TRABALHO DE GRADUAÇÃO

CONTROLE ADAPTATIVO DE PROCESSO DE NÍVEL UTILIZANDO APRENDIZADO POR REFORÇO ATOR-CRÍTICO

Daniel Vicentin Gonçalves

Brasília, Julho de 2016

ENGENHARIA
MECATRÔNICA
UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASILIA
Faculdade de Tecnologia
Curso de Graduação em Engenharia de Controle e Automação

TRABALHO DE GRADUAÇÃO

CONTROLE ADAPTATIVO DE PROCESSO DE NÍVEL UTILIZANDO APRENDIZADO POR REFORÇO ATOR-CRÍTICO

Daniel Vicentin Gonçalves

Relatório submetido como requisito parcial de obtenção de grau de Engenheiro de Controle e Automação

Banca Examinadora

Prof. Adolfo Bauchspiess, ENE/UnB _____
Orientador

Prof. Eduardo Stockler Tognetti, ENE/UnB _____
Co-orientador

Prof. Alexandre Ricardo Soares Romariz, ENE/UnB _____
Examinador interno

Brasília, Julho de 2016

FICHA CATALOGRÁFICA

GONÇALVES, DANIEL V.
CONTROLE ADAPTATIVO DE PROCESSO DE NÍVEL UTILIZANDO APRENDIZADO POR REFORÇO ATOR-CRÍTICO

[Distrito Federal] 2016.

x, 61p., 297 mm (FT/UnB, Engenheiro, Controle e Automação, 2016). Trabalho de Graduação – Universidade de Brasília.Faculdade de Tecnologia.

1. Controle Adaptativo
2. Aprendizado por Reforço
3. Ator-Crítico
4. Diferença Tempora
5. Processo de Quatro Tanques

I. Mecatrônica/FT/UnB

REFERÊNCIA BIBLIOGRÁFICA

GONÇALVES, DANIEL V. Controle Adaptativo De Processo De Nível Utilizando Aprendizado Por Reforço Ator-crítico. Trabalho de Graduação em Engenharia de Controle e Automação, Publicação FT.TG-$n°$000, Faculdade de Tecnologia, Universidade de Brasília, Brasília, DF, 61p.

CESSÃO DE DIREITOS

AUTOR: Daniel Vicentin Gonçalves

TÍTULO DO TRABALHO DE GRADUAÇÃO: Controle Adaptativo De Processo De Nível Utilizando Aprendizado Por Reforço Ator-crítico.

GRAU: Engenheiro ANO: 2016

Controle Adaptativo, Aprendizado por Reforço, Ator-Crítico, Diferença Temporal, Processo de Quatro Tanques

Dedicatória

um beijo pra minha mãe, pro meu pai, e pra você

Daniel Vicentin Gonçalves

Agradecimentos

Agradecimentos!

Daniel Vicentin Gonçalves

RESUMO

O presente trabalho apresenta o projeto de um controlador adaptativo utilizando aprendizado por reforço para controle de nível em um processo de 4 tanques. O controlador utiliza a abordagem Ator-Crítico com aproximação de funções por redes neurais de base radial e treinamento através do gradiente descendente do erro de diferença temporal. Os resultados simulados demonstram um desempenho superior do controlador adaptativo quando comparado a um controlador PI tradicional, porém não foi possível a aplicação no sistema real devido a distúrbios técnicos na planta piloto, tornando as medições realizadas imprecisas. São também apresentadas as dificuldades dessa implementação.

Palavras Chave: Controle Adaptativo, Aprendizado por Reforço, Ator-Crítico, Diferença Temporal, Processo de Quatro Tanques

ABSTRACT

This work presents an adaptive controller project using reinforcement learning for liquid-level control in a quadruple-tank process. The controller was designed with the Actor-Critic method using redial basis networks for function approximation and training through gradient descent of the temporal difference error. The simulation results show superior performance of the adaptive controller when compared to traditional PI controller, but it was not possible to apply it in a real environment because of thecnical disturbances in the pilot plant that made the measurements imprecise. The other challenges found are also presented.

Keywords: Adaptive Control, Reinforcement Learning, Actor-Critic, Temporal Difference, Quadruple Tank Process

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SÍMBOLOS

Símbolos Latinos

Símbolos Gregos

Grupos Adimensionais

Subscritos

Sobrescritos

$\dot{}$	Variação temporal

Siglas

RL	*Reinforcement Learning*
MDP	Processo de decisão de Markov - *Markov Decision Process*

Capítulo 1

Introdução

1.1 Contextualização

Controle de nível de líquidos com a interação de múltiplos tanques é um problema comum em processos industriais [3]. Esses processos apresentam comportamento não linear. Além disso, características muitas vezes desconsideradas durante a modelagem, como saturação, variação de parâmetros ao longo do tempo, atrasos, zonas mortas ou perturbações diversas dificultam o controle desses processos. Um sistema com múltiplos tanques foi proposto por Johansson com o objetivo didático de demonstrar conceitos e propriedades de um sistema de múltiplas entradas e múltiplas saídas, MIMO (do inglês, multiple-input and multiple-output), em especial sistemas de fase não mínima. [4]

Técnicas clássicas de controle podem ser aplicadas a sistemas não lineares a partir da linearização em torno de um ponto de operação, uma vez que pequenas variações em torno desse ponto resultam em uma resposta aproximadamente linear. Porém, para maiores variações, a resposta poderá divergir consideravelmente do esperado no projeto, [5].

Para melhorar a resposta do sistema em todo o domínio de operação existem técnicas que lidam diretamente com modelos não lineares. A maioria dessas técnicas buscam a otimização da resposta em relação a algum critério. Técnicas de controle ótimo possuem uma maior complexidade matemática e exigem uma maior precisão no modelo utilizado.

Na falta de um modelo preciso do sistema, é possível recorrer a técnicas de controle adaptativo, em que o controlador é capaz de se ajustar ao sistema.

Em aprendizado de máquinas, o Aprendizado por Reforço, ou Reinforcement Learning (RL), é um método de resolver problemas de otimização através de um agente interagindo com o meio e adaptando suas ações ou políticas de controle a partir de estímulos recebidos. Na literatura é possível encontrar exemplos em que o controle de líquidos é realizado através de aprendizado por reforço. [6]

1.2 Contextualização do problema

As não-linearidades de um sistema podem existir devido à própria dinâmica do sistema ou de características de atuadores e sensores, como saturação da entrada e dinâmica de resposta do próprio atuador ou sensores.

Sistemas não-lineares demandam uma maior complexidade durante a modelagem e projeto de controladores. Com isso em mente, a maioria dos controladores convencionais considera um modelo linearizado do sistema em um dado ponto de operação. Entretanto, essa abordagem pode comprometer o desempenho do sistema em pontos distantes ao ponto especificado no projeto.

De modo a definir um controlador largamente independente do ponto de operação, metodologias de controle não linear vêm sendo cada vez mais utilizadas, sendo algumas dessas técnicas baseadas em princípios biológicos.

Com esse intuito, foi considerado nesse trabalho a implementação de um controlador a partir de aprendizado por reforço, método de aprendizado intimamente ligado ao controle adaptativo e ao controle ótimo.

Uma vez que o aprendizado por reforço aparenta fornecer resultados promissores na área de controle, o trabalho busca analisar e implementar essas técnicas de aprendizado em um sistema de controle de processos real.

1.3 Objetivos do projeto

O trabalho consiste na implementação e avaliação de um controlador utilizando aprendizado por reforço. A técnica utilizada é a partir da metodologia ator-crítico, a partir do erro de diferença temporal e aproximação de funções por redes neurais de base radial.

O controlador será implementado em uma planta de controle de nível com quatro tanques, considerando o comportamento de duas bombas.

1.4 Resultados obtidos

Os resultados apresentados mostram a veracidade do algoritmo através de simulações do processo de nível. As simulações indicam que o controlador por aprendizado permite o controle do sistema com um tempo de resposta e um sobrepasso menor do que o controlador PI convencional utilizado.

A aplicação no sistema real não foi possível devido a limitações técnicas na planta. São apresentados esses aspectos juntamente com as demais dificuldades apresentadas durante a implementação, como inconsistências nas medidas devido a ruídos e a presença de limitações de atuação.

1.5 Apresentação do manuscrito

Esse trabalho apresentará nos próximos capítulos as ideias e conceitos envolvidos no Aprendizado por Reforço, assim como toda a definição matemática utilizada. Em seguida será apresentado o sistema a ser controlado, com a descrição da Planta piloto e a modelagem do sistema para construção do ambiente de simulação.

No desenvolvimento será apresentada a metodologia do projeto do controlador, assim como os ambientes de simulação e experimento e os desafios encontrados na implementação.

Serão então apresentados os resultados em simulação e em seguida os resultados da implementação no sistema real.

Capítulo 2

Descrição da Planta Piloto

Para a implementação do controlador projetado nesse trabalho, foi utilizada a planta didática com configuração de quatro tanques, DCNV4T, da fabricante DIDATICONTROL [7].

A planta é contruída a partir do processo de tanques quádruplos apresentado por Johansson e criada com o objetivo didático de propor a implementação de controles multivariáveis avançados.

A Figura 2.1 mostra a planta real do processo.

O processo é composto por quatro tanques de acrílico, um reservatório, sensores de nível, válvulas manuais, válvulas motorizadas e duas bombas acionadas por inversores de frequência. A interface de controle e monitoração do processo é composta por um painel elétrico, um controlador lógico programável e um computador.

Figura 2.1: Planta DCNV4T.

A planta visa propiciar o desenvolvimento de conhecimentos na área de modelagem e projeto de sistemas de controle de processos, lidando com múltiplas variáveis e características e condições similares à realidade encontrada na indústria.

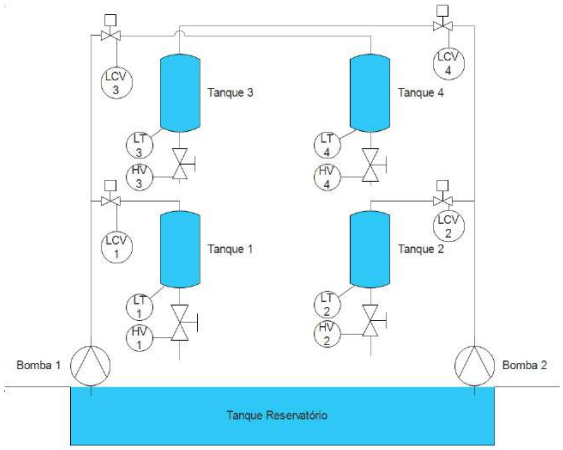

Figura 2.2: Configuração dos tanques

Uma importante característica desse sistema é a não linearidade do processo, além das não linearidades intrínsecas a qualquer sistema real, como saturação de atuadores, tempos de resposta, curva de atuação, entre outros.

Nas seguintes seções serão apresentadas as principais características relacionadas à planta utilizada.

2.1 Tanques

A planta possui um total de cinco tanques, sendo quatro utilizados para o processo em si e o quinto utilizado como reservatório para armazenamento do líquido utilizado.

O reservatório fica localizado na parte inferior da bancada e possui líquido suficiente para o funcionamento do processo sem a necessidade de reabastecimento durante um ensaio.

Os tanques utilizados no processo são compostos por cilindros de acrílico com altura de aproximadamente 90 centímetros e diâmetro de aproximadamente 20 centímetros.

Por questões de segurança e para evitar vazamentos do tanque, por passar da altura máxima, a instrumentação existente possui um acionamento de desarme ao atingir a altura de 68 centímetros, logo essa altura foi definida como altura máxima do tanque.

Os tanques foram enumerados e serão referenciados por essas identificações. A numeração é apresentada na Figura 2.2

2.2 Bombas

O processo utiliza duas bombas centrífugas monoestágio SR-5-25PT da série MARK, fabricadas pela GRUNDFOS. As bombas são acopladas a dois motores elétricos de modelo W48J, fabricados pela WEG. O motor e a bomba são acionados por um inversor de frequências.

Modelo	W48J
Potência	0.5CV
Tensão de Alimentação	220/380V
Corrente	1.8/1.04A
Frequência	60Hz
Rotações	3500RPM

Tabela 2.1: Características nominais do Motor Elétrico WEG

Modelo	SR-5-25PT
Potência	0.5CV
Conexões	1" e 3/4"
Vazão	$8.5m^3/h$
Altura Manométrica	6 mc.a
Rotações	3500RPM

Tabela 2.2: Características nominais da bomba GRUNDFOS

Como vantagens apontadas pelo fabricante, essas bombas proporcionam um menor consumo de energia, uma construção compacta em monobloco, bocal de sucção e recalque centrados e baixo custo de manutenção.

As bombas são conectadas aos tanques através de tubulações 3/4", feitas de PVC. Assim a bomba 1 bombeia o fluido para os tanques 1 e 4, enquanto a bomba 2 bombeia o fluido para os tanques 2 e 3.

As tabelas 2.1 e 2.2 apresentam as características nominais do motor elétrico e da bomba.

Foram levantadas as curvas de cada bomba, a partir de medições nos transmissores de vazão para cada tanque. As curvas apresentadas na Figura 2.4 mostram que as bombas possuem uma diferença considerável de potência, onde a bomba 1 possui aproximadamente 60% da potência da bomba 2.

Foi também observado que a bomba não apresenta vazão antes de atingir aproximadamente 20% da potência total, o que caracteriza uma zona morta do atuador.

O inversor de frequências utilizado possui como máximo uma frequência nominal de 60Hz, porém a empresa configurou o inversor para atingir um máximo de 50Hz na saída.

Figura 2.3: Bomba e motor elétrico

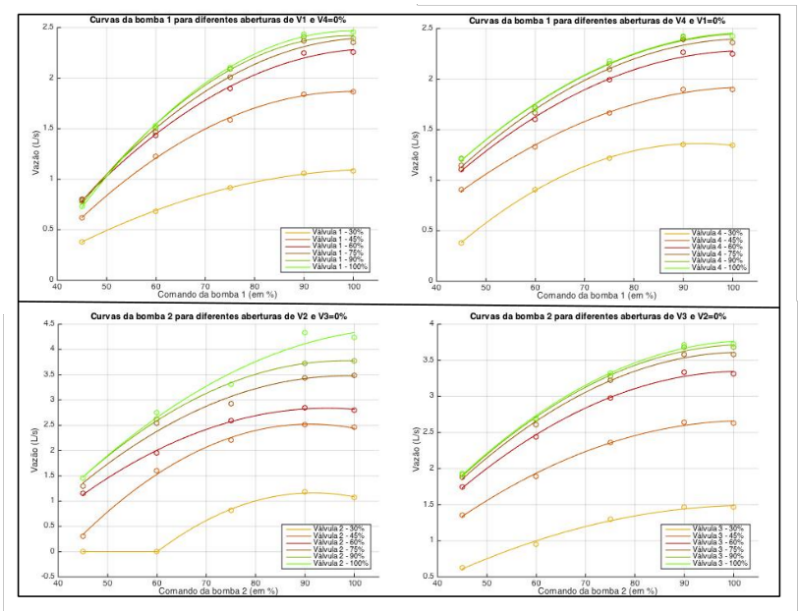

Figura 2.4: Curvas das bombas

Sendo assim, além da não linearidade característica da bomba, o atuador possui uma zona morta de 20% e uma saturação em aproximadamente 90% da potência máxima.

