Apoena Revista Eletrônica

USO DA TÉCNICA DE *REINFORCEMENT LEARNING* PARA CONTROLE DE NÍVEL DE TANQUE

Yuri Matheus Sant'Anna de Oliveira[1]
Iury Silva[2]
David Alan Silva dos Santos[3]
Antônia Ferreira dos Santos Cruz[4]
Tiago de Oliveira Silva[5]
Gilson Amorim Carvalho[6]

RESUMO

Metodologias baseadas em dados são cada vez mais utilizadas para a resolução das mais diversas atividades. Nesse trabalho, a técnica *reinforcement learning* (RL) é aplicada no escopo de controle de processos. Esse mecanismo consiste em treinar um agente, a exemplode um programa, código ou algoritmo, a interagir em um ambiente, por meio de ações, para atingir um objetivo. Nesse sentido, visando sua aplicação, realizou-se um projeto de um controlador baseado na técnica de aprendizado por reforço no ambiente de simulação do software MATLAB, e, adicionalmente foi feita a comparação com um controlador *Proporcional e Integral – PI*. Por meio destas simulações foi possível visualizar as vantagense desvantagens da técnica *reinforcement learning* em relação ao controle *PI* para controle de um tanque linear

Palavras-chave: *Reinforcement Learning*. Controle PI. Matlab

1 INTRODUÇÃO

As pessoas aprendem interagindo com o ambiente desde a infância, o desenvolvimento de determinados movimentos como caminhar ou levantar os braços, sozinhos. E essas atividades são realizadas sem a presença de alguém para ensinar a realizar essa tarefa. Nesse sentido, há a execução de uma interação com o ambiente em que o indivíduo está, mediante

[1] *Bacharelando em Engenharia Elétrica, Centro Universitário Jorge Amado, E-mail: yuri3.oliveira@gmail.com*

[2] *Bacharelando em Engenharia Elétrica, Centro Universitário Jorge Amado, E-mail: iury.158@hotmail.com*
[3] *Bacharelando em Engenharia Elétrica, Centro Universitário Jorge Amado, E-mail: david.alan.s@hotmail.com*
[4] *Orientadora Professora Mestra em Regulação da Indústria de Energia, Professora Titular do Centro Universitário J orge Amado, E-mail:acruz1107@unijorge.pro.br*
[5] *Co-orientador Doutorando em Mecatrônica. Mestre em Mecatrônica Tiago de Oliveira Silva, E-mail: tiagoo.ts@gmail.com*

[6] *Co-orientador,Professor Titular do Centro Universitário Jorge Amado, E-mail:Gilson. Carvalho@unijorge.pro.br*

suas necessidades, com o objetivo de se alcançar determinado resultado.

Com o surgimento da Quarta Revolução Industrial nos últimos anos, definiu-se uma tendência para automação das fábricas, gerando um mundo em que os sistemas de fabricação virtuais e físicos cooperam entre si de uma maneira flexível a nível global. Diante desse cenário, a obtenção de dados mais precisos e confiáveis se tornam cada vez mais acessíveis na implementação dos processos industriais, o que torna essas metodologias baseadas em dados cada vez mais aplicáveis.

Nesse sentido, algumas dessas metodologias utilizadas, como as áreas de *Data Science* e *Reinforcement Learning*, faz-se requeridas, apresentando resultados bastante satisfatórios. No entanto, mesmo sendo um tema bastante explorado nas mais diversas áreas, como em Marketing, Finanças, Recursos Humanos e todos os setores possíveis onde há geração de dados, *Reinforcement Learning* ainda é um conteúdo pouco utilizado no controle de processos.

