ANÁLISE DA DEFINIÇÃO DE PARÂMETROS DO APRENDIZADO POR REFORÇO NO DESEMPENHO DE UM MANIPULADOR ROBÓTICO [1]

André Luiz Carvalho Ottoni[2]
Erivelton Geraldo Nepomuceno[3]
Marcos Santos de Oliveira[4]

RESUMO

O objetivo deste trabalho é analisar os efeitos da definição de parâmetros do Aprendizado por Reforço no desempenho de um manipulador robótico. Para isso, são realizados experimentos com um manipulador simulado do tipo SCARA. Além disso, é adotada a modelagem matemática via Metodologia de Superfície de Resposta. Os resultados apontam uma tendência no aumento de desempenho do manipulador quando a taxa de aprendizado e o fator de desconto se aproximam simultaneamente de zero.

Palavras-chave: Aprendizado por Reforço. Manipulador Robótico. Metodologia de Superfície de Resposta.

1 INTRODUÇÃO

O Aprendizado por Reforço (AR) é um importante campo da Aprendizagem de Máquina (*Machine Learning*) (WATKINS; DAYAN, 1992; SUTTON; BARTO, 1998). No AR, um agente aprende em uma interação direta com um ambiente por meio de sucessos (reforços) e fracassos (penalidades). Nesse sentido, o agente acumula experiência e aprende as melhores ações a serem executadas em cada situação do ambiente (estado) (SUTTON; BARTO, 1998).

Um ponto chave nos experimentos de AR é a definição dos valores de parâmetros dos algoritmos de aprendizado, como taxa de aprendizado (α) e fator de desconto (γ). Nessa linha, diversos trabalhos já demonstraram que o desempenho do AR pode ser influenciado por esses

[1] **Como citar este artigo:**
OTTONI, A. L. C; NEPOMUCENO, E. G; OLIVEIRA, M. S. Análise da definição de parâmetros do aprendizado por reforço no desempenho de um manipulador robótico. **ForScience**: revista científica do IFMG, Formiga, v. 5, n. 3, e00267, jul./dez. 2017.

[2] Mestre em Engenharia Elétrica pela Universidade Federal de São João del-Rei (UFSJ). Professor contratado do Departamento de Engenharia Mecatrônica do CEFET-MG / Campus Divinópolis. Currículo Lattes: <http://lattes.cnpq.br/2003401420560517>. E-mail: andreottoni@ymail.com.

[3] Doutor em Engenharia Elétrica pela Universidade Federal de Minas Gerais (UFMG). Professor Associado do Departamento de Engenharia Elétrica da UFSJ. Currículo Lattes: <http://lattes.cnpq.br/5858842244938381 >. E-mail: nepomuceno@ufsj.edu.br.

[4] Doutor em Estatística e Experimentação Agropecuária pela Universidade Federal de Lavras (UFLA). Professor do Departamento de Matemática e Estatística da UFSJ. Currículo Lattes: <http://lattes.cnpq.br/2561582437711389>. E-mail: mso@ufsj.edu.br.

parâmetros (SUTTON; BARTO, 1998; EVEN-DAR; MANSOUR, 2003; SCHWEIGHOFER; DOYA, 2003). Assim, várias pesquisas apresentaram técnicas para o ajuste dinâmico/adaptativo dos parâmetros do AR de acordo com a situação do aprendizado (SCHWEIGHOFER; DOYA, 2003; MURAKOSHI; MIZUNO, 2004). No entanto, um método muito comum ainda é a definição desses parâmetros em valores constantes durante todo o aprendizado (BECK; SRIKANT, 2012). Nesse aspecto, uma abordagem recente de análise dos efeitos dos parâmetros do AR é a modelagem matemática por meio técnicas estatísticas (OTTONI et al., 2016; OTTONI; NEPOMUCENO; OLIVEIRA, 2016). Ottoni et al. (2016) aplicam Regressão Logística na análise da definição dos parâmetros dos algoritmos Q-learning e SARSA em um ambiente de navegação simulada (*gridworld*). Já Ottoni, Nepomuceno e Oliveira (2016), abordam a Metodologia de Superfície de Resposta (RSM) para a estimação dos valores de α e γ na aplicação do Problema do Caixeiro Viajante.

Além disso, os trabalhos de Ottoni et al. (2016) e Ottoni, Nepomuceno e Oliveira (2016) ressaltam a importância de investigar aplicação da modelagem estatística dos parâmetros em outros domínios tradicionais do AR. Nesse sentido, uma área que possui importantes aplicações do AR é o campo dos manipuladores robóticos (THAM; PRAGER, 1993; HERNANDEZ; LOPE, 2007; PARK; KIM; SONG, 2007; LIN, 2009; KIM et al., 2010; MILJKOVIć et al., 2013; TANG; LIU; TONG, 2014). No entanto, apesar do AR ser frequentemente adotado em pesquisas com braços robóticos, a literatura carece de uma metodologia para a análise dos efeitos dos parâmetros de aprendizado nesse domínio. Tomando como exemplos os trabalhos de Tham e Prager (1993), Park, Kim e Song (2007), Kim et al. (2010), em nenhum deles fica claro qual a estratégia adotada para a definição de α e γ.