Essas características das bombas, juntamente com o tempo de resposta do conjunto de atuação faz com que sejam acrescentadas outras não-linearidades ao sistema e um desafio adicional na implementação do controlador.

2.3 Válvulas

O sistema possui diversas válvulas para alterar sua configuração e para alterar taxas de vazão para uma dada configuração.

O conjunto de válvulas é composto tanto por válvulas manuais como por válvulas motorizadas. As válvulas manuais são mantidas em uma configuração padrão, sendo alteradas apenas para algumas medições específicas.

As válvulas motorizadas utilizam um motor elétrico para acionar um mecanismo de avanço por meio de um trem de engrenagens para abrí-las ou fechá-las. Essa válvula permite um controle sobre a abertura, sendo possível realizar uma alteração com precisão e definir diferentes configurações para o sistema. As características nominais da válvula utilizada são mostradas na tabela 2.3.

A variação de abertura da válvula na estratégia de controle não seria possível, pois o tempo total de abertura é de 65 segundos, sendo muito lenta para utilização no controle em malha fechada.

Modelo	SR13G220032B1-E
Tipo	Duas Vias
Tensão de Alimentação	24VAC
Potência	3.5VA
Frequência	50/60Hz
Fator de Fluxo	3.2
Máximo Diferencial de Pressão Fechada	0.3MPa
Máxima Pressão Estática	2.1MPa
Tempo de abertura	65s

Tabela 2.3: Características nominais das válvulas motorizadas

Modelo	SP21
Tensão de alimentação	12 a 30Vdc
Consumo	22mA
Saída	4 a 20mA
Proteção	Inversão de polaridade
Tipo de Sensor	Sensor piezoresistivo em Aço Inox 316

Tabela 2.4: Características nominais dos sensores

2.4 Sensores de nível

Os sensores utilizados são da marca Sitron e modelo SP21 para monitoração de pressão em líquidos e gases. É um sensor piezoresistivo com uma dinâmica linear, apresentando uma tensão de saída proporcional à pressão exercida no sensor.

A medida de altura é realizada a partir da medição de pressão. Foi realizada uma calibração para converter a variável medida para a variável de trabalho, porém essa calibração foi realizada por fora desse trabalho.

Na planta piloto são usados quatro sensores de nível, uma na base de cada tanque. As características nominais do sensor são apresentados na tabela 2.4.

A reposta dos sensores é bastante influenciada ao se ligar o inversor de frequências que move o motor e a bomba, gerando uma queda na medição e ruídos de medição.

Capítulo 3

Modelagem do Sistema

De forma a se obter um modelo do sistema para simulações e projeto de um controlador por técnicas convencionais, o primeiro passo é a modelagem do sistema. A forma de modelagem utilizada será a modelagem fenomenológica, ou seja, a partir das equações diferenciais que descrevem a dinâmica do sistema.

3.1 Modelagem Fenomenológica

A modelagem fenomenológica do sistema será baseada no trabalho realizado por Johansson[4] e será feita a partir dos princípios de conservação de massa e da equação de Bernoulli para líquidos incompressíveis.

Para cada tanque, o balanco de massas resulta em:

$$\frac{dV}{dt} = A\frac{dh}{dt} = q_{in} - q_{out} \tag{3.1}$$

sendo:

V = volume de água no tanque

A = área da secção transversal do tanque

h = altura da coluna de líquido

q_{in} = fluxo de entrada

q_{out} = fluxo de saída

Pela equacão de Bernoulli para líquidos incompressíveis:

$$\frac{\rho v_w^2}{2} + \rho g h + P = \text{constante} \tag{3.2}$$

sendo:

v_w = velocidade de escoamento da água

g = aceleração da gravidade

P = Pressão

ρ = massa específica da água

Assumindo que na superfície da água a velocidade seja nula ($v_w = 0$) e que na parte inferior de cada tanque a altura da coluna de água é igual a zero ($h = 0$):

$$(v_w = 0) : \rho g h + P = \text{constante}$$

$$(h = 0) : \frac{\rho v^2}{2} + P = \text{constante}$$

Igualando as duas equações, a velocidade da água no fluxo de saída é dada por:

$$\rho g h + P = \frac{\rho v_w^2}{2} + P$$

$$v_w = \sqrt{2gh} \tag{3.3}$$

Como o fluxo de saída do tanque ($q_{out,i}$) é o produto entre o fluxo de saída pela secção transversal do orifício (o):

$$q_{out,i} = o_i v_{w,i} = o_i \sqrt{2gh_i} \tag{3.4}$$

Considerando que as bombas gerem um fluxo linearmente proporcional à tensão aplicada na bomba (u), o fluxo da bomba é dado por:

$$q_{bomba_1} = k_i u_i$$

Considerando γ_i a proporção de abertura das válvulas entre os tanques para cada bomba, o fluxo para os tanques será dado como:

$$q_{bomba_1} = q_{bomba_1 1} + q_{bomba_1 3} = \gamma_1 k_1 u_1 + (1 - \gamma_1) k_1 u_1$$

$$q_{bomba_2} = q_{bomba_2 2} + q_{bomba_2 4} = \gamma_2 k_2 u_2 + (1 - \gamma_2) k_2 u_2$$

A partir de (1) e (4):

$$\frac{dh_i}{dt} = \frac{1}{a}(q_{in,i} - q_{out,i})$$

Logo, para o tanque 1:

$$q_{in,1} = q_{out,3} + q_{bomba1,1} = o_3\sqrt{2gh_3} + \gamma_1 k_1 v_1$$

$$q_{out,1} = o_1\sqrt{2gh_1}$$

Para o tanque 2:

$$q_{in,2} = q_{out,4} + q_{bomba2,2} = o_4\sqrt{2gh_4} + \gamma_2 k_2 v_2$$

$$q_{out,2} = o_2\sqrt{2gh_2}$$

Para o tanque 3:

$$q_{in,3} = q_{bomba2,3} = (1-\gamma_2)k_2 v_2$$

$$q_{out,3} = o_3\sqrt{2gh_3}$$

Para o tanque 4:

$$q_{in,4} = q_{bomba1,4} = (1-\gamma_1)k_1 v_1$$

$$q_{out,4} = o_4\sqrt{2gh_4}$$

Dessa forma, as equações que definem a dinâmica do sistema pode ser dada por:

$$\begin{cases} \frac{dh_1}{dt} = \frac{1}{A_1}(q_{in,1} - q_{out,1}) & = \frac{1}{A_1}(o_3\sqrt{2gh_3} + \gamma_1 k_1 v_1 - o_1\sqrt{2gh_1}) \\ \frac{dh_2}{dt} = \frac{1}{A_2}(q_{in,2} - q_{out,2}) & = \frac{1}{A_2}(o_4\sqrt{2gh_4} + \gamma_2 k_2 v_2 - o_2\sqrt{2gh_2}) \\ \frac{dh_3}{dt} = \frac{1}{A_3}(q_{in,3} - q_{out,3}) & = \frac{1}{A_3}((1-\gamma_2)k_2 v_2 - o_3\sqrt{2gh_3}) \\ \frac{dh_4}{dt} = \frac{1}{A_4}(q_{in,4} - q_{out,4}) & = \frac{1}{A_4}((1-\gamma_1)k_1 v_1 - o_4\sqrt{2gh_4}) \end{cases} \tag{3.5}$$

A partir dessas equações foi construído o ambiente de simulações.

Apesar da modelagem utilizar as propriedades físicas que descrevem a dinâmica do sistema, o sistema real possui um comportamento bastante diferente. Isso ocorre pois diferentes alturas na coluna de líquido, há uma diferença no regime de escoamento, além de coeficientes adicionais que podem afetar um sistema real.

Capítulo 4

Aprendizado por Reforço

Em ambientes dinâmicos, um comportamento ótimo é difícil de ser obtido a partir de um único instante, uma vez que um comportamento favorável à busca de um objetivo pode deixar de sê-lo logo em seguida.

A partir da ideia de aprendizado foi então definido o conceito de aprendizado por reforço no aprendizado de máquinas, onde um agente aprende seu comportamento a partir de experiência adquirida.

O conceito de aprendizado por reforço tem suas raízes na psicologia, porém ganhou amplo espaço na área de inteligência artificial e aprendizado de máquinas, sendo usado para resolver problemas gerais que envolvem escolha de uma política ótima para tomada de decisões, [8, 9, 10].

4.1 Conceito

O aprendizado por reforço difere do aprendizado supervisionado, que é amplamente estudado na área de aprendizado de máquinas, onde o aprendizado é realizado a partir de exemplos fornecidos por um supervisor, como o caso de reconhecimento de padrões e a utilização de redes neurais artificiais.

O aprendizado por reforço visa então aprender a partir de experiências próprias, através de interações com o ambiente. O agente do aprendizado executa ações e avalia o resultado das ações tomadas, atualizando seu comportamento a cada tentativa. Os resultados de cada ação são avaliados a partir de respostas do ambiente, como as mudanças de estados e recompensas obtidas pelo agente.

O aprendizado parte do pressuposto de que uma ação que produz resultados satisfatórios ou positivos deve ser reforçada, enquanto ações com resultados indesejados devem ser reprimidas.

A necessidade de interação com o ambiente leva a um dos grandes desafios dessa forma de aprendizado, o dilema entre exploração e aproveitamento (em inglês, *exploration vs exploitation*). O agente deve escolher as ações que são avaliadas por produzirem uma maior recompensa, mas para encontrar essas ações é necessário explorar outras opções não escolhidas anteriormente. É

necessário buscar o aproveitamento de ações já descobertas, porém também é necessário explorar novas ações para tomar melhores decisões.

Outra característica importante do aprendizado por reforço é fato de considerar o problema como uma interação com um ambiente desconhecido visando um objetivo, o que difere esse método de outros que consideram subproblemas mais gerais, sem o conhecimento de como essa etapa influenciará no resultado final.

Ao utilizar o aprendizado por reforço, todos os agentes já possuem conhecimento do objetivo e devem lidar com as incertezas do ambiente onde operam.

4.2 MDP

Processos de decisão de Markov, ou MDP (do inglês: Markovian Decision Process) fornecem uma estrutura formal adequada para o aprendizado por reforço.

Um processo decisório de Markov pode ser definido por uma quádrupla $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, onde $\mathcal{S}$ representa o espaço de estados possíveis do processo, $\mathcal{A}$ representa o espaço de ações do controlador, $\mathcal{P}$ indica o espaço de probabilidades de transições de estados, dada uma ação, e $\mathcal{R}$ indica o espaço de recompensas a partir dessas transições, [11].

Dessa forma, a probabilidade de uma mudança de estados é representada por:

$$P(s, a, s') = Pr\{s_{t+1} = s' | s_t = s, a_t = a\} \tag{4.1}$$

Onde $P(s, a, s')$ indica a probabilidade de transição de um estado s para cada estado s' caso seja tomada a ação a.

De maneira semelhante,

$$R(s, a, s') = E\{r_{t+1} | s_t = s, a_t = a, s_{t+1} = s'\} \tag{4.2}$$

Onde $R(s, a, s')$ indica a recompensa esperada r_{t+1}após uma ação a que leva o estado s para cada estado s'.

A Figura 4.1 apresenta um exemplo de MDP, onde as ações são dadas por u, enquanto cada estado é dado por x.

Sendo assim, a propriedade de Markov é definida pelo fato de que as transições e recompensas após uma ação a partir do estado atual depende apenas desse estado, independente dos estados anteriores.

Uma política de açõesπ é definida por uma sequência de ações a ser executada a partir de um estado, ou uma função que mapeia o estado atual à ação que deve ser executada em seguida, definindo uma estratégia de controle. Essa política pode ser dita estocástica quando há a probabilidade de uma ação levar a diferentes estados. Caso cada ação possa levar a apenas um estado a

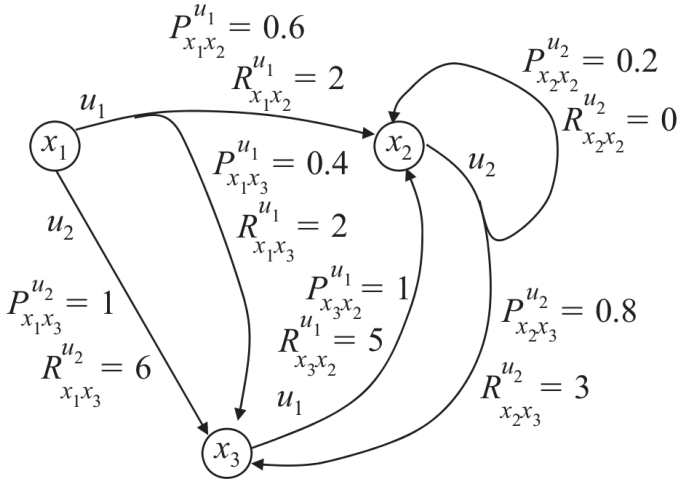

Figura 4.1: Exemplo de MDP com probabilidades e recompensas associadas, [1].

política é dita determinística.

Quando um sistema apresenta a propriedade de Markov é possível então que, a partir de um política de ações π, sejam definidos os estados futuros do sistema e, de forma análoga, partir de um objetivo e definir as ações que levariam o agente a esse estado.

4.3 Recompensas e Função de Valor

Um agente deve ser capaz de perceber o estado s em que se encontra, executar ações para realizar uma mudança de estado e então buscar atingir um objetivo. O aprendizado é importante quando o agente não possui, a princípio, um conhecimento da política ideal a ser adotada para atingir o objetivo.

Em técnicas de aprendizado supervisionado o agente recebe exemplos de comportamento a serem seguidos, utilizando então esses exemplos para atualizar sua política de ações. Porém, no aprendizado por reforço o agente deve aprender através de suas próprias interações com o ambiente.

Dessa forma, é importante a definição do sinal de recompensa utilizado pelo agente. A recompensa r, é definido como um valor numérico disponibilizado ao agente após a execução de uma ação. Essa recompensa é obtida a partir de $R(s, a)$ apresentada anteriormente.

Como exemplo, vamos considerar um jogo de damas. Caso um jogador elimine uma peça do adversário, esse recebe um ponto, +1, porém, caso ele perca uma peça, perderá um ponto, -1. Nesse caso o jogador vai sempre se esforçar para eliminar peças do adversário enquanto evita perder as suas.

Relacionado ao conceito de recompensa está o da função de valor V. A função de valor define a recompensa que se espera acumular a partir do estado atual s ao se adotar uma política de ações π. A função de valor $V^{\pi}(s)$ representa então a análise de desempenho a longo prazo, enquanto a função de recompensa $R(s)$ avalia uma ação a curto prazo.