A abordagem de *Machine Learning* (ML), no meio tecnológico, possibilitou adaptação ao ambiente dinâmico em diversas utilidades, variando entre robôs flexíveis e tarefas automatizadas (ENGELBERGER, 1989). O *Reinforcement Learning* (RL), uma técnica encontrada na *Machine Learning*, aplicável na obtenção de resultados a partir daaprendizagem de tentativa e erro, em que o agente não é informado sobre a ação que deverá ser tomada, mas que deverá realizar testes e descobrir qual a melhor ação que deverá ser tomada para a obtenção de uma maior recompensa. É importante ressaltar que nem sempre a recompensa recebida é imediata. No entanto, a depender da ação, isso influenciará as recompensas futuras. (SUTTON et al., 2014).

1. FUNDAMENTAÇÃO TEÓRICA

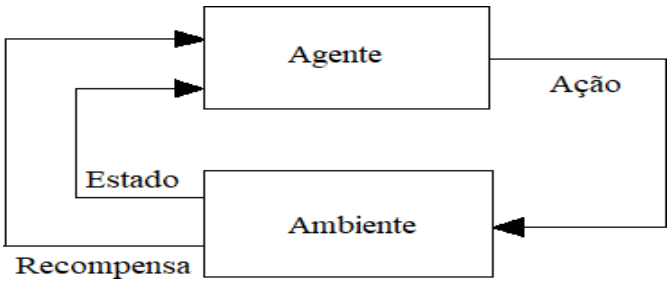
1.1 *Reinforcement Learning* (RL)

Aprendizado por reforço ou *Reinforcement Learning* é um ensaio de modelos de aprendizado de máquina para tomar uma sequência de decisões, onde o agente deve tomar medidas para maximizar as recompensas cumulativas, aprender a assimilar uma meta em um ambiente indeterminado e potencialmente complexo. Ao alavancar o poder da pesquisa e de muitas tentativas, o aprendizado por reforço é atualmente a maneira mais eficaz de sugerir a criatividade da máquina.

O objetivo do *Reinforcement Learning* (RL) é aprender uma estratégia para o agente a partir de tentativas empíricas e um retorno simples relativo recebido. Com a estratégia ideal, o

agente é capaz de se adaptar agilmente ao ambiente para maximizar recompensas iminentes.

Figura 1: Esquema metodológico do *Reinforcement Learning*.



Fonte: Adaptado de Sutton et al, 1998.

Conforme pode-se observar na figura 1, o agente comporta-se em um ambiente, como o meio reage a certas ações é definido por um molde que podemos ou não conhecer. O operador pode ficar em um de vários estados no ambiente e escolher suceder uma das muitas práticas para mudar de um estado para outro. A condição em que o agente chegará é resultado das probabilidades de transição entre os estados. Dessa maneira assim que uma atividade é realizada, o ambiente oferece uma retribuição como *feedback*.

1.2 Controlador PI

O controlador PI é constituído por duas partes em sua ação de controle, uma proporcional ao erro e outra proporcional à integral do erro. Em determinados sistemas que requerem um controle mais preciso, habitualmente é aplicado um controlador proporcional,no qual, o sinal do controle é proporcional ao sinal de entrada, ou seja, o retorno será de pequeno valor, se o sinal de entrada for pequeno e saída será grande se a entrada for de valor grande, sendo basicamente um amplificador.