Dessa forma, o objetivo deste trabalho é analisar os efeitos da definição dos parâmetros do AR no desempenho de simulações de um manipulador SCARA, um tipo de braço robótico muito adotado na indústria e em pesquisas (VISIOLI; LEGNANI, 2002; HERNANDEZ; LOPE, 2007; LIN, 2009). Para isso, será proposta uma metodologia baseada no trabalho de Ottoni, Nepomuceno e Oliveira (2016), adotando a modelagem matemática via RSM (MYERS; MONTGOMERY; Anderson-Cook, 2009) para a estimação dos parâmetros de aprendizado.

Este trabalho está organizado em seções. A seção 2 apresenta conceitos teóricos iniciais de AR. Em seguida, a seção 3 descreve a metodologia do trabalho. Já a seção 4 descreve a análise dos resultados. Finalmente, na seção 5 são apresentadas as conclusões.

2 APRENDIZADO POR REFORÇO

O Aprendizado por Reforço é formulado a partir dos Processos de Decisão de Markov (PDM) (SUTTON; BARTO, 1998). Em um PDM, uma regra de decisão é o mapeamento de

estados em ações (PELLEGRINI; WAINER, 2007). Assim, o AR consiste em aprender as melhores ações para as situações do sistema (estados), com o objetivo de maximizar ao longo do tempo o valor da recompensa (SUTTON; BARTO, 1998).

Um dos algoritmos mais adotados no AR é o SARSA (SUTTON; BARTO, 1998). O SARSA é uma modificação de outro tradicional algoritmo, o Q-learning (WATKINS; DAYAN, 1992). O algoritmo SARSA recebeu esse nome pois envolve na sua atualização os termos: s (estado no instante t), a (ação executada no instante t), $r(s,a)$ (reforço para o par $s \times a$), s' (estado no instante $t+1$) e a' (ação executada no instante $t+1$).

A Equação (1) descreve a atualização da matriz de aprendizado Q pelo SARSA, com a execução da ação a no estado s:

$$Q_{t+1} = Q_t(s,a) + \alpha[r(s,a) + \gamma Q_t(s',a') - Q_t(s,a)]. \quad (1)$$

O Algoritmo 1 retrata o Algoritmo SARSA, em que, α, γ e ε são os parâmetros de aprendizado, definidos entre 0 e 1:

- α é a taxa de aprendizado. Responsável por controlar a velocidade que as novas informações incidem sobre as experiências já armazenadas na matriz de aprendizado.
- γ é o fator de desconto. Controla o grau de influência das recompensas futuras sobre a recompensa no instante t.
- ε é o parâmetro da política de seleção de ações ε-greedy. Essa política controla o nível de exploração (seleção de ações aleatórias para o estado) e explotação (seleção gulosa).

Algoritmo 1 – Algoritmo SARSA.

Definir os parâmetros: α, γ e ε
Para cada ***s,a*** inicialize ***Q(s,a)***=0
Observe o estado ***s***
Selecione a ação ***a*** usando a política ε-greedy
Enquanto o critério de parada não for satisfeito {
    Execute a ação ***a***
    Receba a recompensa imediata ***r(s,a)***
    Observe o novo estado ***s'***
    Selecione a nova ação ***a'*** usando a política ε-greedy
    Atualize Q(s,a) com a Equação (1)
    s = s'
    a = a'
}

3 METODOLOGIA

A metodologia proposta neste trabalho visa analisar os efeitos dos parâmetros do AR no desempenho de um manipulador robótico e compreende cinco etapas:

1. Definição do ambiente de estudo do manipulador robótico (simulador).
2. Planejamento e realização dos experimentos.
3. Análise dos resultados gerais.
4. Modelagem matemática via Metodologia de Superfície de Resposta.
5. Análise dos resultados do modelo de RSM proposto.

Na sequência, as duas primeiras etapas da metodologia são detalhadas. Já os tópicos de 3 à 5 são abordados na seção 4 de análise dos resultados.

3.1 Simulador adotado

O simulador adotado neste trabalho foi desenvolvido pelo Prof. José Antônio Martin Hernández da *Universidad Complutense de Madrid* (UCM). O código, escrito em *MATLAB*®, está disponível para *download* na página pessoal do Prof. Hernández (HERNANDEZ,).

Em linhas gerais, o ambiente de simulação escolhido permite a realização de experimentos de AR com um manipulador SCARA. Já o algoritmo padrão implementado no código é o SARSA.

A Figura 1 apresenta a interface gráfica gerada pelo simulador adotado, em que é mostrado o modelo físico do manipulador SCARA simulado (HERNANDEZ; LOPE, 2007).