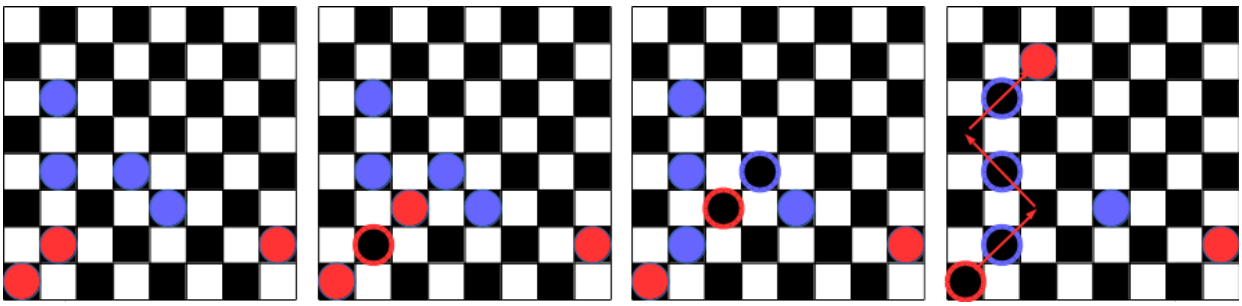

Figura 4.2: Exemplo de diferença entre recompensa e Valor

Muitas vezes, uma política pode levar inicialmente a recompensas ruins, porém levar a um maior valor V futuro. Utilizando novamente o exemplo exibido, um jogador poderia sacrificar uma de suas peças, perdendo um ponto, mas eliminando em seguida peças do adversário, totalizando um saldo positivo.

É perceptível que as decisões do agente devem ser tomadas buscando maximizar o valor obtido, em vez de simplesmente maximizar a recompensa, já que muitas vezes um sacrifício imediato pode levar a maiores ganhos futuros.

A maioria dos algoritmos de aprendizado por reforço é baseado na estimação da função de valor, para determinar quão desejável seria para o agente estar em determinado estado, [12]. A partir da função de valor é então definida a política a ser seguida.

4.4 Diferença Temporal

A função de valor V é a base para seleção da política de ações adotada, porém essa função é desconhecida pelo agente, sendo necessário aprender a estimá-la a partir das recompensas obtidas através das interações com o ambiente.

O método de aprendizado por diferença temporal (TD, *temporal difference*) utiliza as recompensas obtidas ao para estimar o valor dessa política, fazendo com que o agente aprenda a melhorar seu comportamento a cada ação executada, [13].

Ao utilizar experiência por interações, os métodos por diferença temporal removem a necessidade de um modelo explícito do sistema, podendo ser aplicados em sistemas com parâmetros ou dinâmica desconhecidos, [13].

Outra vantagem na utilização desse método está na natureza incremental de sua implementação, sendo utilizados para aprendizado online. A cada passo realizado, as recompensas obtidas são suficiente para executar o próximo passo, diferente de métodos como Monte Carlo, que finalizam um episódio de treinamentos para então obter a estimativa de recompensa, [13].

O erro de diferença temporal pode ser calculado através da Equação:

$$\delta_{TD} = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{4.3}$$

A partir desse erro, pode ser realizado o método mais simples de aprendizado por diferença temporal, conhecido como TD(0) [13],

$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

onde s_{t+1}representa o estado atingido após a execução de uma ação a partindo de um estado s_t, recebendo uma recompensa r_{t+1} correspondente. Ao receber essa nova recompensa a função de valor$V$ do novo estado é utilizada para atualizar a do estado anterior.

O fator α é chamado fator de aprendizado define o passo de aprendizado adotado, definindo a taxa de variação para $V(s_t)$, sendo $0 < \alpha < 1$. Outro parâmetro do aprendizado é o fator de desconto γ, limitado ao intervalo $0 < \gamma < 1$, que define a importância dada às recompensas futuras, sendo que, quanto menor γ, maior a importância das recompensas a curto prazo e menos importante serão as recompensas futuras.

4.5 Traços de Elegibilidade

O método TD(0) é o primeiro passo do aprendizado com diferença temporal, porém esse método leva em conta apenas o Valor a partir do último estado do sistema, não possuindo nenhuma espécie de memória e, consequentemente, não atribuindo o efeito de recompensas a estados passados.

Como em muitos sistemas reais a recompensa pode ser recebida do ambiente com um atraso em relação à ação executada, é necessário possuir conhecimento em relação aos estados que contribuíram para o resultado.

O método de traços de elegibilidade estudado foi o TD(λ), onde lambda representa um fator de atenuação, para definir o atraso do sinal de reforço.

A implementação dos traços pode estar definida de duas formas [11]: traços acumulativos e traços substitutivos.

A formulação matemática dos traços acumulativos é dada por:

$$e_t \leftarrow \begin{cases} \gamma\lambda e_t, & x \neq x_k \\ \gamma\lambda e_t + 1, & x = x_k \end{cases} \tag{4.4}$$

para todos os estados não terminais.

Já no caso de traços substitutivos, o traço de cada estado visitado é diretamente substituído por 1.

$$e_t \leftarrow \begin{cases} \gamma\lambda e_t, & x \neq x_k \\ 1, & x = x_k \end{cases} \tag{4.5}$$

A diferença apresentada por esses dois métodos é apresentado na figura 2.

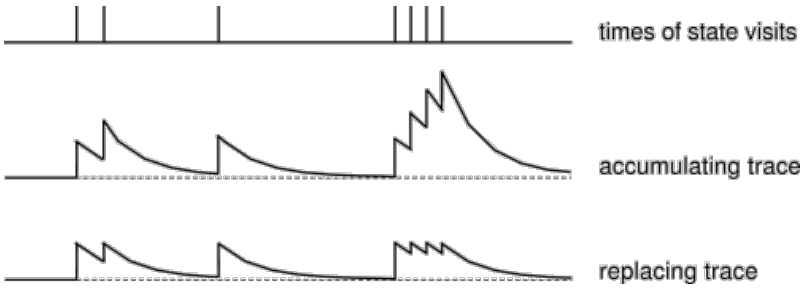

Figura 4.3: Diferença entre traços acumulativos e substitutivos [2].

É importante notar que o método de traços acumulativos possui problemas de convergência, tornando traços substitutivos mais utilizados na prática, [2].

Ao se utilizar traços de elegibilidade, a regra de treinamento TD(λ) passa a ser dada por:

$$V(s_t) \leftarrow V(s_t) + \alpha \delta_{TD} e_t \tag{4.6}$$

4.6 Métodos de Aprendizado

Muitos problemas em diversos campos possuem uma grande dimensionalidade do espaço de estados, não sendo possível tratar cada estado de maneira explícita e fazendo-se necessário o uso de aproximadores de funções para definir as funções de valor e política, [11].

Com o passar do tempo, foram desenvolvidos diversos algoritmos de aprendizado por reforço, que podem ser divididos em três grupos: Método Crítico, Método Ator, e Método Ator-Crítico[14]

Os métodos Críticos, como Q-Learning [15] e Sarsa [16], utilizam uma função de valor espaço-estado sem uma função explícita para representar a política adotada. Dessa forma, esses métodos utilizam geralmente uma discretização do espaço de estados e ações, sendo uma aproximação de todo o espaço de possibilidades.

Métodos Atores, por sua vez, utilizam políticas parametrizadas e não possuem nenhuma forma de memória da função de valor. Esses métodos buscam aprimorar a política através do gradiente dos parâmetros. A vantagem desses métodos é a capacidade de atingir um espaço de ações contínuo, porém esses métodos apresentam uma alta variância de parâmetros, uma vez que os cálculos não levam em conta estimativas anteriores, [17].

4.6.1 Método Ator-Crítico

O método ator-crítico foi introduzido por Witten [18] e então por Barto *et al. [19]*. Essa classe de métodos utiliza duas estruturas distintas para realizar o aprendizado. A primeira estrutura é o Ator, utilizado para definir a política de ações aplicadas e a segunda estrutura é o Crítico, que estima a função de valor e avalia todas as ações executadas pelo ator.

Uma vez que o Ator possui uma política de ações explícita, calcular a ação a ser seguida

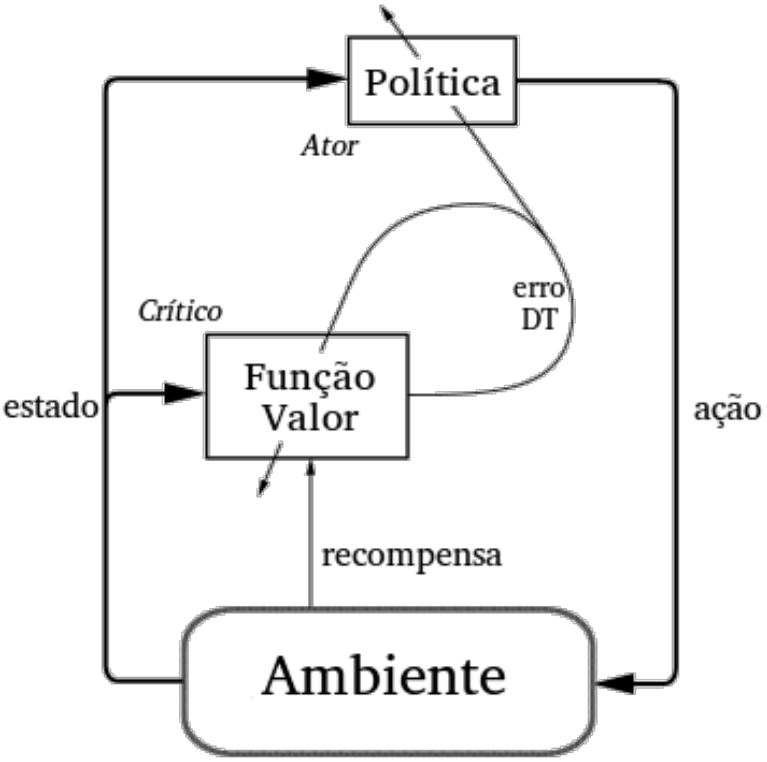

Figura 4.4: Estrutura Ator-Crítico

exige pouco esforço, sendo possível trabalhar com um espaço contínuo de ações. Esses métodos também possuem a vantagem de conseguir aprender políticas estocásticas, obtendo as melhores probabilidades a partir da seleção de diversas ações.

O crítico avalia as ações adotadas através do método de diferença temporal e utiliza o erro DT, da equação 4.3, para aprimorar sua estimação de valor e a política do crítico. Sendo V a função implementada pelo Crítico.

4.7 Aproximação Linear e Gradiente-descendente

A estrutura do Ator e do crítico não pode ser apenas um mapeamento direto dos estados e ações para o valor e a política, uma vez que para um grande espaço as possibilidades de mapeamento podem ser infinitas. Para definir então a estrutura do Ator e do Crítico de forma a abranger todo o espaço de estado e ações, inclusive para o caso contínuo, é necessário o uso de aproximadores de funções.

Dentre os métodos de aproximação possíveis, uma escolha adequada é a de métodos de aproximação linear, uma vez que apresentam uma grande simplicidade de implementação e baixo custo de execução, características desejáveis para o algoritmo de aprendizado.

Os métodos de aproximação linear podem ser caracterizados por um vetor de parâmetros

$$\vec{v_t} = (v_t(1), v(2), ..., v(n))^T$$

e por um vetor de características

$$\vec{\phi_s} = (\phi_s(1), \phi_s(2), ..., \phi_s(n))^T$$

a partir do estado atual do sistema. Esses dois vetores em conjunto definem a influência $\upsilon(i)$ da característica $\phi(i)$ sobre a aproximação da função.

Dessa forma, podemos definir a função de valor por:

$$V(s_t) = \vec{\upsilon_t}\vec{\phi_s} = \sum_{i=1}^{n} \upsilon_t(i)\phi_s(i)$$

As funções de valor definida é linear em relação aos parâmetros adotados.

Na aproximação linear, o aprendizado ocorre através da atualização do vetor de parâmetros, que pode ser realizada através do gradiente-descendente. O método consiste em uma pequena variação dos parâmetros na direção que mais reduzir o erro daquela aproximação.

Dessa forma,

$$\vec{\upsilon_{t+1}} \leftarrow \vec{\upsilon_t} + \alpha \nabla_{\vec{\upsilon_t}} V(s_t) \tag{4.7}$$

representam o aprendizado dos parâmetros pelo gradiente descendente, sendo α a taxa de aprendizado.

Como foi realizada uma aproximação linear, temos que:

$$\nabla_{\vec{\upsilon_t}} V(s_t) = \vec{\phi_s} \tag{4.8}$$

Ao se utilizar o erro de diferença temporal e traços de elegibilidade, juntamente com a aproximação de funções, pelo gradiente descendente o aprendizado se torna:

$$\vec{\upsilon}_{t+1} \leftarrow \vec{\upsilon_t} + \alpha \delta_{DT} \vec{e}_t \tag{4.9}$$

$$\vec{e}_{t+1} \leftarrow \gamma\lambda \vec{e}_t + \nabla_{\vec{\upsilon}} V(s_t) \tag{4.10}$$

4.7.1 Funções de Base Radial

Para definir as características adotadas na aproximação linear de funções, podemos utilizar funções de base radial (FBR), uma vez que essas funções podem ser usadas para aproximar qualquer função contínua, sendo consideradas aproximadores universais, [20].

As funções de base radial são dadas por funções que seu valor depende apenas de sua distância da origem. Para o trabalho, foram utilizadas funções gaussianas, da forma:

$$f(x) = \exp\left(-\frac{\|x - \mu_i\|^2}{2\sigma_i^2}\right) \tag{4.11}$$

Sendoμ a média da gaussiana eσ o desvio padrão da distribuição.

Apesar das funções de base radial serem aproximadores universais, Moody [21] afirma que a estimação por FBR comum foi pobre, exceto para dados de treinamento, enquanto FBR normalizadas cobriam um maior espaço de entrada e proporcionavam uma melhor aproximação.

Sendo assim, uma função de base radial normalizada pode ser dada por:

$$Nf(x) = \frac{f_i(x)}{\sum_{j=1}^{h} f_j(x)} \tag{4.12}$$

Onde $f_i(x)$ representa cada elemento de um vetor de FBR, como na camada oculta de uma rede neural.

Capítulo 5

Desenvolvimento

5.1 Sistema Simulado

De forma a se projetar um controlador, é bastante útil iniciar o projeto por um ambiente de simulação, onde pode ser realizado experimentos de forma mais rápida e se obter uma noção inicial em relação a parâmetros utilizados. Para isso foi utilizado como ambiente de simulações o Simulink®

A partir da modelagem do sistema foi desenvolvido o ambiente de simulações onde foi implementado o algoritmo de controle através de aprendizado por reforço, apresentado na figura 5.1.

Para as simulações iniciais foi considerado o modelo não-linear completo, com blocos saturadores delimitando a altura de cada tanque e os valores do sinal de entrada.

O sistema simulado, apesar de seguir a modelagem fenomenológica do sistema real, possui parâmetros aproximados e não leva em conta outras fontes de não-linearidade, como as curvas da bomba e sua dinâmica de atuação.

5.2 Estrutura do Controlador

Nessa etapa serão descritos todos os componentes utilizados para integrar o controlador. A estrutura geral é apresentada na figura

5.2.1 Variáveis de entrada

O controlador utilizado partiu da abordagem de ajustar os parâmetros de um controlador PID, composto por um canal proporcional, um derivativo e um integral. Para isso foi realizada uma realimentação negativa e utilizado um conversor de estados para tornar as variáveis de erro operáveis. As conversões realizadas foram a obtenção da integral do erro e a derivada do mesmo.