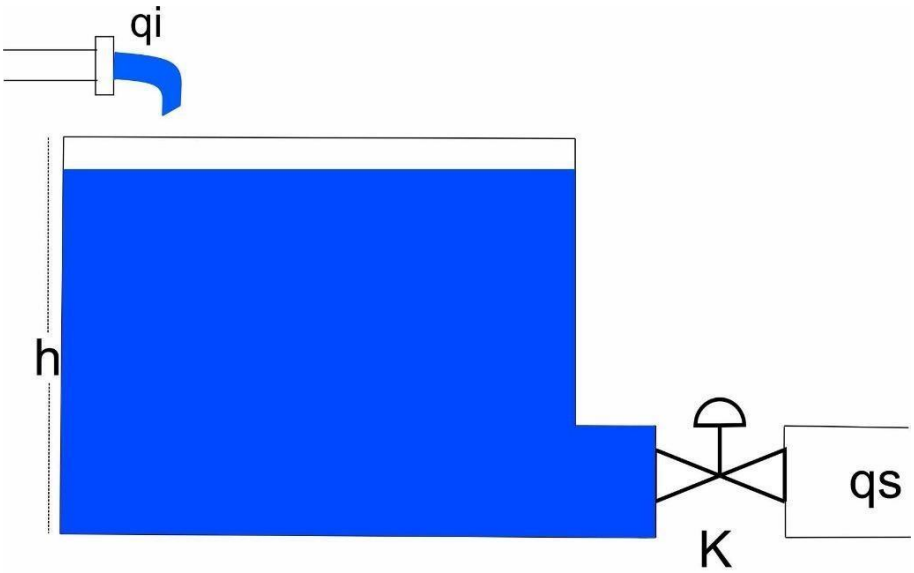
A elementar função da ação integral é fazer com que meios sigam, com erro nulo, um sinal de orientação do tipo degrau. Contudo, a ação integral se empregue isoladamente tende a piorar a estabilidade relativa do sistema. Para igualar este fato, a ação integral é em geral utilizada em conjunto com a ação proporcional estabelecendo o controlador PI, desta forma o controlador é dado pela Equação 1:

$$a(t) = Kp\, e(t) + ki \int e(t)\, d \qquad \text{[Eq. 1]}$$

2. Resultados

Neste capítulo, apresentamos os resultados das simulações realizadas com *reinforcement learning* para controle de nível de um tanque linear. O tanque projetado possui 8 metros de altura (h) e um registro (K) na saída, conforme Figura 2. Inicialmente, a vazão de entrada (qi) foi definida como 1 m³/min, como a área do tanque é constante, definiu-se a constante C = 1,0, o registro K = 0,5 $m^{2.5}/min$.

Figura 2: Tanque simulado no projeto



Fonte: Adaptado de Ogata (2011)

2.1 Equações

Seguindo a lei da conservação de massa, é possível descrever a variação volumétrica do tanque através da Equação 2:

$$\Delta V = qi - qs \qquad \text{[Eq. 2]}$$

A variação citada na Equação 2 significa que se a vazão que entra (qi) é igual a vazão que sai (qs), não há alteração no nível (h). Em caso de uma vazão na entrada maior que a saída, o nível aumenta, porém, se elevar apenas a vazão de saída, o nível irá diminuir.

A vazão de saída depende, além do nível, da posição do registro (K). A vazão de saída é proporcional, levando em conta a constante de proporção K, a diferença de pressão sobre o registro.

$$qs = K\sqrt{\Delta P} \quad \text{[Eq. 3]}$$

Levando em consideração que a saída do tanque está aberta, obtém-se uma pressão atmosférica após o registro. No tanque, a pressão é realizada pelo nível com acréscimo da atmosfera, entendendo que a equação 3 utiliza a diferença de pressão e a pressão atmosférica está nos dois pontos do registro, é possível cancelar na equação e reescrevê-la da seguinte maneira:

$$qs = K^* \sqrt{h(t)} \quad \text{[Eq. 4]}$$

No caso do tanque com uma área constante (C), o volume é dado por:

$$C^*\Delta h(t) = qi - qs \quad \text{[Eq. 5]}$$

Para um modelo simulatório, tem-se a seguinte equação final do nível do tanque:

