	-5.00
-7.00	-6.00	-5.00	*	-3.00	-4.00
*	-5.00	-4.00	-3.00	-2.00	*
-7.00	-6.00	*	-2.00	-1.00	0.00

Tabela 5: Função valor obtida por iteração de política.

RD	RD	D	*	D	DL
RD	RD	D	*	D	DL
RD	RD	RD	R	D	DL
R	RD	D	*	D	L
*	R	R	RD	D	*
R	U	*	R	R	SURD

Tabela 6: Política resultante da iteração de política.

2.2 Caso $p_c = 0.8$ e $\gamma = 0.98$

2.2.1 Avaliação de Política

-47.19	-47.11	-47.01	*	-45.13	-45.15
-46.97	-46.81	-46.60	*	-44.58	-44.65
-46.58	-46.21	-45.62	-44.79	-43.40	-43.63
-46.20	-45.41	-44.42	*	-39.87	-42.17
*	-44.31	-41.64	-35.28	-32.96	*
-45.73	-45.28	*	-29.68	-21.88	0.00

Tabela 7: Função valor obtida por avaliação de política.

SURDL	SURDL	SURDL	*	SURDL	SURDL
SURDL	SURDL	SURDL	*	SURDL	SURDL
SURDL	SURDL	SURDL	SURDL	SURDL	SURDL
SURDL	SURDL	SURDL	*	SURDL	SURDL
*	SURDL	SURDL	SURDL	SURDL	*
SURDL	SURDL	*	SURDL	SURDL	S

Tabela 8: Política utilizada na avaliação de política.

2.2.2 Iteracão de Valor

-11.65	-10.78	-9.86	*	-7.79	-8.53
-10.72	-9.78	-8.78	*	-6.67	-7.52
-9.72	-8.70	-7.59	-6.61	-5.44	-6.42
-8.70	-7.58	-6.43	*	-4.09	-5.30
*	-6.43	-5.17	-3.87	-2.76	*
-8.63	-7.58	*	-2.69	-1.40	0.00

Tabela 9: Função valor obtida por iteração de valor.

D	D	D	*	D	D
D	D	D	*	D	D
RD	D	D	R	D	D
R	RD	D	*	D	L
*	R	R	D	D	*
R	U	*	R	R	S

Tabela 10: Política derivada da iteração de valor.

2.2.3 Iteração de Política

-11.65	-10.78	-9.86	*	-7.79	-8.53
-10.72	-9.78	-8.78	*	-6.67	-7.52
-9.72	-8.70	-7.59	-6.61	-5.44	-6.42
-8.70	-7.58	-6.43	*	-4.09	-5.30
*	-6.43	-5.17	-3.87	-2.76	*
-8.63	-7.58	*	-2.69	-1.40	0.00

Tabela 11: Função valor obtida por iteração de política.

D	D	D	*	D	D
D	D	D	*	D	D
R	D	D	R	D	D
R	D	D	*	D	L
*	R	R	D	D	*
R	U	*	R	R	S

Tabela 12: Política resultante da iteração de política.

3 Discussão dos Resultados

Foram realizados experimentos em dois cenários distintos: um ambiente determinístico com $p_c = 1.0$ e fator de desconto $\gamma = 1.0$, e um ambiente estocástico com $p_c = 0.8$ e $\gamma = 0.98$. A comparação entre os resultados obtidos permite observar o impacto da incerteza e da depreciação temporal na função de valor e nas políticas ótimas.

Na **avaliação de política**, os valores do primeiro cenário atingem até -380, refletindo longos caminhos esperados sob uma política aleatória e sem desconto. Já no ambiente estocástico, os valores são mais suaves (por volta de -45), pois o desconto atenua o custo acumulado.

Tanto na **iteração de valor** quanto na **iteração de política**, observou-se convergência para a mesma política ótima nos dois cenários, indicando robustez dos métodos mesmo sob incerteza. A diferença principal aparece na funçao de valor, que possui menos ruído no ambiente determinístico.