Sendo assim, os estados de entrada para o controlador foram:

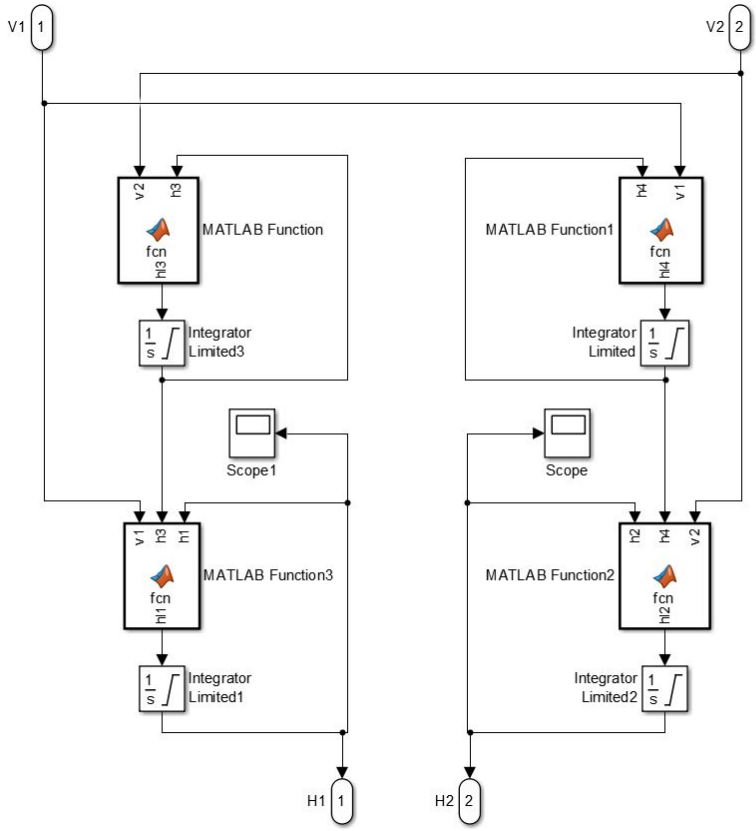

Figura 5.1: Sistema Simulado

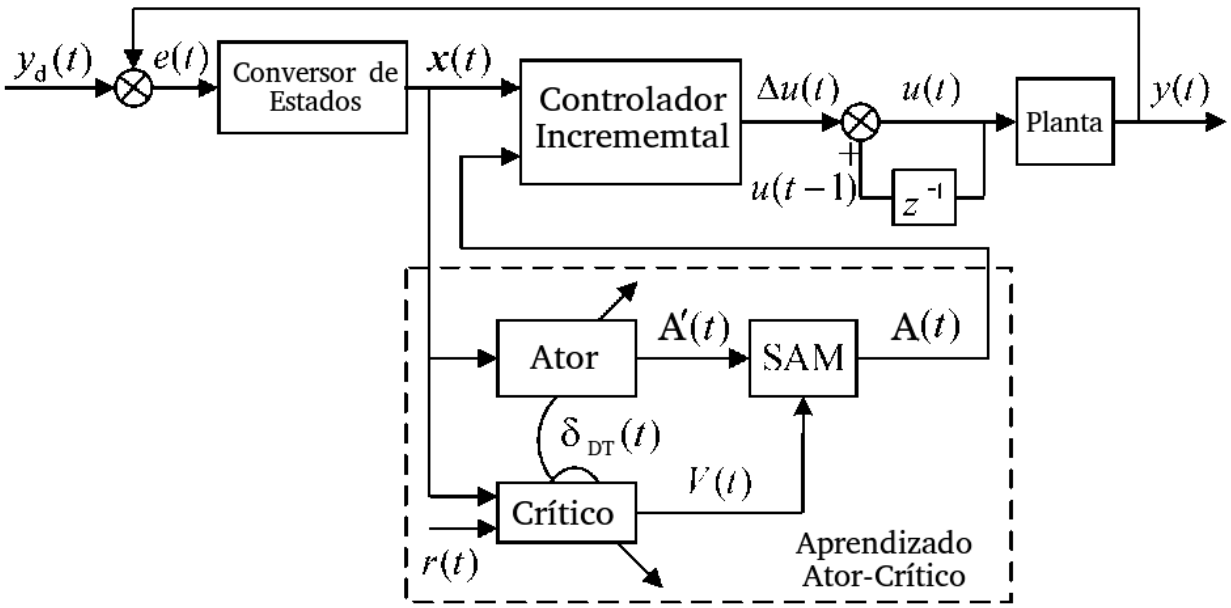

Figura 5.2: Sistema completo

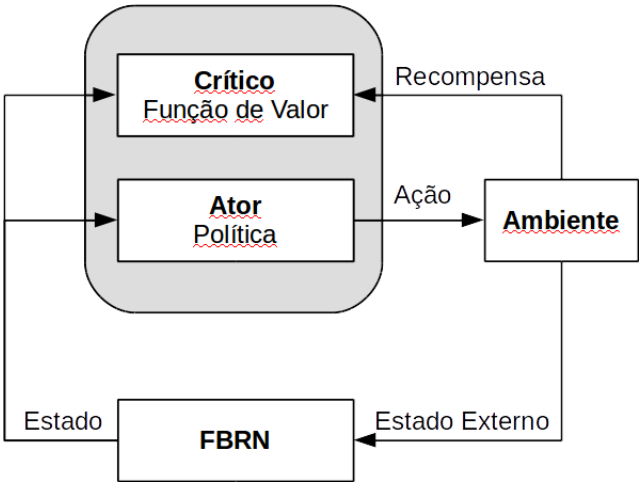

Figura 5.3: Arquitetura do ator-crítico

$$s_t = \left[\begin{array}{ccc} e(t) & \frac{de(t)}{dt} & \int_0^t e(\tau)d\tau \end{array} \right] \tag{5.1}$$

5.2.2 Arquitetura Ator-Crítico

A estratégia de controle utilizada possui a estrutura Ator-Crítico utilizando uma rede de funções de base radial para estimação dos estados do sistema, o crítico para avaliação da função de valor e o ator gerando a política de controle.

Como proposto por Cheng *et al.* [22], foi implementada a estrutura Ator-Crítico através de uma única rede de funções de Base Radial com entradas e camada oculta compartilhada. Isso é possível pois a entrada dos dois componentes é um vetor de características extraídas do ambiente, sendo necessário apenas alterar suas saídas.

Essa implementação, a partir de uma única rede pode acelerar o processo de aprendizado e reduzir a repetição de cálculos das saídas dos neurônios, [22]. As médias e desvios padrões da rede precisam ser treinados uma única vez, já os parâmetros do Ator e do Crítico devem ser treinados de forma diferenciada.

5.2.3 Rede de Funções de Base Radial Normalizadas

As funções de base radial normalizadas são utilizadas para condicionamento do sinal e aproximação de funções, havendo duas camadas inclusas na camada oculta, sendo a primeira a aplicação da função de ativação e a segunda a normalização.

As funções de ativação de base radial possuem a forma da equação 4.11:

$$\varphi_j(x) = \exp\left(-\frac{\left\|s_t - \vec{\mu}_j\right\|^2}{2\sigma_j^2}\right), \quad j = 1, 2, ..., h \tag{5.2}$$

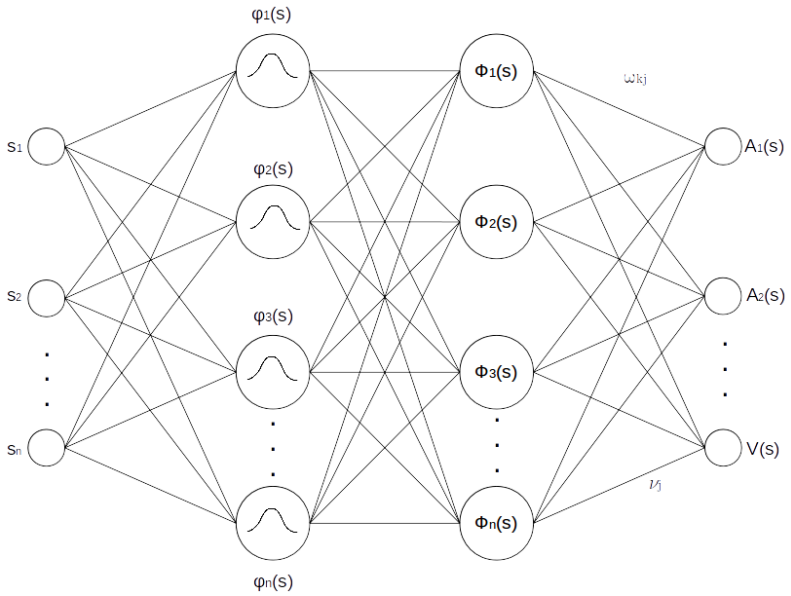

Figura 5.4: Topologia da rede

Sendo $\overrightarrow{\mu}_j = \begin{bmatrix} \mu_{1j} & \mu_{2j} & \mu_{3j} \end{bmatrix}$. A normalização é dada por:

$$\phi_j(s_t) = \frac{\varphi_j(s_t)}{\sum_{l=1}^{h} \varphi_l(s_t)} \tag{5.3}$$

Após a aplicação dos pesos na rede, a saída do ator é dada por:

$$A_k(s_t) = \sum_{j=1}^{h} \omega_{kj} \phi_j(s_t) \tag{5.4}$$

enquanto a saída do crítico é dada por:

$$V(s_t) = \sum_{j=1}^{h} \upsilon_j \phi_j(s_t) \tag{5.5}$$

5.2.4 Exploração e Aproveitamento

De forma a lidar com o dilema da exploração e aproveitamento, a saída do ator não é diretamente utilizada, passando antes por um modificador de ação estocástico (SAM, do inglês *Stochastic Action Modifier*), onde é adicionado um ruído gaussiano ao sinal de controle. Dessa forma, a nova saída do ator é dada por:

$$\hat{A}(s_t) = A(s_t) + \mathcal{N}(0, \sigma_v) \tag{5.6}$$

Ou seja, a saída do Ator é acrescida de um ruído gaussiano de média 0 e desvio padrão σ_v. Sendo σ_v dado por:

$$\sigma_v = \frac{k_1}{1 + \exp(k_2 V(s_t))} \tag{5.7}$$

Onde k_1e k_2 são constantes. É importante perceber que σ_v é decrescente, fazendo com que maiores valores de $V(s_t)$ gerem menores valores de σ_v e fazendo então a saída do SAM ser aproximadamente a saída recomendada pelo treinamento.

5.2.5 Erro de diferença temporal e recompensa imediata

De forma a utilizar o aprendizado sobre o ator e o crítico, é necessário inicialmente definir como será calculado o erro de diferença temporal.

Sabemos que o erro de diferença temporal é dado pela equação 4.3, porém ainda não foi definido o tipo de recompensa utilizada.

A recompensa adotada visa manter o erro dentro de uma faixa definida como aceitável, além de visar sempre uma redução desse erro. Sendo assim, a recompensa imediata é calculada considerando-se o erro do sistema e a taxa de variação do erro simultaneamente, de forma que o seu valor é dado por

$$r(t) \quad = \quad \alpha r_e(t) + \beta r_{ec}(t) \tag{5.8}$$

Em que α e β são coeficientes que ponderam o quanto cada um desses fatores é importante para o sistema. $r_e(t)$ e $r_{et}(t)$ são definidos da seguinte maneira

$$\begin{aligned} r_e(t) &= \begin{cases} 0, & |e(t)| \leq \varepsilon \\ -0.5, & c.c \end{cases} \\ r_{ec}(t) &= \begin{cases} 0, & |e(t)| \leq |e(t-1)| \\ -0.5, & c.c \end{cases} \end{aligned} \tag{5.9}$$

Em que ε é um parâmetro que define a faixa de tolerância [23]. Utilizou-se $\alpha = 0.4$, $\beta = 0.6$ e $\varepsilon = 0.3$.

Dessa forma, há condições suficientes para se prosseguir com o aprendizado do ator e do crítico.

5.2.6 Aprendizado do Ator

O ator é o responsável por definir a política de controle adotada. Pela abordagem utilizada, o Ator apresenta três saídas, sendo essas multiplicadas pelo erro, pela integral e pela derivada desse erro.

O método de aprendizado é a partir da diferença temporal, dada pela Equação 4.3 e utiliza como recompensa as definições apresentadas em 5.8.

O γ adotado foi 0.98, uma vez que um valor maior de γ indica uma maior importância dada a recompensas acumuladas.

A partir das equações 5.4 e 5.6 já obtidas para definição do ator e pela técnica de gradiente descendente, a variação de parâmetros do ator é dada por:

$$\Delta\omega_{kj} = \alpha_A \delta_{TD} \left[\frac{\hat{A}_k - A_k}{\sigma_v}\right]_{t=1} \phi_j(s_{t-1}) \tag{5.10}$$

onde α_A é a taxa de aprendizado do ator.

5.2.7 Aprendizado do Crítico

O crítico é responsável por avaliar a política adotada, além de ser extremamente importante na estimação do gradiente do ator, através do termo σ_v.

Muitas vezes, durante o aprendizado do crítico, a recompensa pode ser recebida com um atraso em relação às ações realizadas. Para isso utiliza-se o método TD(λ), apresentado nas equações 4.9 e 4.10:

$$\Delta\overrightarrow{v} = \alpha_C \delta_{DT} \overrightarrow{e}_{t-1}$$

$$\overrightarrow{e}_t \leftarrow \gamma\lambda\overrightarrow{e}_{t-1} + \nabla_{\overrightarrow{v}} V(s_{t-1})$$

Pela equação 4.8, a variação de pesos para o crítico é dada por:

$$\Delta\overrightarrow{v} = \alpha_C \delta_{TD}(\overrightarrow{\phi}(s_{t-1}) + \gamma\lambda\overrightarrow{e}_{t-1}) \tag{5.11}$$

5.2.8 Média e desvio padrão

Para melhor ajuste das funções de base radial na aproximação de parâmetros, foi utilizado treinamento para correção da média e desvio das gaussianas. O treinamento foi definido pelo gradiente descendente:

$$\mu_{ij}(t+1) = \mu_{ij}(t) + \eta_\mu \delta_{TD} \nabla_\mu V(s_t)$$

$$\mu_{ij}(t+1) = \mu_{ij}(t) + \eta_\mu \delta_{TD} \upsilon_j \phi_j(s_t) \frac{s_t(i) - \mu_{ij}(t)}{\sigma_j^2(t)} \tag{5.12}$$

Da mesma forma, para o desvio padrão:

$$\sigma_j(t+1) = \sigma_j(t) + \eta_\sigma \delta_{DT} \upsilon \Phi_j(s_t) \frac{\|s_t(j) - \mu_{i,j}(t)\|^2}{\sigma_j^3(t)} \tag{5.13}$$

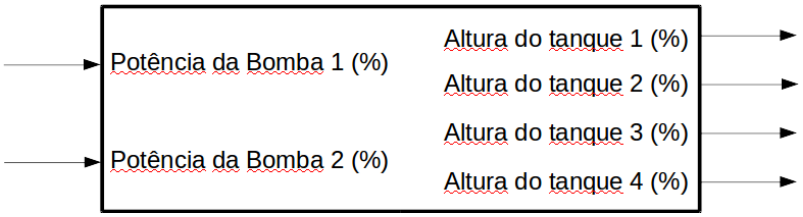

Figura 5.5: Entradas e saídas do sistema

5.3 Implementação

5.3.1 Simulação

Foi implementado inicialmente o controlador proposto em um ambiente simulado e, de forma a se estabelecer parâmetros comparativos, foram comparados os resultados com um controlador PI convencional.