$$\Delta h(t) = \frac{qi - K*\sqrt{h(t)}}{C} \quad \text{[Eq. 6]}$$

3.3 Software de simulação

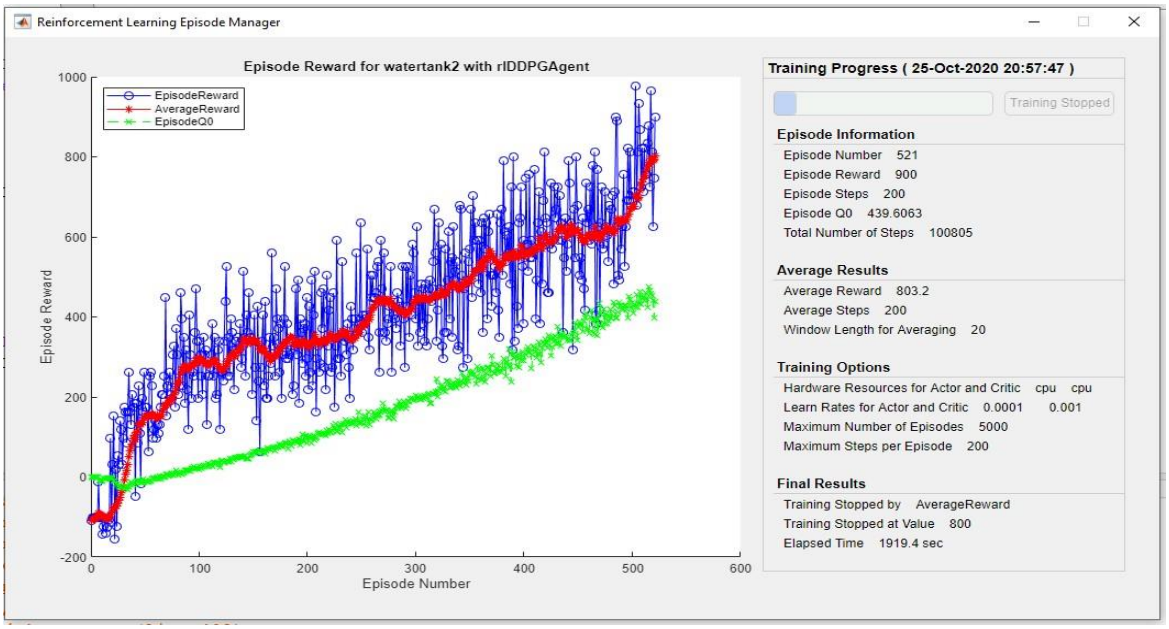
O *matlab* é um software de programação por linguagem de blocos por meio do*simulink*, que trabalha com uma linguagem de códigos derivadas do C ++, com análise numérica, cálculo de matrizes e processamento de sinais, permitindo produzir estruturas e funções simples e complexas. Esta ferramenta possibilita implementar e testar soluções com uma maior facilidade e precisão. A simulação das atribuições é integrada no *simulink*, dispositivo do *matlab* que utiliza uma área de interação gráfica ante a forma de diagrama de blocos, permitindo modelar, simular e analisar sistemas, possibilitando, projetar e ensaiar inúmeros sistemas.

3.5 Controle do tanque utilizando Reinforcement Learning

A técnica de controle *reinforcement learning* utiliza o treinamento para o aprendizado do agente. O agente necessita da prática para que aprenda com seus erros e assume as melhores decisões para o controle do sistema, que ocorre quando uma recompensa é enviada durante o treinamento para que o agente entenda qual estado deva operar, e uma punição, compreendendo que não pode seguir o estado atual.

A análise de cada episódio do treinamento é baseada na observação inicial do ambiente, no qual computa a ação inicial. Com a ação inicial, é determinada a próxima observação que gera uma nova ação para o ambiente, se a condição de pré-determinada para afinalização do treinamento do agente for atendida, entende-se que o objetivo foi alcançado. Em situação oposta, o algoritmo continuará analisando os episódios subsequentes.