Conforme apresentado por Hernandez e Lope (2007), os parâmetros de Denavit-Hartenberg (*D-H parameters*) (DENAVIT; HARTENBERG, 1955) para o manipulador SCARA simulado são definidos na Tabela 1. Já a matriz de cinemática direta é apresentada na Equação (2) (HERNANDEZ; LOPE, 2007).

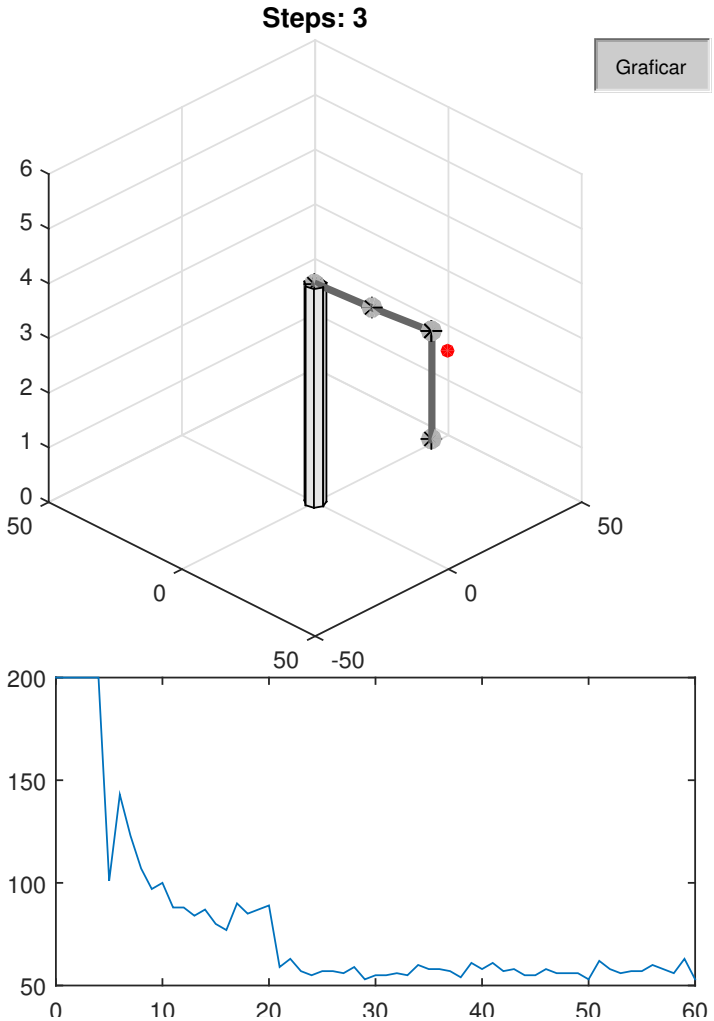

Figura 1 – Interface gráfica gerada pelo simulador adotado no *software* MATLAB. A interface gráfica é composta por duas imagens geradas em conjunto. O gráfico superior da interface plota o manipulador em três dimensões e mostra a sequência de passos executados. Já a imagem inferior, apresenta o gráfico de evolução do aprendizado (episódios $\times$ número de passos necessários para alcançar o objetivo).

Tabela 1 – Parâmetros de Denavit-Hartenberg para o manipulador SCARA simulado.

i	θ_i	d_i	a_i	α_i
1	θ_1	0	$a_1 = 20$	0
2	θ_2	0	$a_2 = 20$	π
3	θ_3	0	0	0
4	0	d_4	0	0

Fonte: Hernandez e Lope (2007).

$$T^4 = \begin{pmatrix} C_{12-3} & -S_{12-3} & 0 & a_1C_1 + a_2C_{12} \\ S_{12-3} & C_{12-3} & 0 & a_1S_1 + a_2S_{12} \\ 0 & 0 & -1 & -d_4 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad (2)$$

em que, θ_1, θ_2 e θ_3 são os ângulos das articulações do manipulador; a_1 e a_2 são os comprimentos dos elos dos braços; C_{12-3} é $cos(\theta_1 + \theta_2 - \theta_3)$, S_{12-3} é $sen(\theta_1 + \theta_2 - \theta_3)$, C_1 é $cos(\theta_1)$, S_1 é $sen(\theta_1)$, C_{12} é $cos(\theta_1 + \theta_2)$ e S_{12} é $sen(\theta_1 + \theta_2)$ (HERNANDEZ; LOPE, 2007).