O projeto do PI foi feito de maneira empírica, sendo estimado um valor adequado para os respectivos ganhos.

5.3.2 Planta Piloto

A implementação na planta real foi realizada através do Simulink, utilizando comunicação via OPC. A configuração das variáveis alteráveis, como potência das bombas e alturas nos tanques, não foi realizada como parte do trabalho, sendo possível apenas acessá-las. Para a implementação desse trabalho, considera-se invisível interação do computador por fora do Simulink com o controlador da planta.

As variáveis acessadas como medições são as alturas do tanques, em porcentagem e as vazões das bombas diretamente em cada tanque. Já as variáveis controláveis são a porcentagem de potência das bombas e a abertura de cada válvula automática. A figura 5.5 apresenta as entradas e saídas do sistema.

Como as válvulas motorizadas possuem um alto tempo para alteração, foi utilizado um valor fixado de abertura antes de cada experimento.

Capítulo 6

Resultados

Nesse capítulo serão apresentados os resultados em simulação do controlador proposto e serão apresentadas as dificuldades de implementação na planta piloto.

6.1 Resultados Simulados

Foram escolhidos 2 valores de nível para o tanque 2, de forma que, enquanto esse tanque mantém o nível desejado, o tanque 1 varia entre alguns níveis arbitrários.

O sinal de controle foi saturado para que fique dentro da faixa de alimentação do inversor que controla a tensão da bomba. A faixa de operação utilizada foi entre 0 e 100%.

O sinal de referência dura 30 segundos para cada nível de avaliação no tanque 1, e dura 150 segundos para cada nível de avaliação no tanque 2.

As Figuras 6.1 e 6.3 mostram, respectivamente, a performance dos dois controladores para o tanque 1 e o erro do nível do tanque 1 em relação à referência. Já as figuras 6.2 e 6.4 mostram a performance dos controladores para o tanque 2 e o erro apresentado. Na 6.5 verificam-se os sinais saturados enviados ao sistema e os sinais não saturados calculados pelos controladores.

Das Figuras 6.1 e 6.2 percebe-se que o controlador adaptativo tem performance superior no que diz respeito à velocidade de resposta e sobrepasso. Na maior parte dos pontos de operação, o controlador adaptativo proposto é tão ou mais rápido que o controlador PI, não apresentando o mesmo sobrepasso.

É possível verificar na Figura 6.2 que o controlador proposto é mais resistente a perturbações. Para cada variação de nível no tanque 1, o nível no tanque 2, para o sistema com um controlador tradicional, oscila significativamente, ao passo que o sistema com o controlador proposto oscila pouco, ou não oscila.

A Figura 6.3 mostra que as curvas de erro se assemelham, mas o erro quadrático médio mostra uma diferença considerável entre a performance do controlador PI e do controlador adaptativo. Os erros quadráticos médios obtidos foram 13.11 e 10.83 para o controlador PI e para o controlador

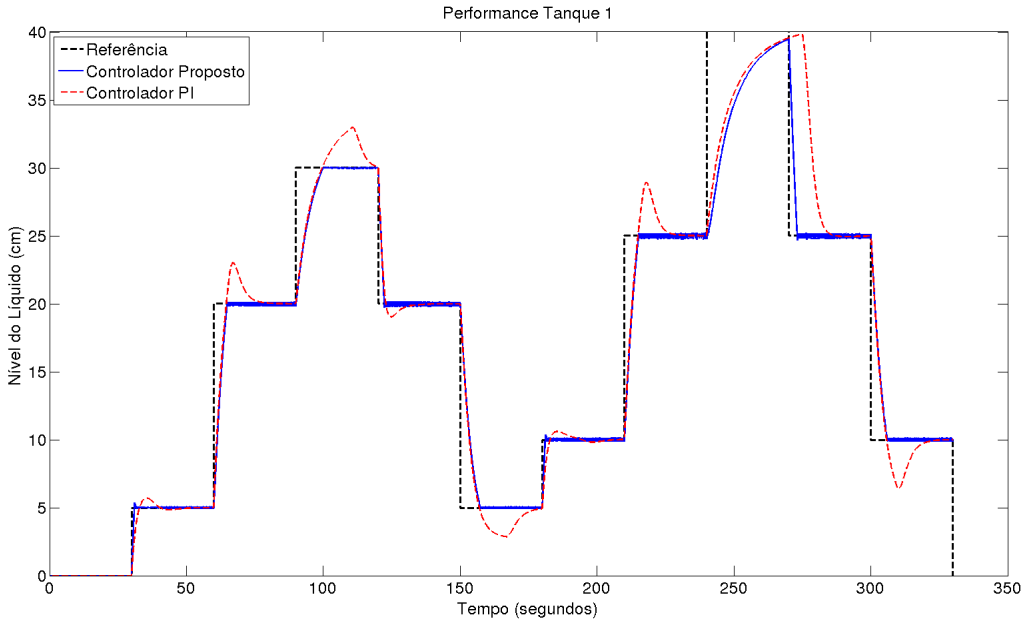

Figura 6.1: Performance do Tanque 1

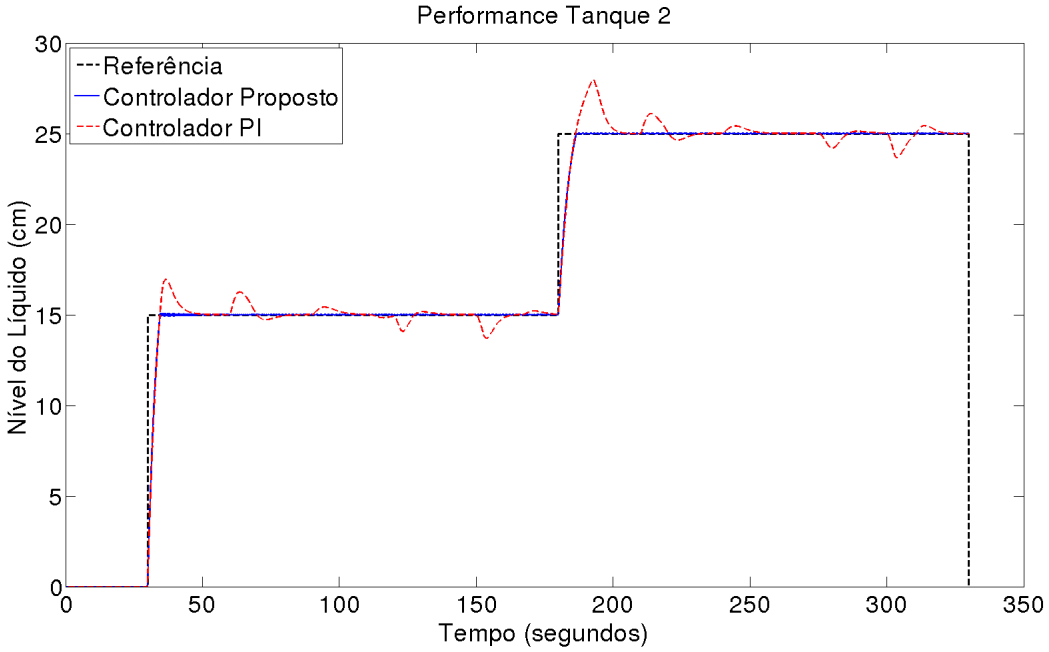

Figura 6.2: Performance do Tanque 2

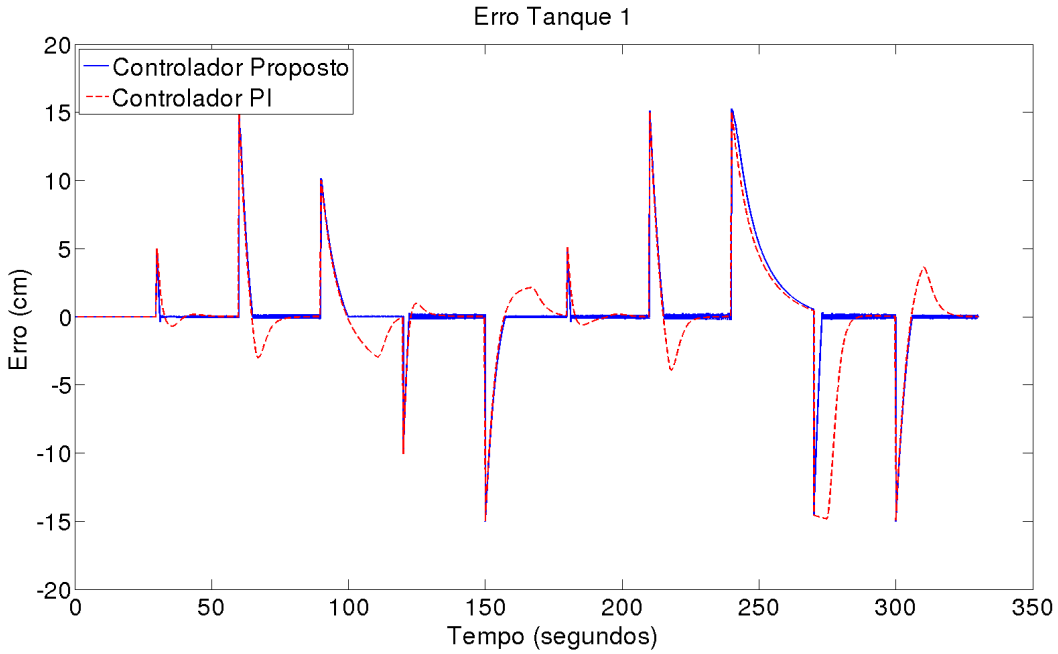

Figura 6.3: Erro do Tanque 1

adaptativo, respectivamente.

É importante perceber que em diversos momentos, os dois sinais de saída são idênticos. Observando a figura 6.5 percebe-se que isso ocorre pelo fato de que nesses momentos o sinal de controle está saturado nos dois casos, de modo que as coincidências ocorrem, na verdade, no limite da saturação.

Para efeitos de simulação, considerou-se o inversor da bomba com resposta ideal, no entanto, pela Figura 6.5 percebe-se que o sinal de controle tem variações muito abruptas , de forma que, o sinal simulado não é adequado para aplicação no sistema real.

6.1.1 Dificuldades encontradas

O primeiro passo da aplicação no sistema real foi a análise de perturbações observadas. Primeiramente foi notado que, com os tanques vazios, a medição apresenta um valor diferente de zero. A causa mais provável desse offset é a altura do sensor dentro do tanque, que faz com que valores muito baixos de altura sejam desconsiderados.

Apesar das diversas vantagens de uso do inversor [24], foi notado que a leitura dos sensores é alterada no momento que os inversores são ligados. A utilização do inversor gera ruídos, além de causar alteração nas leituras dos sensores.

A figura 6.6 mostra a leitura dos tanques vazios. A partir de 150 segundos da leitura, o inversor é ligado a um valor de potência de 10%, situado na zona morta de atuação e fazendo com que não haja entrada líquido nos tanques. Além da alteração no valor medido há uma alta taxa de ruídos, que prejudica as medições mesmo com a aplicação de um filtro passa baixas. O filtro aplicado

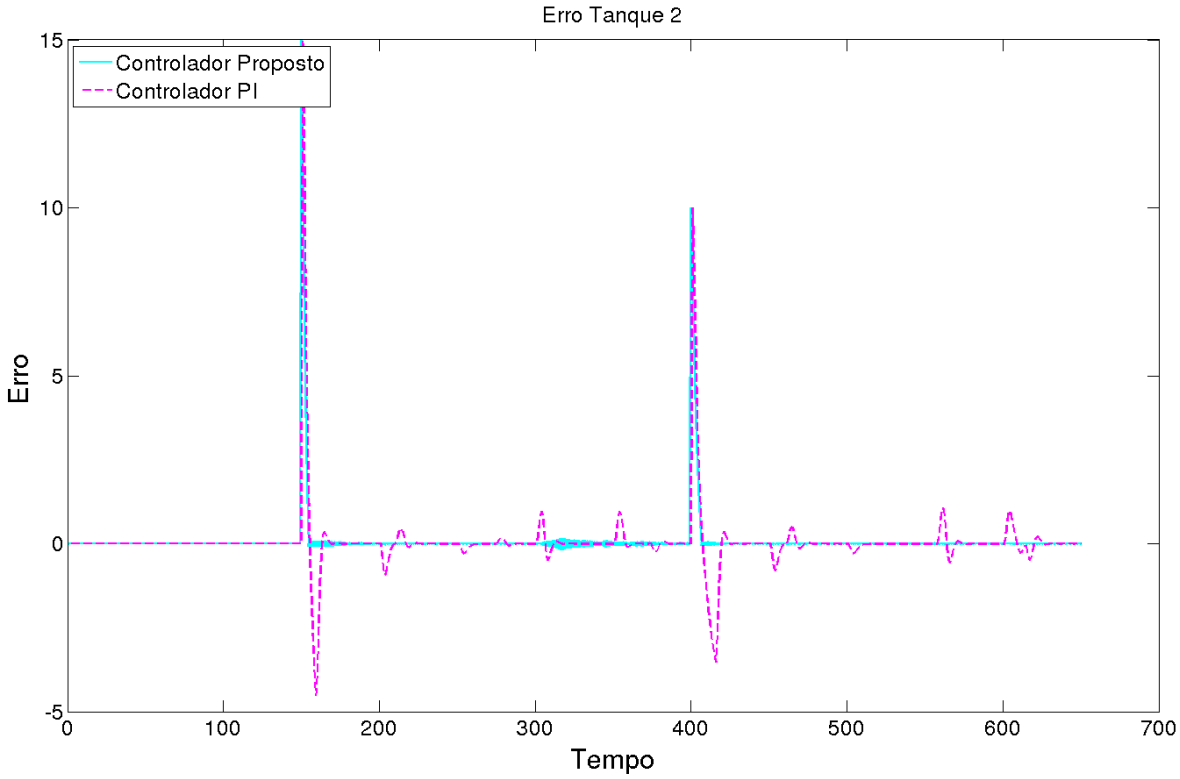

Figura 6.4: Erro do Tanque 2

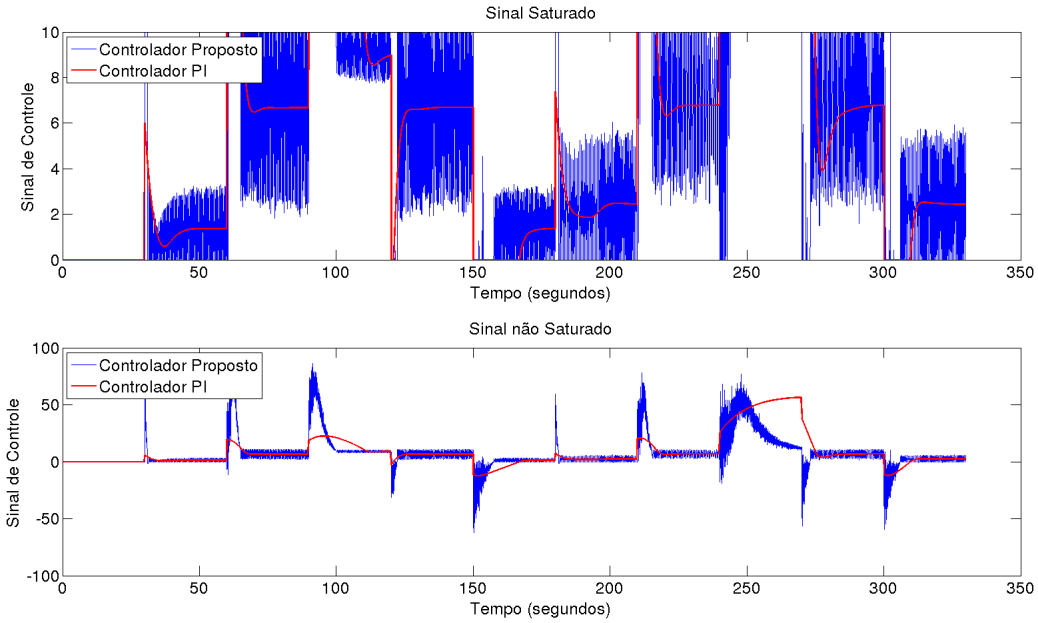

Figura 6.5: Sinal de Controle para o tanque 1

	Bomba 1	Bomba 2
Limite inferior	24%	29%
Limite Superior	90%	90%
Vazão Máxima (L/s)	4,2	2,4

Tabela 6.1: Limites dos atuadores

possui função de transferência $\frac{1}{2s+1}$.