Figura 3: Treinamento do agente



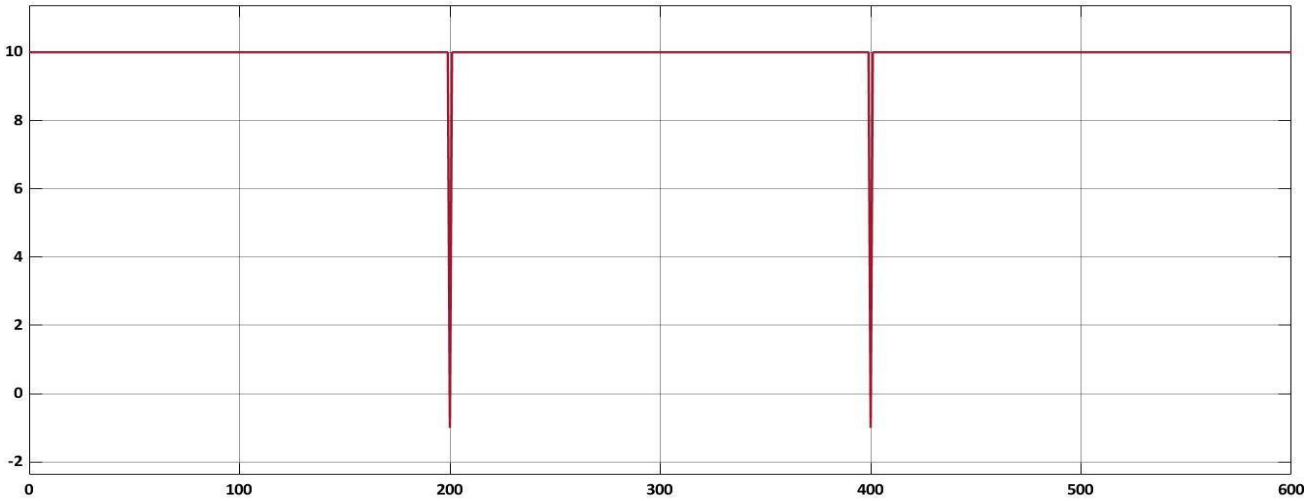
Fonte: Elaboração própria

Conforme a Figura 3, a cada episódio, o agente recebe um valor de recompensa ou punição. A condição de encerramento do treinamento do agente foi uma janela de 20 episódios com recompensa média de, no mínimo, r(t) = 800, por episódio. O limite de episódios determinado foi 5000, porém, com o pré-requisito estipulado, a finalização ocorreu no episódio 521. A média de recompensa (*Average Reward*) foi 803.2, e a recompensa futura (*Episode* Q0) 439.6063 na conclusão.

Na opção do treinamento, especificou-se a taxa de aprendizagem em 0.0001 para o *actor* e 0.001 para o *critic*. O *critic* realiza atualizações e acumula para a *Value Function*, o *actor* executa função similar ao *critic*, mas ao invés do *Value Function*, opera na *policy*. Quanto menor a taxa de aprendizagem, mais preciso será o resultado.

A partir do momento em que o agente está treinado e o mesmo aprendeu os parâmetros para o controle da planta, pode-se definir os níveis desejados para o tanque. Inicialmente o sistema funcionou com três níveis de referência, h = (4 m; 4,5 m; 4 m) com t = (0 min; 200 min; 400 min) respectivamente. Para que o controle atinja os objetivos de referência, utilizou-se a recompensa imediata, quando o nível do tanque alcança ou aproxima- se do sinal de entrada, um valor R=10 é enviado ao agente para que possa compreender que está no ponto ideal ou equivalente ao solicitado. Em contraponto, quando a condição está distante da pretendida, o agente recebe uma punição, R = -1, conforme Figura 4.