3.2 Experimentos realizados

Os experimentos realizados visaram avaliar como a definição dos parâmetros α e γ pode influenciar no desempenho de aprendizado do manipulador simulado. Para isso, foi abordada uma metodologia experimental baseada em trabalhos recentes (OTTONI; NEPOMUCENO; OLIVEIRA, 2016; OTTONI et al., 2016). Nesse sentido, foram realizados experimentos envolvendo um conjunto de 64 combinações dos parâmetros taxa de aprendizado e fator de desconto. Os valores para α e γ adotados são:

- α: [0,01, 0,15, 0,30, 0,45, 0,60, 0,75, 0,90 e 0,99].
- γ: [0,01, 0,15, 0,30, 0,45, 0,60, 0,75, 0,90 e 0,99].

Além disso, cada combinação foi simulada em cinco épocas (repetições independentes) com 1000 episódios. Cada época é uma repetição independente, ou seja, o aprendizado é acumulado ao longo dos 1000 episódios e zerado sempre ao início de uma época. Vale ressaltar também que, a medida de desempenho analisada é o número de passos (movimentos) necessários para o manipulador sair da posição inicial (0, 0, 0) e alcançar o ponto final da trajetória (30, 20, 3) em um episódio. Ou seja, cada episódio de aprendizado é composto por interações, em que, para cada interação, o robô executa um movimento referente a atuação de uma das três juntas. Nesse sentido, o objetivo é que o manipulador aprenda uma sequência de atuação das juntas de modo a minimizar o número de movimentos para que o robô alcance o fim da trajetória em um episódio.

O número de movimentos do manipulador em um episódio foi limitado em 200 passos. Assim, um episódio é encerrado se o manipulador finaliza a trajetória ou alcança o número máximo de movimentos permitidos.

A função de reforço pré-definida no simulador adotado é apresentada no trabalho de Hernandez e Lope (2007) e dada pela Equação (3):

$$R = \frac{\beta}{1 + d^n}. \tag{3}$$

De acordo com Hernandez e Lope (2007), β é um escalar que controla a magnitude absoluta do reforço, d é a distância Euclidiana entre a posição atual do manipulador e a posição do objetivo, e n é o expoente utilizado para controlar a forma como a função de reforço responde a distância Euclidiana. Neste trabalho, foram adotados os valores pré-definidos no simulador para β e n, sendo $\beta = 10^8$ e $n = 2$.

Quanto à política de seleção de ações $\varepsilon - greedy$, o cálculo do parâmetro ε foi mantido como pré-definido no simulador, e é apresentado na Equação (4):

$$\varepsilon_t = \varepsilon_{t-1} \times k_\varepsilon, \tag{4}$$

em que, o parâmetro ε_t (instante t) é igual ao valor de ε_{t-1} (instante $t-1$) vezes uma constante de decaimento ($k_\varepsilon = 0{,}99$). No instante $t = 0$, $\varepsilon_0 = 0{,}1$.

4 ANÁLISE DOS RESULTADOS

4.1 Resultados gerais

Nesta seção, são apresentados os resultados gerais dos experimentos realizados. Para cada época (repetição) de uma combinação foi calculada a média de passos necessários para o manipulador chegar ao fim da trajetória ao longo dos 1000 episódios. Dessa forma, para cada combinação foram calculados cinco valores de média. Em seguida, foi considerado o melhor desempenho (menor média de movimentos) de cada uma das combinações. A Tabela 2 apresenta esses valores médios encontrados.

Tabela 2 – Menor média de movimentos por combinação.

$\alpha \mid \gamma$	**0,01**	**0,15**	**0,30**	**0,45**	**0,60**	**0,75**	**0,90**	**0,99**
0,01	42,467	50,932	44,986	46,690	56,857	52,679	67,355	47,416
0,15	55,555	58,649	64,623	66,161	67,284	71,678	71,008	48,921
0,30	63,673	80,522	86,026	82,824	96,549	83,139	77,811	94,139
0,45	65,296	72,190	76,014	76,518	86,776	82,684	82,541	86,386
0,60	85,534	74,141	74,894	87,589	86,957	88,787	88,513	113,791
0,75	77,471	85,316	89,685	74,088	86,598	90,479	91,642	98,492
0,90	96,323	88,379	85,658	97,054	83,405	89,373	99,012	107,781
0,99	74,698	87,470	108,361	89,412	73,964	90,427	111,015	125,212

Pode-se observar pela Tabela 2 que o melhor desempenho foi alcançado pela combinação $\alpha = 0{,}01$ e $\gamma = 0{,}01$, com uma média igual à 42,467 movimentos. Já a combinação $\alpha = 0{,}99$ e $\gamma = 0{,}99$, obteve o maior valor médio de passos necessários para o manipulador chegar ao fim da trajetória (125,122).

É possível observar também a variação de desempenho alterando apenas o valor de um dos parâmetros. Por exemplo, tomando a primeira linha de dados da Tabela 2, referente aos valores médios das combinações com $\alpha = 0{,}01$, os resultados variam entre 42,467 e 67,355 movimentos. Nesse aspecto, a Figura 2 apresenta o gráfico de aprendizado (passos $\times$ episódios) para duas combinações: (i) $\alpha = 0{,}01$ e $\gamma = 0{,}01$ e (ii) $\alpha = 0{,}01$ e $\gamma = 0{,}90$.