Essa variação na leitura do sistema, enquanto o nível do líquido não se altera, gera problemas sérios na implementação do controlador, uma vez que a leitura não está confiável e não é possível observar com precisão o regime permanente para a altura do tanque.

Além dos ruídos percebidos sobre as medições do tanque vazio, foi observada a resposta do sistema a um degrau de 50% da potência total. Foram utilizadas as configurações de fase mínima e não-mínima do sistema, onde as válvulas para o primeiro caso se encontram a uma proporção $\gamma_1 = \gamma_2 = 0.7$ e no segundo $\gamma_1 = \gamma_2 = 0.3$. A figura 6.7 mostra que o regime permanente é alcançado, porém as medidas realizadas pelos sensores são bastante ruidosas.

Um fator importante é a presença de uma zona morta do atuador. Para verificar a zona-morta, foi aplicada uma rampa na entrada do sistema, variando a potência de 0 a 100% em 300 segundos. De forma a se verificar a resposta do atuador diretamente sobre a saída do sistema, foi utilizado $\gamma_1 = \gamma_2 = 1$, ou seja, os tanques inferiores com abertura das válvulas a 100% e os superiores a 0%.

Na figura 6.8, é apresentado o ponto onde a variação de cada bomba tem efeito sobre as saídas. É possível então delimitar uma zona morta de aproximadamente 24% para a bomba 2 e de aproximadamente 29% para a bomba 1. Para lidar com esse problema, foi aplicado um offset à entrada, mantendo a bomba sempre no limiar de funcionamento.

Além da zona morta, o atuador é saturado em 90% da potência total. Essa limitação foi implementada pelo fabricante da planta no momento da instalação do sistema e faz com que valores acima de 90% sejam desconsiderados. Essa saturação precisou ser incluída antes da planta, para evitar operar em uma faixa que não tenha resposta na saída do sistema. Foi notado também que manter a potência a 100%, ou tentar incluir valores ainda maiores, faz com que o atuador seja desarmado. Dessa forma, os limites dos atuadores foram definidos na tabela 6.1.

Devido a todas essas limitações, o controlador adaptativo não foi implementado com sucesso. Os problemas na medição, juntamente com uma bomba de vazão reduzida, fazem com que a faixa de operação seja muito baixa e com uma alta variação nas leituras.

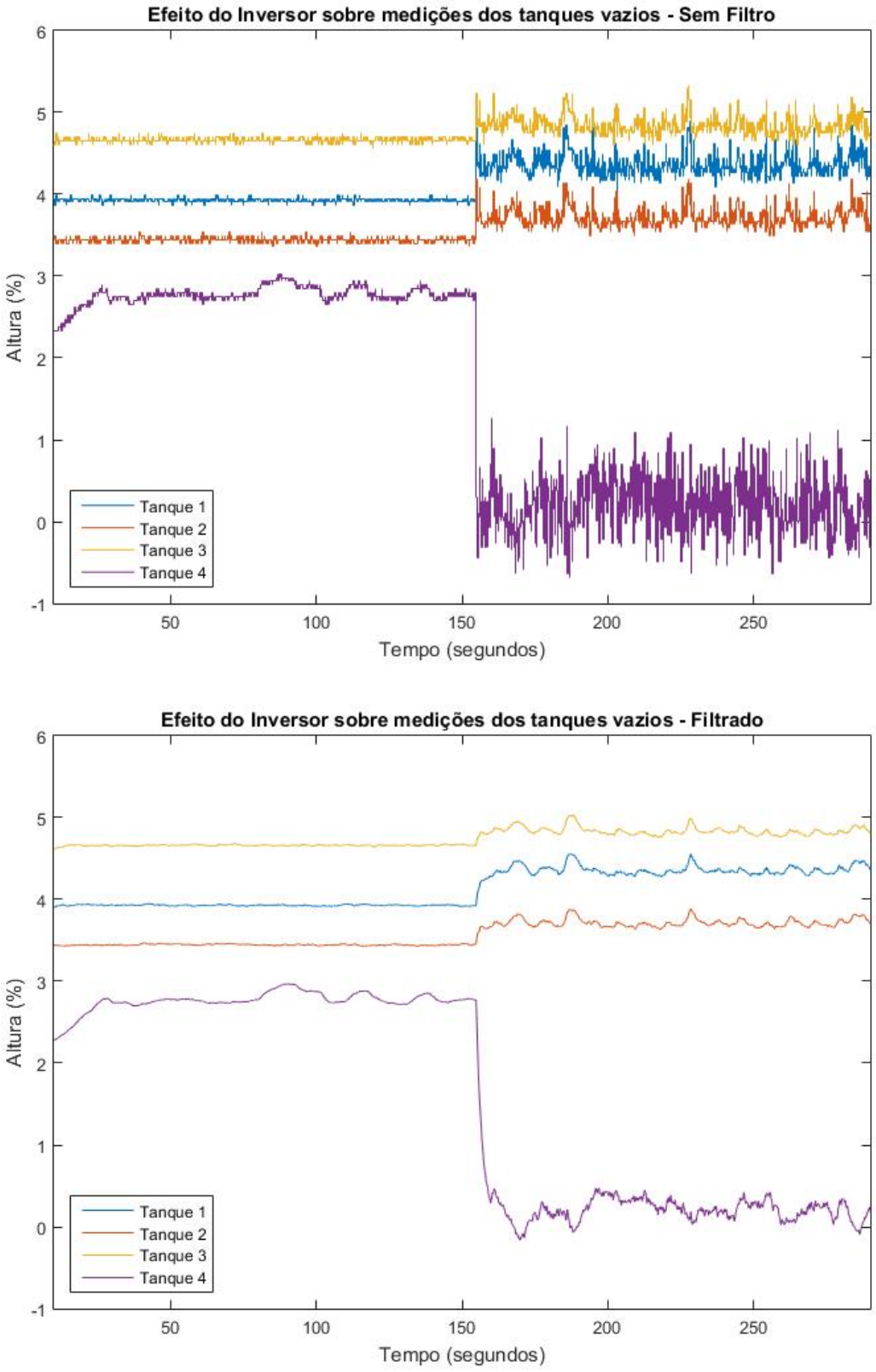

Figura 6.6: Efeito do inversor sobre as medições

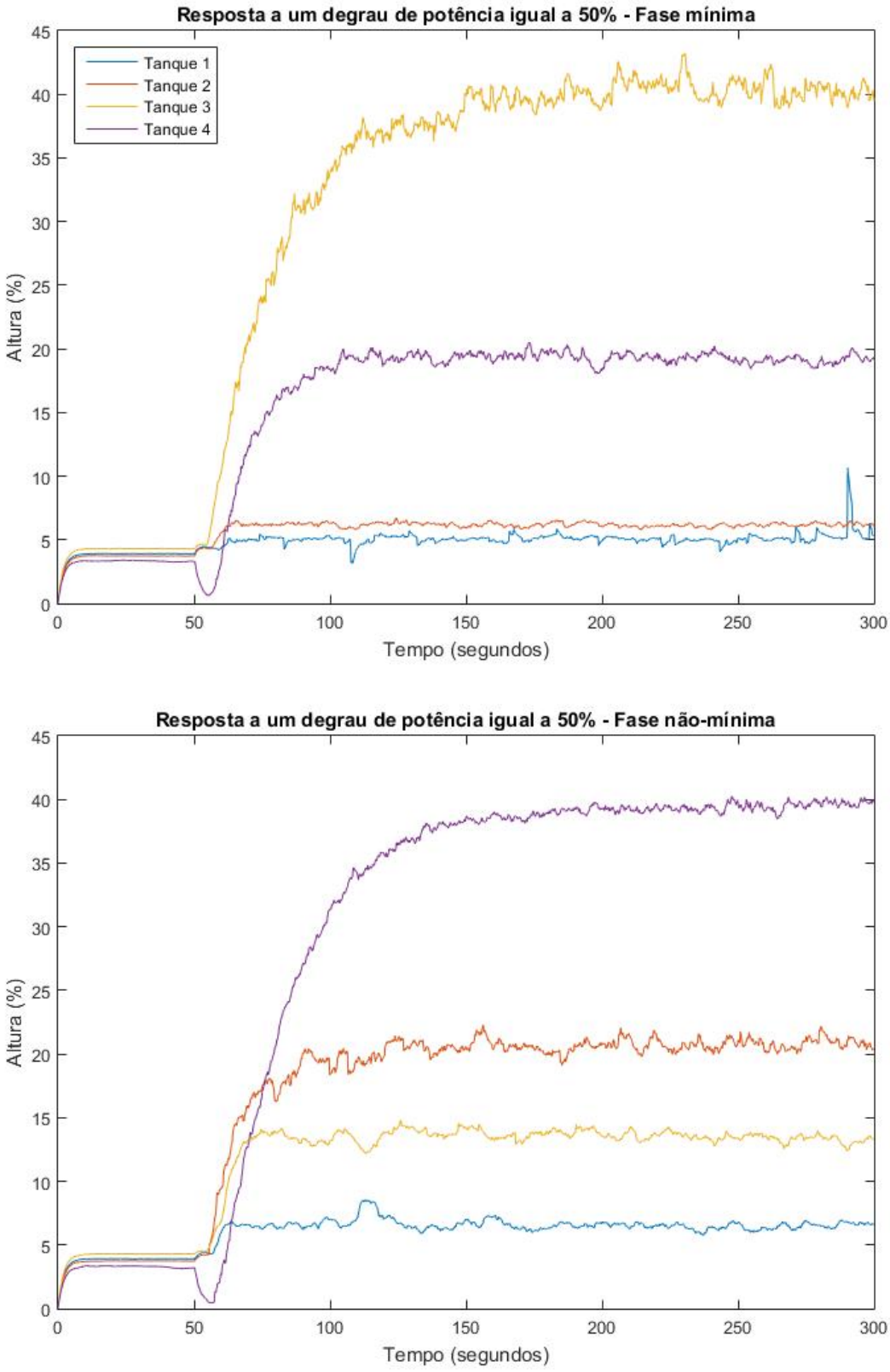

Figura 6.7: Resposta ao degrau

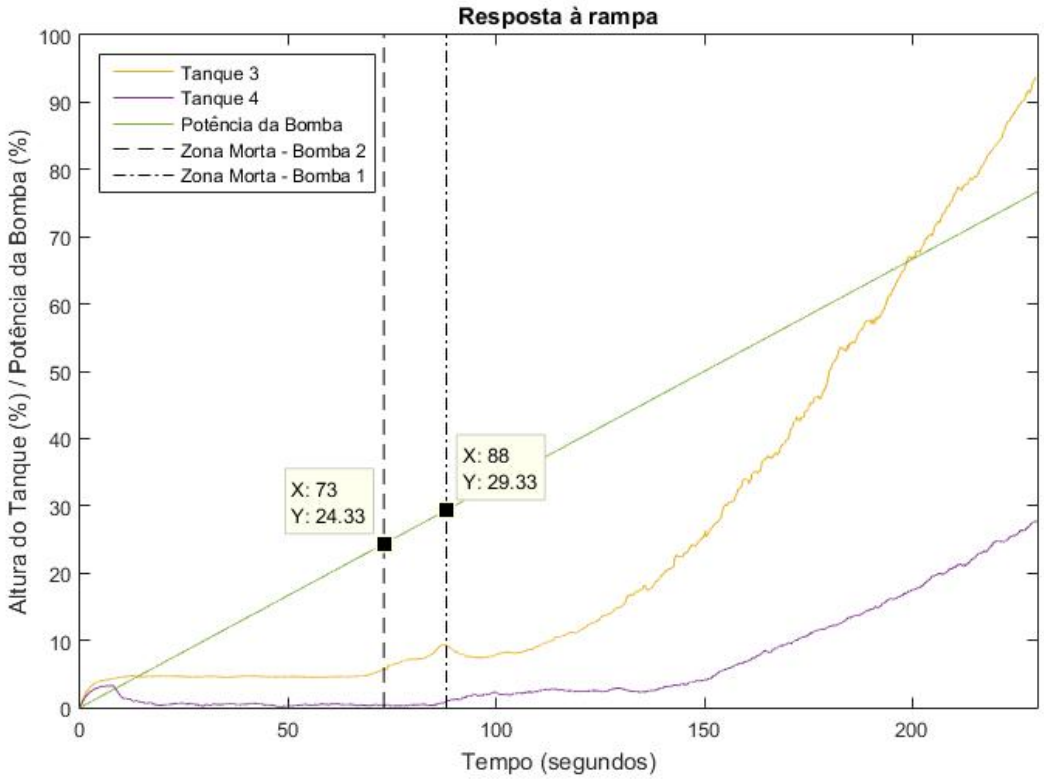

Figura 6.8: Resposta do sistema a uma rampa

Capítulo 7

Conclusões

A partir dos resultados apresentados, foi mostrada a veracidade matemática do algoritmo utilizado. O aprendizado por reforço é uma técnica adequada para controle de processos, porém a implementação na planta-piloto não pôde comprovar tal fato, uma vez que o processo real não cumpre alguns requisitos importantes para a implementação.

Em um ambiente sem falhas de medição, como o caso da simulação, se mostrou possível o uso do controlador utilizando aprendizado por reforço. As falhas de medição notadas durante a implementação do sistema real geram uma oscilação grande no valor medido, fazendo com que não haja confiança em relação à altura da coluna de água no tanque, tornando difícil a aplicação de qualquer técnica de controle e em especial técnicas como a proposta, que observa a saída e a tendência de alteração do sinal para gerar sua estratégia de controle.