Figura 4: Gráfico de recompensa

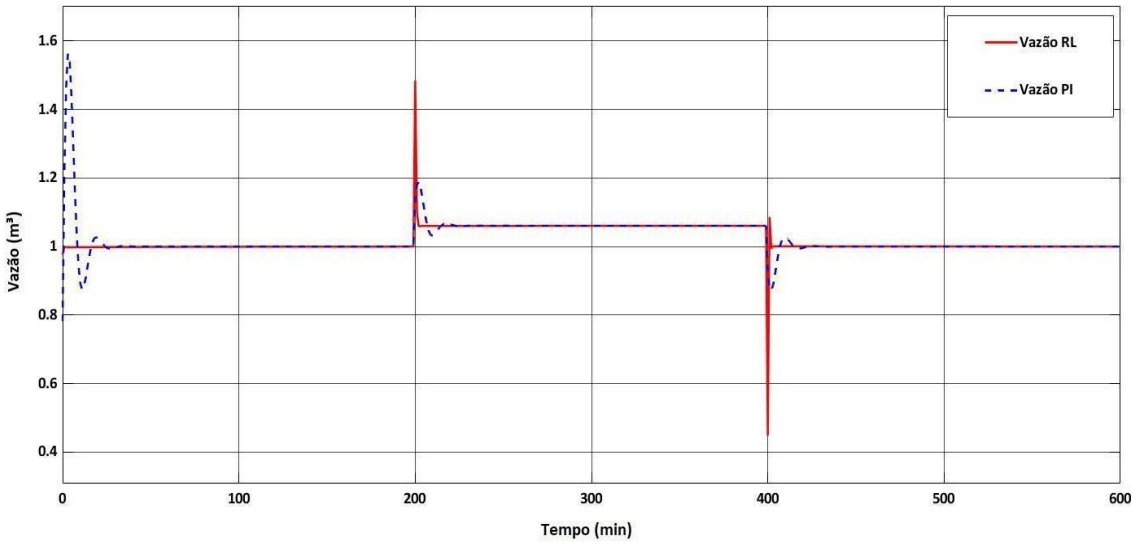


Fonte: Elaboração própria

A partir da Figura 4, percebe-se que durante as mudanças de níveis, t= 200 min e t = 400 min, o agente recebe punições, pois, é o período em que se exige mais do controle, devido aos distúrbios causados pela vazão, e o sistema distancia-se da referência. As recompensas são encaminhadas no momento em que existe uma aproximação do sinal de controle com a referência.

Para os devidos valores de referência no projeto, o agente determinou, de acordo com a Figura 5, vazões de entrada no tanque, com o propósito de adequar o volume no tanque com o objetivo definido.

Figura 5: Vazão volumétrica de entrada



Fonte: Elaboração própria

De acordo com a Figura 5, quando projetado um degrau de acréscimo no sistema, a vazão de entrada liberada no tanque aumenta o suficiente com intenção de aproximar-se da melhor maneira possível ao nível. Em comparação entre os modelos de controle *reinforcement learning* e controle PI, consegue-se constatar as diferenças como os sistemas secomportam. Durante a simulação com a técnica de aprendizagem por reforço, a vazão de entrada inicia com qi = 0,976 m³/min e vazão máxima qi_m = 1 m³/min que representa a vazão esperada na simulação, enquanto o método Proporcional e Integral a vazão inicial foi qi = 0,782 e alcançou qi_m= 1,54 m³/min com estabilização da vazão após t = 24 min, porém, durante as mudanças de níveis do tanque, a vazão tem picos distintos dos iniciais para os modelos. No tempo t = 200 min, o controle RL tem pico de = 1,48 m³/min e estabilização depois de t = 2 min, em contrapartida, o controle PI atingiu qi_m = 1,19 m³/min e controlou a vazão logo após t = 20 min.

Ao encontrar uma redução de nível do tanque, as vazões de entrada reagiram de maneira semelhante às anteriores, porém, inversamente. O método *reinforcement learning* atingiu vazão de entrada mínima qi_{min} = 0,448 m³/min e retornou ao ponto estável em t = 2 min, enquanto o controle Proporcional e Integral reduziu a vazão à qi_{min}= 0,873, estabilizando após t = 14 min.