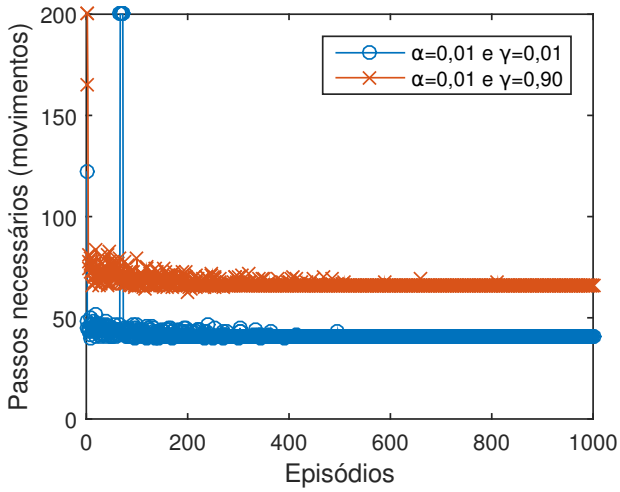

Figura 2 – Evolução do número de passos necessários (movimentos) para o manipulador alcançar o objetivo (fim da trajetória) de acordo com as combinações de parâmetros: (i) $\alpha = 0{,}01$ e $\gamma = 0{,}01$ e (ii) $\alpha = 0{,}01$ e $\gamma = 0{,}90$.

A Figura 2 mostra que a combinação $\alpha = 0{,}01$ e $\gamma = 0{,}01$ convergiu para um valor de passos igual à 41, número menor do que o alcançado pelos parâmetros $\alpha = 0{,}01$ e $\gamma = 0{,}90$ igual à 61 movimentos, ou seja, apenas variando o valor de γ.

Em seguida, os resultados serão avaliados adotando a modelagem matemática via Metodologia de Superfície de Resposta (OTTONI; NEPOMUCENO; OLIVEIRA, 2016).

4.2 Modelagem via superfície de resposta

A Metodologia de Superfície de Resposta (RSM) é uma ferramenta estatística voltada para a análise de problemas de otimização (MYERS; MONTGOMERY; Anderson-Cook, 2009). Nesse sentido, Ottoni, Nepomuceno e Oliveira (2016) apresentam uma modelagem ma-

temática via RSM para a análise dos efeitos da definição dos parâmetros α e γ no desempenho do AR. A estrutura do modelo proposta por Ottoni, Nepomuceno e Oliveira (2016) é apresentada na Equação (5):

$$\hat{y} = \beta_0 + \beta_1\alpha + \beta_2\gamma + \beta_3\alpha^2 + \beta_4\gamma^2 + \beta_5\alpha\gamma, \tag{5}$$

em que, $\hat{y}$ é a variável resposta do modelo ajustado (medida de desempenho); β_0, β_1, β_2, β_3, β_4 e β_5 são os coeficientes; α e γ são as variáveis independentes, referentes à taxa de aprendizado e ao fator de desconto, respectivamente.

Assim, neste trabalho foi ajustado um modelo RSM para auxiliar na análise do desempenho do manipulador SCARA de acordo com a seleção dos parâmetros α e γ. Para o ajuste do modelo foram adotados os dados apresentados na Tabela 2 e o pacote RSM do *software* estatístico R (LENTH, 2009; R Core Team, 2013). A Equação (5) apresenta o modelo ajustado:

$$\hat{y} = 48{,}067 + 79{,}453\alpha + 4{,}087\gamma - 45{,}970\alpha^2 + 5{,}872\gamma^2 + 13{,}313\alpha\gamma. \tag{6}$$

Para avaliar a adequação do modelo aos dados foram analisadas algumas medidas de ajuste. A primeira análise visa verificar se os resíduos do modelo seguem uma distribuição normal. Adotando o teste de Kolmogorov-Smirmov (KS), a hipótese inicial (H_0) é que os resíduos seguem uma distribuição normal ($p_{KS} > 0{,}05$), e a hipótese alternativa (H_1) que não seguem ($p_{KS} < 0{,}05$) (LOPES, 2011). Adotando o teste KS foi confirmada H_0 com $p_{KS} = 0{,}3243$. Também foi confirmada a significância do modelo com $p - valor = 1{,}041 \times 10^{-15}$. Nesse teste, a hipótese inicial é que o modelo não é significante ($p - valor > 0{,}05$) e significante se aceita a hipótese alternativa ($p - valor < 0{,}05$) (MYERS; MONTGOMERY; Anderson-Cook, 2009). Já os valores dos coeficientes de determinação múltipla (R^2) e coeficiente de determinação múltipla ajustada (R^2_a) foram 0,7384 e 0,7159, respectivamente. Esses coeficientes são ajustados entre 0 e 1, e quanto mais próximo de 1, evidencia um bom modelo (MYERS; MONTGOMERY; Anderson-Cook, 2009).