O inversor de frequência fornece diversas vantagens no acionamento de um motor elétrico, como a remoção da alta corrente de partida do motor, porém se não for tratado com cuidado pode interferir em outras partes da instrumentação do processo, como no processo utilizado. O inversor de frequências exerceu forte efeito sobre as medições, mesmo com alguns passos realizados tentando minimizá-lo, como por exemplo o aterramento do inversor e o isolamento elétrico da fonte de alimentação para instrumentação.

Foi perceptível que, apesar da técnica de controle se mostrar realizável e promissora em simulações, a instrumentação de medição e atuação sobre o sistema pode tornar esse controle muito mais difícil ou até irrealizável.

7.1 Perspectivas Futuras

Durante esse trabalho foram realizados testes de aterramento da planta e isolamento da fonte dos instrumentos de medição, porém não foi percebida grande alteração nas medições ruidosas. Como proposta de trabalho futuro, seria adequada a verificação da instalação elétrica da planta, uma vez que os ruídos gerados pelo inversor podem ocorrer devido a fugas de tensão, e o isolamento do inversor de frequências, uma vez que esse pode causar interferências tanto conduzidas pela rede

elétrica quanto irradiadas devido a sua alta taxa de chaveamento [25].

Outro ponto percebido como possível causa de ruídos seria a forma como a água escoa para os tanques. A entrada de líquido ocorre pelo topo do tanque, gerando variações de medição ao atingir a coluna de água. Como sugestão também seria adequada a instalação de um sistema para redução dessas turbulências, como anteparos para redução da velocidade da água e redução de ondas de presso sobre o sensor.

REFERÊNCIAS BIBLIOGRÁFICAS

[1] VRABIE, D.; VAMVOUDAKIS, K. G.; LEWIS, F. L. *Optimal adaptive control and differential games by reinforcement learning principles*. [S.l.]: IET, 2013.

[2] SINGH, S. P.; SUTTON, R. S. Reinforcement learning with replacing eligibility traces. *Machine learning*, Springer, v. 22, n. 1-3, p. 123–158, 1996.

[3] GARCIA, C. *Modelagem e Simulação de Processos Industriais e de Sistemas Eletromecânicos Vol. 1*. [S.l.]: EdUSP, 2005.

[4] JOHANSSON, K. H. The quadruple-tank process: a multivariable laboratory process with an adjustable zero. *IEEE Transactions on control systems technology*, IEEE, v. 8, n. 3, p. 456–465, 2000.

[5] OGATA, K.; YANG, Y. Modern control engineering. Prentice-Hall Englewood Cliffs, 1970.

[6] NOEL, M. M.; PANDIAN, B. J. Control of a nonlinear liquid level system using a new artificial neural network based reinforcement learning approach. *Applied Soft Computing*, Elsevier, v. 23, p. 444–451, 2014.

[7] DIDATICONTROL. *DIDATICONTROL Automatizando a Ciencia*. 2016. Disponível em: <http://didaticontrol.com.br>.

[8] SYAFIIE, S. et al. Learning control for batch thermal sterilization of canned foods. *ISA transactions*, Elsevier, v. 50, n. 1, p. 82–90, 2011.

[9] LIN, W.-S.; ZHENG, C.-H. Constrained adaptive optimal control using a reinforcement learning agent. *Automatica*, Elsevier, v. 48, n. 10, p. 2614–2619, 2012.

[10] SYAFIIE, S. et al. Model-free control based on reinforcement learning for a wastewater treatment problem. *Applied Soft Computing*, Elsevier, v. 11, n. 1, p. 73–82, 2011.

[11] BAYIZ, Y. E. *Multi-Agent Actor-Critic Reinforcement Learning for Cooperative Tasks*. Tese (Doutorado) — TU Delft, Delft University of Technology, 2014.

[12] SUTTON, R. S.; BARTO, A. G. *Reinforcement learning: An introduction*. [S.l.]: MIT press Cambridge, 1998.

[13] SUTTON, R. S. Learning to predict by the methods of temporal differences. *Machine learning*, Springer, v. 3, n. 1, p. 9–44, 1988.

[14] KONDA, V. R.; TSITSIKLIS, J. N. Actor-critic algorithms. In: *NIPS*. [S.l.: s.n.], 1999. v. 13, p. 1008–1014.

[15] WATKINS, C. J.; DAYAN, P. Q-learning. *Machine learning*, Springer, v. 8, n. 3-4, p. 279–292, 1992.

[16] RUMMERY, G. A.; NIRANJAN, M. *On-line Q-learning using connectionist systems*. [S.l.]: University of Cambridge, Department of Engineering, 1994.

[17] GRONDMAN, I. et al. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, IEEE, v. 42, n. 6, p. 1291–1307, 2012.

[18] WITTEN, I. H. An adaptive optimal controller for discrete-time markov environments. *Information and control*, Elsevier, v. 34, n. 4, p. 286–295, 1977.

[19] BARTO, A. G.; SUTTON, R. S.; ANDERSON, C. W. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, IEEE, n. 5, p. 834–846, 1983.

[20] HAYKIN, S. S. Redes neurais artificiais: princípio e prática. *2ª Edição, Bookman, São Paulo, Brasil*, 2000.

[21] MOODY, J.; DARKEN, C. J. Fast learning in networks of locally-tuned processing units. *Neural computation*, MIT Press, v. 1, n. 2, p. 281–294, 1989.

[22] CHENG, Y.-H.; YI, J.-Q.; ZHAO, D.-B. Application of actor-critic learning to adaptive state space construction. In: IEEE. *Machine Learning and Cybernetics, 2004. Proceedings of 2004 International Conference on*. [S.l.], 2004. v. 5, p. 2985–2990.

[23] WANG, X.-S.; CHENG, Y.-H.; WEI, S. A proposal of adaptive pid controller based on reinforcement learning. *Journal of China University of Mining and Technology*, Elsevier, v. 17, n. 1, p. 40–44, 2007.

[24] RODRIGUES, W. et al. Critérios para o uso eficiente de inversores de frequência em sistemas de bombeamento de água. *Campinas: Tese de Doutorado, FEC, Unicamp*, 2007.

[25] SOUSA, A. O.; SILVA, M. M. da; PIRES, I. A. Estudo de interferências na alimentação elétrica de motores de indução por inversores de frequência. *e-xacta*, v. 5, n. 1, 2012.

ANEXOS

I. DESCRIÇÃO DO CONTEÚDO DO CD

Descrever CD.

II. PROGRAMAS UTILIZADOS

Quais programas foram utilizados?

II. ARTIGO SUBMETIDO AO CBA

CONTROLE ADAPTATIVO DE NÍVEL DE LÍQUIDOS UTILIZANDO APRENDIZADO POR REFORÇO

Email:

Abstract— This work presents an adaptative controller project using reinforcement learning and compares its performance against traditional controllers for a quadruple-tank process. The controller was designed with the Actor-Critic method using radial basis networks. The training was performed using the gradient descent of the temporal difference error. The results show superior performance of the adaptive controller regarding response speed and showing smaller mean squared error, also, the adaptative controller can manage system parameter variation while traditional controllers cannot, thus using adaptative controllers brings significative advantage.

Keywords— Adaptative Control, Reinforcement Learning, Actor-Critic, Temporal Difference, Quadruple-tank Process.

Resumo— O presente trabalho apresenta o projeto de um controlador adaptativo que utiliza aprendizado por reforço e compara sua performance com controladores tradicionais em um processo de 4 tanques. O controlador utiliza a abordagem Ator-Crítico com redes neurais de base radial e o treinamento é feito através do gradiente descendente do erro de diferença temporal. Os resultados demonstram performance superior para o controlador adaptativo, com resposta mais rápida e menor erro quadrático médio, além disso, o controlador adaptativo é capaz de lidar com variações de parâmetros do sistema, enquanto controladores tradicionais não possuem essa capacidade, de forma que a utilização de controladores adaptativos apresenta vantagens significativas.

Palavras-chave— Controle Adaptativo, Aprendizado por Reforço, Ator-Crítico, Diferença Temporal, Processo de 4 tanques.

1 Introdução

Controle de nível de líquidos com a interação de múltiplos tanques é um problema comum em processos industriais, principalmente nos ramos químico, petroquímico, de celulose e de alimentos. Um exemplo de sistema de controle com quatro tanques foi proposto por Johansson (2000).

O sistema proposto possui comportamento não linear, fazendo com que estratégias de controle convencionais, como o controlador PID, apresentem uma alteração de comportamento caso a referência esteja longe de um ponto de operação arbitrário, uma vez que esses controladores são projetados para o o sistema linearizado nesse ponto.

Utilizou-se a abordagem de controle adaptativo, que apresenta uma melhor resposta se comparada a métodos de controle convencional devido à sua característica incremental e capacidade de aprendizado.

O método utilizado para adaptação foi aprendizado por reforço, no qual o controlador utiliza dados experimentais para avaliar sua política de ações e corrigir seus parâmetros. Técnicas de aprendizado por reforço em controle já foram demonstradas com sucesso por Deisenroth et al (2009), Syafiie et al (2011) e Wei-Song Lin et al (2012).

A estrutura de aprendizado é baseada na abordagem Ator-Crítico, proposta por Barto et al (1983), em que há um agente que avalia o desempenho do sistema e outro que realiza a política de controle. A cada passo os agentes corrigem seus parâmetros e buscam otimizar a resposta.

Além de melhor performance, a abordagem utilizada permite o controle de um sistema sem qualquer conhecimento de sua dinâmica, sendo uma opção comum no caso de sistemas muito complexos ou com parâmetros desconhecidos.

2 Aprendizado por Reforço

2.1 Conceito

Como apresentado por Sutton e Barto (1988), o conceito de aprendizado por reforço tem suas raízes na psicologia, porém ganhou amplo espaço na área de inteligência artificial e aprendizado de máquinas. O aprendizado acontece reforçando-se uma ação que produz resultados satisfatórios ou positivos, enquanto ações com resultados indesejados devem ser reprimidas.

A atividade de aprendizado pode ser definida por um agente executando ações e interagindo com o ambiente. Cada ação gera uma resposta do ambiente, que por sua vez, gera uma recompensa, calculada por um função de recompensa, podendo o resultado dessa função ser positivo ou negativo. A partir dessa resposta o agente irá moldar suas ações de modo a acumular o máximo de recompensas positivas. A recompensa representa o quão desejável é aquele estado para o comportamento ideal do agente, penalizando escolhas ruins. Como a função de recompensa é utilizada para avaliar o comportamento do agente, o mesmo não pode alterá-la, e ela será utilizada para atualizar sua

política de decisões.

As ações tomadas a partir de cada estado definem a política de comportamento do agente, podendo ser dada tanto por funções, tabelas, espaços de busca, ou, em geral, políticas estocásticas. Essa política será alterada durante o processo de aprendizagem.

A recompensa define a qualidade imediata de uma ação, porém o agente tem como objetivo obter a máxima recompensa acumulada, dessa forma, é definido o conceito de função de valor, que determina a qualidade de uma política ao longo do tempo. A função de valor define a recompensa esperada pela tomada de determinadas ações, podendo ser avaliada a partir da recompensa por atingir um dado estado. Ações podem gerar uma baixa recompensa e ainda assim apresentar um alto valor, o que ocorre caso as demais ações executadas sejam significantemente positivas, dessa forma, apesar do objetivo do aprendizado ser minimizar o custo, a função de valor tem caráter mais significativo, pois representa a possibilidade de maiores recompensas futuras ao se aplicar uma política.

2.2 Aprendizado por Diferença Temporal

Dentre os diversos métodos de aprendizado por reforço, um que possui grande destaque é o aprendizado por diferença temporal (ADT). Esse método consiste na utilização de experiências obtidas para atualizar a função de valor. O processo usa estimações anteriores para realizar novas estimações, técnica conhecida como *bootstrapping*.

Ao utilizar experiência por interação, o ADT remove a necessidade de um modelo do sistema, podendo ser aplicado em sistemas com dinâmica desconhecida.

Após a execução de uma ação, a recompensa e o valor do novo estado são utilizados para estimar o erro de diferença temporal δ_{DT}, em que δ_{DT}é definido como

$$\delta_{DT} = r_{t+1} + \gamma V(s_{t+1}) - V(s_t) \quad (1)$$

O termo r_{t+1} é a recompensa observada após a tomada de ação, $V(s_{t+1})$ é a função de valor estimada, $V(s_t)$ é a função de valor do estado e γ é o coeficiente que garante que o valor de um estado mais distante seja menos relevante. Os termos s_{t+1} e s_t são o estado seguinte estimado e o estado atual, respectivamente. O índice t significa o instante em que esses valores são utilizados.

A cada passo, o erro de diferença temporal afeta o valor de todas as ações realizadas anteriormente. Sendo assim, o valor de cada estado pode ser atualizado de acordo com

$$\Delta V_t(s_t) = \alpha \delta_t e_t(s_t) \quad (2)$$

O termo $\Delta V_t(s)$ representa a variação no valor de cada estado, α é a de aprendizado, $e(s_t)$ representa o quanto o estado é significativo para a ação atual, para que estados antigos não sejam afetados por ações mais recentes.

3 Redes Neurais de Base Radial

3.1 Funções de Base Radial

Uma Função de Base Radial, ou FBR, é uma função cujo valor depende exclusivamente da distância a um ponto central, sendo que a FBR cresce ou decresce monotonicamente com a distância ao centro.

Uma Função de Base Radial muito comum é a Gaussiana, Equação (3). Os parâmetros da função são o centro c_i, tido como referência, e o seu desvio padrão σ_i que define a dispersão da função a partir do centro. (Orr, 1996; Papierok, et al. 2008).

$$f(x) = exp(-\frac{||x - c_i||^2}{2\sigma_i^2}) \quad (3)$$

3.2 Aproximação de Funções com Redes FBR

Como visto na Subseção 2.2, o método ADT utiliza uma determinada função de valor para calcular δ_{DT}, e também, como o método tem capacidade de utilizar sua experiência imediatamente, é necessária alguma técnica que aproxime a função de valor a cada passo.

Uma FBR é tipicamente utilizada para fazer a aproximação de funções na forma apresentada na equação (4), onde a função é representada como um soma de Gaussianas, cada uma com um centro c_i e ponderadas por um coeficiente w_i adequado. (Orr, 1996; Papierok et al 2008).

$$Q_t(s, a) = \sum_{i=1}^{n} w_i f(||x - c_i||) \quad (4)$$

Essa aproximação pode ser interpretada como uma rede neural simples, de apenas uma camada, e é conhecida como Rede de Funções de Base Radial (RFBR). A Figura 1 mostra a estrutura de uma RFBR utilizada para aproximação da função de valor. (Papierok et al 2008).

Uma característica importante das RFBR é que a função aproximada deve ser sempre diferenciável com relação aos coeficiente w_i, de forma que seja possível utilizar qualquer técnica tradicional de aprendizado, como por exemplo o método de gradiente decrescente, para a atualização desses coeficientes. (Orr, 1996; Papierok et al 2008).