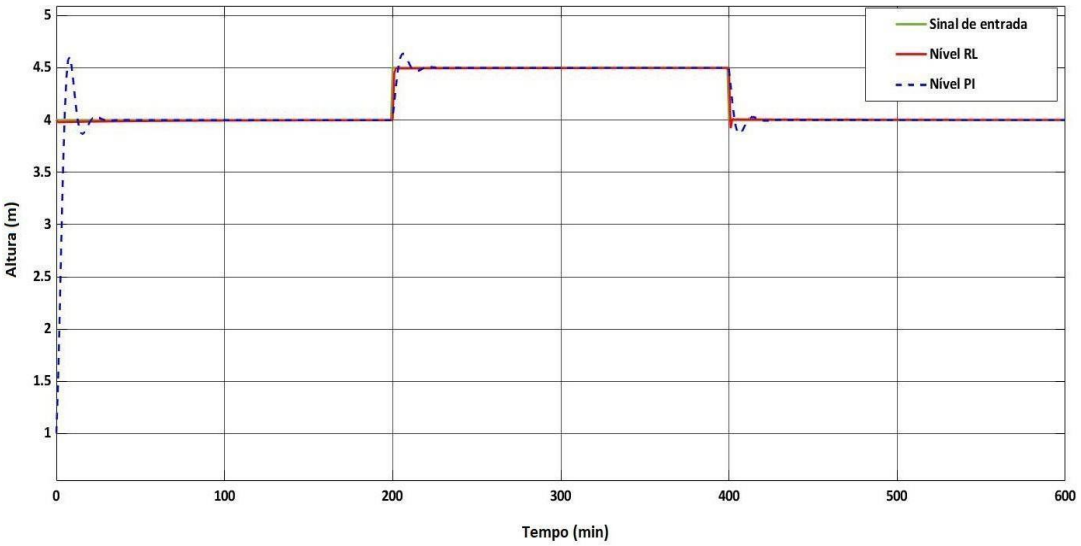
A partir da Figura 8 compreende-se que quando o nível estabiliza, a vazão tende a reagir de maneira semelhante, para o nível do tanque aumentar, precisa-se de um maior volume na

vazão de entrada e quando se necessita de um nível menor no tanque, a vazão de entrada reduz de forma que encontre a altura desejada e logo após, estabiliza-se.

O controle RL fornece um pico na vazão maior que o controle PI, porém, o *reinforcement learning* tem um período de resposta mais rápido que o Proporcional e Integral para estabilizar a vazão de entrada no tanque.

O controlador através do aprendizado por reforço teve um comportamento mais eficaz que o controlador Proporcional-Integral durante a estabilização do nível do tanque. Como pode-se observar na Figura 6, o PI apresenta uma fase para normalização da altura do fluido no tanque muito maior que o controlador com *reinforcement learning.*

Figura 6: Gráfico de nível do tanque



Fonte: Elaboração própria

Com a Figura 6, compreende-se a diferença do erro de estado estacionário (e_{ss}) entre os dois modelos de controle e o seu tempo de estabilização do nível no tanque. É perceptível o contraste entre os controles logo no primeiro momento, com t = 0 min, o erro estacionário do modelo RL foi nulo, enquanto o PI mostrou e_{ss} = 3 m com período de 29 minutos para adequação ao nível esperado. No instante da primeira mudança de nível (h = 4,5 m), t = 200 min, a curva do *reinforcement learning* apresentou uma imprecisão com e_{ss} = 0,501 m ao mesmo tempo que o Proporcional e Integral teve e_{ss} = 0,478 m, porém, enquanto o aprendizado por reforço torna-se estável com 2 minutos, o PI demora 18 minutos. Quando foi solicitado que a altura do fluido no tanque retornasse para h = 4 m, o RL amostrou um erro (e_{ss}) com - 0,499 m, o controle PI demonstrou erro de - 0,478 m, neste ponto, *o reinforcement learning* exibiu um controle mais preciso com estabilização em 1 minuto, ao passo que o

controle Proporcional e Integral durou 11 minutos para restabelecer ao nível desejado.

O erro de estado estacionário inicial do controlador PI é um demonstrativo do comportamento do sistema com uma alteração abrupta de nível, enquanto o RL originou-se com erro nulo, contudo, os dois controles não seguem uma trajetória uniforme e apresentam diferenças mínimas com o sinal de referência, com máximo de $e_{ss} = 10^{-3}$ m.

3. CONSIDERAÇÕES FINAIS

O objetivo desse trabalho deu-se em explorar, através de simulação com software Matlab/Simulink, uma aplicação com o modelo de controle *reinforcement learning* em um tanque linear. Os resultados foram comparados com o controle *Proporcional e Integral*, aplicado no mesmo modelo de tanque, afim de demonstração de vantagens e desvantagens dos métodos.