Em seguida, foi realizada uma análise dos resultados via duas ferramentas gráficas da RSM: gráfico de contornos e superfície de resposta. O gráfico de contornos apresenta em duas dimensões (2D) a relação entre as duas variáveis independentes (α e γ) e a variável resposta ($\hat{y}$). Esse gráfico é semelhante a um mapa topográfico. Dessa forma, a partir das linhas de contornos é possível identificar regiões de mínimo/máximo da resposta ajustada. O gráfico de contornos para o modelo ajustado é apresentado na Figura 3.

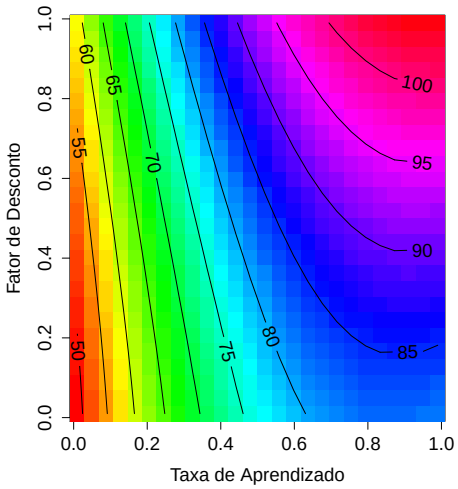

Figura 3 – Gráfico de contornos (2D) para modelo RSM ajustado.

Já o gráfico de superfície de resposta apresenta em três dimensões (3D) a relação entre α, γ e $\hat{y}$. A superfície de resposta para o modelo ajustado é apresentada na Figura 4.

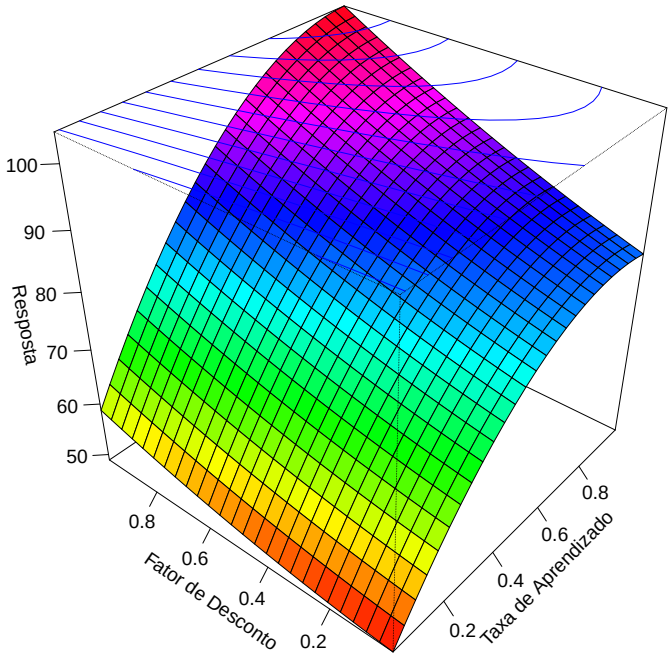

Figura 4 – Superfície de reposta (3D) para modelo RSM ajustado.

Analisando os gráficos 2D e 3D é possível avaliar como os parâmetros α e γ podem influenciar no desempenho do manipulador simulado. Por exemplo, as regiões em vermelho nos gráficos indicam que resposta ajustada está mais próxima de ser minimizada. Nesse aspecto, é possível observar uma tendência de minimizar a resposta (número de passos para alcançar o objetivo) quando $\alpha \to 0$ e $\gamma \to 0$. Por outro lado, quando $\alpha \to 1$ e $\gamma \to 1$, a resposta ajustada tende a valores mais altos (região rosa), distanciando do mínimo de $\hat{y}$.

4.3 Resultados com parâmetros de outros trabalhos

Nesta seção, são apresentados os resultados obtidos nas simulações com a adoção de parâmetros de outros trabalhos que também aplicaram o AR em um manipulador robótico (THAM; PRAGER, 1993; PARK; KIM; SONG, 2007; KIM et al., 2010). Foram realizados experimentos adotando os seguintes parâmetros:

- $\alpha = 0{,}50$ e $\gamma = 0{,}95$ (THAM; PRAGER, 1993),
- $\alpha = 0{,}50$ e $\gamma = 0{,}50$ (PARK; KIM; SONG, 2007),
- $\alpha = 0{,}05$ e $\gamma = 0{,}99$ (KIM et al., 2010).

Essas combinações também foram simuladas em cinco épocas (repetições) de 1000 episódios. Os resultados dos valores médios de movimentos em cada repetição são apresentados na Tabela 3. Além disso, para fins de comparação também são apresentados os resultados para a combinação $\alpha = 0{,}01$ e $\gamma = 0{,}01$ na Tabela 3.