3.3 Método Actor-Critic

O Método *Actor-Critic*, ou Ator-Crítico definido por Sutton e Barto (1988), é uma técnica de ADT

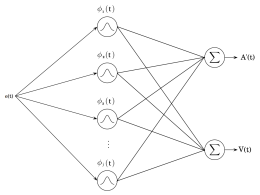

Figura 1: Aproximação de Funções por uma RFBR

que diferencia explicitamente a política da função de valor.

A estrutura que detém a política é chamada de Ator e é responsável pela decisão e tomada de ação, e a estimação da função de valor é chamada de Crítico, responsável por avaliar as ações tomadas pelo Ator. O Crítico é utilizado para o cálculo de δ_{DT}, definido pela Equação 1, que é, por sua vez, utilizado para fazer a avaliação da ação tomada, verificando se o resultado obtido foi melhor ou pior e gerando a mudança de políticas e mudança de sua própria avaliação, como é verificado na Figura 2.

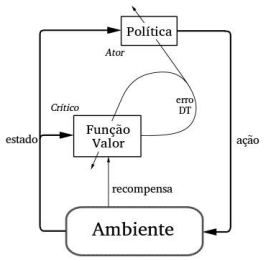

Figura 2: Estrutura do Método Ator-Crítico

4 Implementação em um Sistema de 4 Tanques.

4.1 Definição do Processo

A Figura 3 apresenta o sistema de 4 tanques proposto por Johansson (2000). As válvulas de

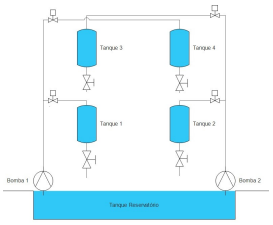

Figura 3: Sistema de 4 Tanques Utilizado

controle da vazão são reguláveis, de modo que a proporção da vazão da bomba que vai para cada tanque é configurável. O Modelo foi implementado como simulação para verificar a possibilidade de implementação em um sistema real.

Da figura nota-se que, a menos que as válvulas de controle de vazão estejam fechadas, ambas as bombas influenciam o nível nos dois tanques inferiores de maneira direta ou indireta através dos tanques superiores, de forma que, para controlar um desses tanques, gera-se perturbação no outro. Pelo mesmo fato, existem algumas configurações de nível, estados do sistema, que são inalcançáveis. Pecebe-se, por exemplo, que a menos que as válvulas estejam completamente abertas ou fechadas é impossível que um tanque inferior tenha um certo valor de nível e o outro esteja completamente vazio, portanto, o controle será executado para configurações de nível que sejam fisicamente possíveis. A faixa de valores de níveis admissíveis muda de acordo com a configuração das válvulas de controle de vazão.

O sistema é modelado por

$$\begin{aligned} h_1' &= \frac{1}{A_1}(o_3\sqrt{2gh_3} + \lambda_1 k_1 u_1 - o_1\sqrt{2gh_1}) \\ h_2' &= \frac{1}{A_2}(o_4\sqrt{2gh_4} + \lambda_2 k_2 u_2 - o_2\sqrt{2gh_2}) \quad (5) \\ h_3' &= \frac{1}{A_3}((1-\lambda_2) k_2 u_2 - o_3\sqrt{2gh_3}) \\ h_4' &= \frac{1}{A_4}((1-\lambda_1) k_1 u_1 - o_4\sqrt{2gh_4}) \end{aligned}$$

Em que h_i' indica a derivada do nível h_i, g é a aceleração da gravidade, aproximada para $1000\frac{cm}{s^2}$. O parâmetro λ_i é a representação da configuração das válvulas e define o quanto da vazão da bomba cada tanque recebe. Utilizou-se para as duas válvulas $\lambda = 0.7$. O fator u_i é a tensão de controle de cada bomba, variando de 0 a 10 e k_i é a constante que relaciona a tensão aplicada com a vazão de saída da bomba. Para as duas bombas, $k = 166.67cm^3/\text{sV}$, com resposta ideal. O

parâmetro A_i é a área da base do tanque i. O valor dessa área é $200.96cm^2$ para todos os tanques. O parâmetro o_i é a área do canal de escoamento de cada tanque. O valor é $5.06cm^2$ para todos os tanques.

4.2 Implementação

Para efeito de comparação, um controlador PI foi implementado empiricamente para o ponto de operação. Os ganhos do canal Proporcional e Integrador utilizados foram 1.2 e 0.4 respectivamente.

Cada controlador utiliza o erro do nível de um dos tanques e envia um sinal de controle para a bomba de número respectivo.

A entrada da RFBR é o erro da saída do sistema em relação à referência. A RFBR, por sua vez, define o Ator e o Crítico. O Crítico é utilizado para calcular δ_{DT} e também, junto com o Ator, calcular o sinal de controle através de um modificador de ação estocástico (*Stochastic Action Modifier* - SAM). O processo completo pode ser verificado na Figura 4.

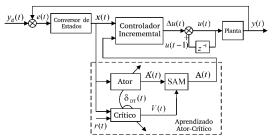

Figura 4: Processo de Controle Adaptativo com Aprendizado por Reforço

A recompensa imediata é calculada considerando-se o erro do sistema e a taxa de variação do erro simultaneamente, de forma que o seu valor é dado por

$$r(t) = \alpha r_e(t) + \beta r_{ec}(t) \quad (6)$$

Em que α e β são coeficientes que ponderam o quanto cada um desses fatores é importante para o sistema. $r_e(t)$ e $r_{et}(t)$ são definidos da seguinte maneira

$$r_e(t) = \begin{cases} 0, & |e(t)| \leq \varepsilon \\ -0.5, & c.c \end{cases} \quad (7)$$

$$r_{ec}(t) = \begin{cases} 0, & |e(t)| \leq |e(t-1)| \\ -0.5, & c.c \end{cases}$$

Em que ε é um parâmetro que define a faixa de tolerância (Wang et al 2007). Utilizou-se $\alpha = 0.4$, $\beta = 0.6$ e $\varepsilon = 0.3$.

Pela sua estrutura, uma RFBR pode ser utilizada para calcular o aprendizado do Atuador e do Crítico simultaneamente, como proposto por Wang et al (2007). A saída da camada escondida da rede é definida como segue,

$$\Phi_j(t) = exp(-\frac{||x(t) - \mu_j(t)||^2}{\sigma_j^2(t)}) \quad (8)$$

Em que os vetores $\mu_j(t)$ e $\sigma(t)$ contem os centros e as variâncias, respectivamente, das Gaussianas a serem utilizadas para a aproximação.

Dados os vetores de peso $w(t)$ e $v(t)$, para o Ator e o Crítico, respectivamente, a saída da rede é dada por

$$A'(t) = \Phi(t)w(t) \quad (9)$$

$$V(t) = \Phi(t)v(t) \quad (10)$$

O modificador de ação estocástico define o ganho do controlador a partir de

$$u(t) = e(t)(A'(t) + \sigma_v(t)) \quad (11)$$

Em que σ_v é um valor aleatório definido por

$$\sigma_v(t) = \frac{1}{1 + exp(RV(t))} \quad (12)$$

Em que R é um parâmetro arbitrário.

Para a atualização dos pesos, a técnica de aprendizado por gradiente decrescente é suficientemente eficiente. Wang et al (2007) propõe que δ_{DT} seja apenas um indicador de quão boa foi a decisão tomada pelo atuado, ou seja, é necessária a definição de um índice de performance, arbitrário. Por simplicidade, define-se o índice de performance $E(t)$ como

$$E(t) = \frac{1}{2}\delta_{dt}^2(t) \quad (13)$$

Utilizando o índice de performance escolhido como função de custo e o método do gradiente decrescente, as equações de atualização dos parâmetros resultam em

$$w_j(t+1) = w_j(t) + \eta_A \delta_{dt}(t) \frac{A(t) - A'(t)}{\sigma_v(t)} \Phi_j(t) \quad (14)$$

$$v_j(t+1) = v_j(t) + \eta_C \delta_{dt}(t) \Phi_j(t) \quad (15)$$

$$\mu_{ij}(t+1) = \mu_{ij}(t) + \eta_\mu \delta_{dt}(t) v_j(t) \Phi_j(t) \frac{x_i(t) - \mu_{ij}(t)}{\sigma_j^2(t)} \quad (16)$$

$$\sigma_j(t+1) = \sigma_j(t) + \eta_\sigma \delta_{dt}(t) v_j(t) \Phi_j(t) \frac{||x(t) - \mu_j(t)||^2}{\sigma_j^3(t)} \quad (17)$$

Em que os parâmetros η_A, η_C, η_μ, η_σ, são as taxas de aprendizado dos parâmetros da rede. As taxas de aprendizado utilizadas foram $\eta_A = 0.03$, $\eta_C = 0.1$, $\eta_\mu = 0.1$ e $\eta_\sigma = 0.05$. Esse valores foram definidos de maneira empírica.

4.3 Algoritmo

O algoritmo para a implementação segue os passos abaixo.

1. Inicializar os parâmetros: $w(0)$, $v(0)$, $\mu(0)$, $\sigma(0)$, η_A, η_C, η_μ, η_σ, α, β, γ e ε.
2. Definir valores iniciais para o Ator (consequentemente de entrada do sistema), Crítico, Recompensa e δ_{DT}.
3. Aplicar o sinal de entrada e verificar a saída do próximo estado. Calcular a Recompensa imediata $r(t+1)$.Utilizar equação (6).
4. Calcular a Saída do Ator e do Crítico. $K(t+1)$, $V(t+1)$ e Utilizar o modificador estocástico para calcular o sinal de entrada $u(t+1)$. Utilizar as equações (9), (10) e (11).
5. Calcular o δ_{DT}. Utilizar equação (1).
6. Atualizar os parâmetros da rede; μ, σ, w, v. Utilizar as equações (14), (15), (16) e (17).

5 Resultados

Foram escolhidos 2 valores de nível para o tanque 2, de forma que enquanto esse tanque mantém o nível desejado, o tanque 1 varia entre alguns níveis arbitrários.

O sinal de controle foi saturado para que fique dentro da faixa de alimentação do inversor que controla a tensão da bomba. A faixa de operação é de 0 a 10V.

O controlador proposto foi um Proporcional, Adaptativo e Incremental a ser comparado com um outro controlador PI tradicional. O sinal de referência dura 30 segundos para cada nível de avaliação no tanque 1, e dura 150 segundos para cada nível de avaliação no tanque 2.

As Figuras 5, 7 e 8 mostram, respectivamente, a performance dos dois controladores para o tanque 1, o erro do nível do tanque 1 em relação à referência e a performance dos controladores para o tanque 2. Na 6 verificam-se os sinais saturados enviados ao sistema e os sinais não saturados calculados pelos controladores.

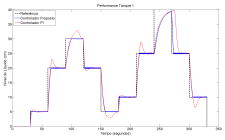

Figura 5: Performance dos Controladores para o Tanque 1

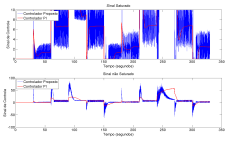

Figura 6: Sinais de Controle para o Tanque 1.

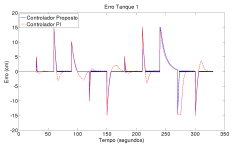

Figura 7: Erro - Tanque 1

Das Figuras 5 e 8 percebe-se que o controlador adaptativo tem performance superior no que diz respeito à velocidade de resposta e sobrepasso. Na maior parte dos pontos de operação, o controlador adaptativo proposto é tão ou mais rápido que o controlador PI, não apresentando o mesmo sobrepasso.

É possível verificar na Figura 8 que o controlador proposto é mais resistente a perturbações. Para cada variação de nível no tanque 1, o nível no

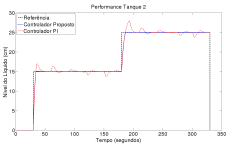

Figura 8: Performance dos Controladores para o Tanque 2

tanque 2, para o sistema com um controlador tradicional, oscila significativamente, ao passo que o sistema com o controlador proposto oscila pouco, ou não oscila.

A Figura 7 mostra que as curvas de erro se assemelham, mas o erro quadrático médio mostra uma diferença considerável entre a performance do controlador PI e do controlador adaptativo. Os erros quadráticos médios obtidos foram 13.11 e 10.83 para o controlador PI e para o controlador adaptativo, respectivamente.

É importante perceber que em diversos momentos, os dois sinais de saída são idênticos. Observando a figura 6 percebe-se que isso ocorre pelo fato de que nesses momentos o sinal de controle está saturado nos dois casos, de modo que as coincidências ocorrem, na verdade, no limite da saturação.

Para efeitos de simulação, como comentado na Subseção 4.1, considerou-se o inversor da bomba com resposta ideal, no entanto, pela Figura 6 percebe-se que o sinal de controle tem variações muito abruptas , de forma que, antes de de ser possível a continuidade do trabalho para aplicação em um sistema real, é necessária a aplicação de técnicas para suavizar o sinal de controle enviado ao inversor.

6 Conclusões

A proposta deste trabalho foi aplicar um controlador adaptativo a um sistema não linear e comparar a sua performance à de um controlador tradicional.

Devido às características de aprendizado adotadas, o controlador é capaz de obter uma resposta rápida ao se mudar a referência, mantendo um erro quadrático médio menor que o controlador tradicional. O controlador adaptativo também apresenta um comportamento mais robusto, obtendo uma boa resposta a perturbações.

Outra vantagem no uso de um controlador adaptativo é a capacidade de se ajustar ao sistema no caso de variação de parâmetros, como alteração da abertura das válvulas. O controle tradicional foi projetado para uma configuração específica das válvulas e outros parâmetros, sendo necessário o reprojeto em caso de alterações.

Um ponto importante a ser considerado na implementação do controlador proposto é a alta taxa de variação no sinal de controle, uma vez que atuadores reais possuem um tempo de resposta a excitações de entrada. Como parte da pesquisa, ainda será necessário implementar o controlador no sistema real e avaliar as características que interferem a resposta e assim realizar os ajustes necessários, caso existam, no algoritmo utilizado.

Referências

Barto, A. G., Sutton, R.S., and Anderson, C. W. (1983). *Neurolike Adaptative Elements That Can Solve Difficult Learning Control Problems.*

Deisenroth, M.P. and Rasmussen C. E. (2009). Efficient Reinforcement Learning for Motor Control

Johansson, K. H. (2000). T*he Quadruple-Tank Process: A Multivariable Laboratory Process with an Adjustable Zero.*

Lin, W. S. and Zheng, C. H. (2012). *Constrained adaptive optimal control using a reinforcement learning agent.*

Orr, M. J. L. (1996). *Introduction to Radial Basis Function Network.*

Papierok, S., Noglik, A. and Pauli, J. (2008). *Application of Reinforcement Learning in a Real Enviroment Using an RBF Network.*

Sutton, R. S. and Barto, A. G. (1988). *Reinforcemente Learning: An Introduction.*

Syafiie, S., Tadeo, F., Martinez, E. and Alvarez, T. (2011). *Model-free control based on reinforcement learning for a wastewater treatment problem.*

Syafiie, S., Tadeo, F., Villafin, M. and Alonso, A. A. (2011). *Learning control for batch thermal sterilization of canned foods.*

Wang, X., Cheng, Y and Sun, W. (2007). *A Proposal of Adaptative PID Controller Based on Reinforcement Learning.*