Os resultados do trabalho demonstraram que o método *RL* foi eficaz e vantajoso quando contrastado com o controlador *PI* para utilização no modelo do tanque especificado. Apesar do erro no nível não obterem uma grande diferença durante a maior parte do tempo, a técnica de aprendizagem por reforço mostrou-se um tempo de resposta mais eficiente para estabilização do nível.

O *reinforcement learning* indica ser uma alternativa interessante para o controle de processos, entendendo que a técnica, diferente do controle *Proporcional e Integral*, não necessita de modelagem e atualização da planta por parte dos operadores para execução. O agente do *reinforcement learning* continua com a aprendizagem durante o processo, o que o torna adaptável às mudanças de dados e parâmetros do sistema.

REFERÊNCIAS

BIKMUKHAMETOV, Timur. JASCHKE, Johannes. First Principles and Machine Learning Virtual Flow Metering: A Literature Review. Journal of Petroleum Science and Engineering. Department of Chemical Engineering, Norwegian University of Science and Technology, 7034, Sem Sælandsvei 4, Trondheim, Norway. Publicado em 2020.

DOGRUER, Tufan. TAN, Nusret. Design of PI Controller using Optimization Method in Fractional Order Control Systems. Science Direct. Graziosmanpasa University, Department of Electronic and Automation, 60250, Tokat, TURKEY. Inonu University, Department of Electrical and Electronics, 44280, Malatya, TURKEY. Publicado em 1998.

FRANÇA, Mário. Curso de introdução ao MATLAB. Faculdade de Ciência e Tecnologia da Universidade de Coimbra. Publicado em fevereiro de 2011.

IDE, Hidenori. KURITA, Takio. Improvement of learning for CNN with ReLU activation by sparse regularization.. Hiroshima University. Publicado em maio de 2017.

JING, Hong. Reinforcement Learning – The Value Function. Disponível em: https://jinglescode.github.io/2019/06/30/reinforcement-learning-value-function/. Acesso em: 28 de setembro de 2020.

KAELBLING, Pack Leslie. LITTMAN, Michael L. MOORE, Andrew W. Reinforcement Learning: A Survey. Journal of Artificial Intelligence Research 4. Computer Science Department, Box 1910, Brown University Provinde, RI 02912-1910 USA. Smith Hall 221, Carnegie Mellon University, 5000 Forbes Avenue Pittsburgh, PA 15213 USA. Publicado em maio de 1996.

KIEFER, Nicholas M. A Value Function Arising in the Economics of Information. Journal of Economic Dynamics and Control. Publicado em abril de 1989.

LAU, Thomas. LI, Haoqian. Reinforcement Learning: Predction, Control and Value Function Aproximation. Columbia University. Point Zero One Technology. Publicado em 29 de agosto de 2019.

MINOARIVELO, Henintsoa Onivola. Application of Markov Decision Processes to the Control of a Traffic Intersection. University of Barcelona, Spain. Publicado em 22 de maio de2009.

SCHWAB, Klauss. A Quarta Revolução Industrial. Instituto Federal de Tecnlogia de Zurique (ETH Zurich). Publicado em 2016.

SHIPMAN, William J. COETZEE, Loutjie C. Reinforcement Learning and Deep Neural Networks for PI Controller Tuning. Science Direct. MINTEK, Johannesburg, South Africa.Publicado em 2019.

SULTAN, Hossam H. SALEM, Nancy M. Multi-Classification of Brain Tumor Images UsingDeep Neural Network.. Helwan University. Publicado em maio de 2019.

SUTTON, Richard S. BARTO, Andrew G. Reinforcement Learning: An Introduction. TheMIT Press Cambridge, Massachusetts London, England. Publicado em 2014.

WATKINS, Hellaby Cornish John Christopher. Learning from Delayed Rewards. King'sCollege. Publicado em maio de 1996. Acesso em: 28 de setembro de 2020.