Tabela 3 – Média de movimentos por época (repetição) para parâmetros adotados em outros trabalhos (THAM; PRAGER, 1993; PARK; KIM; SONG, 2007; KIM et al., 2010) e para a combinação $\alpha = 0{,}01$ e $\gamma = 0{,}01$.

Parâmetros		**Média por repetição**				
α	γ	**1**	**2**	**3**	**4**	**5**
0,50	0,95	109,0	102,3	117,8	103,8	118,8
0,50	0,50	90,7	84,7	78,1	87,2	146,8
0,05	0,99	44,8	52,3	75,6	114,9	135,8
0,01	0,01	46,1	44,0	85,4	58,6	42,5

A Tabela 3 revela que a menor média de movimentos (42,5) foi alcançada pela combinação $\alpha = 0{,}01$ e $\gamma = 0{,}01$. Nesse sentido, é importante destacar a capacidade do modelo RSM ajustado indicar regiões com bons parâmetros. Por exemplo, o ponto definido pelos parâmetros utilizados por Tham e Prager (1993) está na região de contorno azul escuro da Figura 3, longe

da área vermelha, na qual é possível encontrar o ponto ($\alpha = 0{,}01$ e $\gamma = 0{,}01$). Nesse mesmo aspecto, o ponto definido pelos parâmetros de Park, Kim e Song (2007) ($\alpha = 0{,}50$ e $\gamma = 0{,}50$) está na área azul claro da Figura 3. Por outro lado, o segundo melhor desempenho na Tabela 3 foi obtido pelos parâmetros de Kim et al. (2010) ($\alpha = 0{,}05$ e $\gamma = 0{,}99$). Os parâmetros de Kim et al. (2010) estão na região amarela da Figura 3, bem mais próxima da área com tendência de minimizar a reposta do modelo RSM (região vermelha).

5 CONCLUSÃO

Neste trabalho, o objetivo foi analisar os efeitos da definição dos parâmetros do AR no desempenho de aprendizado de um manipulador robótico simulado. Para o manipulador SCARA simulado, os resultados apontam que existe uma tendência de aumento de desempenho quando a taxa de aprendizado e o fator de desconto tendem simultaneamente a valores próximos de zero ($\alpha \to 0$ e $\gamma \to 0$). Nesse aspecto, o modelo RSM ajustado fortalece essa hipótese.

Assim, a principal contribuição deste trabalho é a proposta de uma metodologia para a definição de parâmetros do AR para o domínio dos manipuladores robóticos, tendo em vista que, a literatura carece de uma estratégia para a estimação de parâmetros do AR em aplicações de braços robóticos, conforme evidenciando nos trabalhos de Tham e Prager (1993), Park, Kim e Song (2007), Kim et al. (2010). Dessa forma, como destacado por Ottoni, Nepomuceno e Oliveira (2016), os gráficos de contornos e superfície de resposta permitem visualizar regiões de parâmetros que minimizam/maximizam a resposta ajustada. Como por exemplo, é possível observar que os parâmetros definidos nos trabalhos de Tham e Prager (1993) ($\alpha = 0{,}50$ e $\gamma = 0{,}95$) e Park, Kim e Song (2007) ($\alpha = 0{,}50$ e $\gamma = 0{,}50$), estão na região azul dos gráficos (Figuras 3 e 4), longe área vermelha (melhores combinações de acordo com o modelo RSM proposto).

Em trabalhos futuros, sugere-se analisar os efeitos da definição de parâmetros do AR sobre outros tipos de manipuladores robóticos (simulados e reais) e outros algoritmos tradicionais, como o Q-learning. Além disso, analisar experimentos mais complexos, por exemplo, com o objetivo aleatório durante o treinamento e com obstáculos na trajetória.

Agradecimentos: Agradecemos à CAPES, CNPq, FAPEMIG, UFSJ e CEFET-MG.

ANALYSIS OF THE DEFITION OF REINFORCEMENT LEARNING PARAMETERS IN THE ROBOTIC MANIPULATOR PERFORMANCE

ABSTRACT

The objective of this work is to analyze the effects of the definition of Reinforcement Learning parameters on the performance of a robotic manipulator. For this experiments are performed with a simulated manipulator of the type SCARA. Furthermore it adopted the mathematical modeling via Response Surface Methodology. The results point to a tendency in the manipulator performance increase when the learning rate and discount factor are simultaneously approaching zero.

Keywords: Reinforcement Learning. Robotic Manipulator. Response Surface Methodology.

REFERÊNCIAS

BECK, C.; SRIKANT, R. Error bounds for constant step-size Q-learning. **Systems and Control Letters**, v. 61, n. 12, p. 1203–1208, 2012.

DENAVIT, J.; HARTENBERG, R. S. A kinematic notation for lower pair mechanisms based on matrices. **Journal of Applied Mechanics**, v. 77, n. 2, p. 215–221, 1955.

EVEN-DAR, E.; MANSOUR, Y. Learning Rates for Q-learning. **Journal of Machine Learning Research**, v. 5, p. 1–25, 2003.

HERNANDEZ, J. A. M. Software tools for reinforcement learning, artificial neural networks and robotics (matlab and python). Acesso em: 03 ago. 2017. Disponível em: <https://jamh-web.appspot.com/download.htm>.

HERNANDEZ, J. A. M.; LOPE, J. A distributed reinforcement learning control architecture for multi-link robots: Experimental validation. **ICINCO 2007 - International Conference on Informatics in Control, Automation and Robotics**, p. 192–197, 2007.

KIM, B. et al. Impedance learning for robotic contact tasks using natural actor-critic algorithm. **IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)**, v. 40, n. 2, p. 433–443, April 2010. ISSN 1083-4419.

LENTH, R. V. Response-Surface Methods in R, using rsm. **Journal of Statistical Software**, v. 32, n. 7, p. 1–17, 2009.

LIN, C.-K. H∞ reinforcement learning control of robot manipulators using fuzzy wavelet networks. **Fuzzy Sets and Systems**, v. 160, n. 12, p. 1765 – 1786, 2009. ISSN 0165-0114.

LOPES, R. H. C. Kolmogorov-smirnov test. In: ____. **International encyclopedia of statistical science**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011. p. 718–720. ISBN 978-3-642-04898-2.

MILJKOVIć, Z. et al. Neural network reinforcement learning for visual control of robot manipulators. **Expert Systems with Applications**, v. 40, n. 5, p. 1721 – 1736, 2013. ISSN 0957-4174.

MURAKOSHI, K.; MIZUNO, J. A parameter control method in reinforcement learning to rapidly follow unexpected environmental changes. **Biosystems**, v. 77, n. 1-3, p. 109 – 117, 2004. ISSN 0303-2647.

MYERS, R. H.; MONTGOMERY, D. C.; Anderson-Cook, C. M. **Response surface methodology: process and product optimization using designed experiments**. [S.l.]: John Wiley & Sons, 3 ed, 2009.

OTTONI, A. L. C.; NEPOMUCENO, E. G.; OLIVEIRA, M. S. Análise de sensibilidade dos parâmetros do aprendizado por reforço na solução do problema do caixeiro viajante: modelagem via superfície de resposta. In: **CONGRESSO BRASILEIRO DE AUTOMÁTICA, 21., 2016, Vitória.ES. Anais... Vitória, ES: SBA**. [S.l.: s.n.], 2016. p. 513–518.

OTTONI, A. L. C. et al. Análise da influência da taxa de aprendizado e do fator de desconto sobre o desempenho dos algoritmos Q-learning e SARSA: aplicação do aprendizado por reforço na navegação autônoma. **Revista Brasileira de Computação Aplicada**, v. 8, n. 2, p. 44–59, 2016.

PARK, J.-J.; KIM, J.-H.; SONG, J.-B. Path planning for a robot manipulator based on probabilistic roadmap and reinforcement learning. **International Journal of Control Automation and Systems**, Korean Institute of Electrical Engineers, v. 5, n. 6, p. 674–680, 2007.

PELLEGRINI, J.; WAINER, J. Processos de Decisão de Markov: um tutorial. **RITA - Revista de Informática Teórica e Aplicada**, v. 14, n. 2, p. 133–179, 2007.

R Core Team. **R: A language and environment for statistical computing**. Vienna, Austria, 2013.

SCHWEIGHOFER, N.; DOYA, K. Meta-learning in reinforcement learning. **Neural Networks**, v. 16, n. 1, p. 5–9, 2003.

SUTTON, R.; BARTO, A. **Reinforcement learning: an introduction**. 1st. ed. [S.l.]: Cambridge, MA: MIT Press, 1998.

TANG, L.; LIU, Y.-J.; TONG, S. Adaptive neural control using reinforcement learning for a class of robot manipulator. **Neural Computing and Applications**, Springer, v. 25, n. 1, p. 135–141, 2014.

THAM, C. K.; PRAGER, R. W. Reinforcement learning methods for multi-linked manipulator obstacle avoidance and control. In: **ASIA-PACIFIC WORKSHOP ON ADVANCES IN MOTION CONTROL, 1993. Proceedings...** [S.l.: s.n.], 1993. p. 140–145.

VISIOLI, A.; LEGNANI, G. On the trajectory tracking control of industrial scara robot manipulators. **IEEE Transactions on Industrial Electronics**, v. 49, n. 1, p. 224–232, Feb 2002. ISSN 0278-0046.

WATKINS, C. J.; DAYAN, P. Technical note Q-learning. **Machine Learning**, v. 8, n. 3, p. 279–292, 1992.

Recebido em: 04/07/2017

Aprovado em: 07/08/2017

Publicado em: 13/11/2017