Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração (PROFICAM)
Escola de Minas, Universidade Federal de Ouro Preto (UFOP)
Associação Instituto Tecnológico Vale (ITV)

Dissertação

CONTROLE DE ESPESSADORES CONTÍNUOS CONVENCIONAIS VIA OTIMIZAÇÃO POR POLÍTICA PROXIMAL

Jonathan Ribeiro Silva

Ouro Preto
Minas Gerais, Brasil
2023

Jonathan Ribeiro Silva

CONTROLE DE ESPESSADORES CONTÍNUOS CONVENCIONAIS VIA OTIMIZAÇÃO POR POLÍTICA PROXIMAL

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração da Universidade Federal de Ouro Preto e do Instituto Tecnológico Vale, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Controle e Automação.

Orientador: Prof. Márcio Feliciano Braga, D.Sc.
Coorientador: Prof. Thiago Antonio Melo Euzébio, D.Sc.

Ouro Preto
2023

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

S586c Silva, Jonathan Ribeiro.
Controle de espessadores contínuos convencionais via otimização por política proximal. [manuscrito] / Jonathan Ribeiro Silva. - 2023.
58 f.: il.: color., gráf., tab..

Orientador: Prof. Dr. Márcio Feliciano Braga.
Coorientador: Prof. Dr. Thiago Antonio Melo Euzébio.
Dissertação (Mestrado Profissional). Universidade Federal de Ouro Preto. Programa de Mestrado Profissional em Instrumentação, Controle e Automação de Processos de Mineração. Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração.
Área de Concentração: Engenharia de Controle e Automação de Processos Minerais.

1. Aprendizado do computador - Aprendizado por reforço. 2. Otimização de Política Proximal (PPO). 3. Sistemas de controle - Espessadores. I. Braga, Márcio Feliciano. II. Euzébio, Thiago Antonio Melo. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 681.5:622.2

Bibliotecário(a) Responsável: Maristela Sanches Lima Mesquita - CRB-1716

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E APLICADAS
DEPARTAMENTO DE ENGENHARIA ELETRICA

FOLHA DE APROVAÇÃO

Jonathan Ribeiro Silva

Controle de espessadores contínuos convencionais via otimização por política proximal

Dissertação apresentada ao Programa de Pós-Graduação em Instrumentação, Controle e Automação de Processos de Mineração da Universidade Federal de Ouro Preto como requisito parcial para obtenção do título de mestre em Engenharia de Controle e Automação.

Aprovada em 03 de outubro de 2023.

Membros da banca

D.Sc. - Márcio Feliciano Braga - Orientador (Universidade Federal de Ouro Preto)
D.Sc. - Thiago Antonio Melo Euzébio - Coorientador (Helmholtz-Zentrum Dresden Rossendorf)
D.Sc. - Wendy Yadira Eras Herrera - (Universidade Federal de Ouro Preto)
D.Sc. - Márcio Júnior Lacerda - (Universidade Federal de São João del-Rei)

Márcio Feliciano Braga, orientador do trabalho, aprovou a versão final e autorizou seu depósito no Repositório Institucional da UFOP em 11/12/2023.

Documento assinado eletronicamente por **Marcio Feliciano Braga**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 11/12/2023, às 16:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0639886** e o código CRC **53ADBB6D**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3808-0818 - www.ufop.br

Dedico esse trabalho a Deus, a minha esposa Taynara, a meu filho Miguel, a meu pai Ronaldo e a minha mãe Vanilma.

Agradecimentos

Agradeço primeiramente a Deus por ter me capacitado e dado oportunidades ao longo da vida para estar chegando até o momento da produção deste trabalho. Agradeço a minha esposa, Taynara, pelo amor, confiança e paciência durante esse período. Agradeço aos meus pais, Ronaldo e Vanilma, por acreditarem em mim e me darem as condições para me dedicar aos estudos. Agradeço a meu irmão, Jonas, pelo companheirismo. Agradeço a todos meus familiares e amigos, por tornar o caminho mais fácil, mesmo diante dos obstáculos. Agradeço especialmente aos meus orientadores, Márcio e Thiago, por me guiarem nesta missão, pelas dicas e sugestões, e por incentivarem a não desistir.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brasil (CAPES), Código de Financiamento 001; do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq); da Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG); e da Vale SA.

“Tudo quanto fizerdes, fazei-o de todo o coração, como para o Senhor e não para homens, cientes de que recebereis do Senhor a recompensa da herança. A Cristo, o Senhor, é que estais servindo.” — Colossenses 3:23-24 ARA

Resumo

Resumo da Dissertação apresentada ao Programa de Pós Graduação em Instrumentação, Controle e Automação de Processos de Mineração como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

CONTROLE DE ESPESSADORES CONTÍNUOS CONVENCIONAIS VIA OTIMIZAÇÃO POR POLÍTICA PROXIMAL

Jonathan Ribeiro Silva

Outubro/2023

Orientadores: Márcio Feliciano Braga
Thiago Antonio Melo Euzébio

A aprendizagem por reforço é uma técnica de aprendizagem de máquina que vem sendo aplicada com grande sucesso, principalmente no contexto de jogos eletrônicos e robótica, sendo capaz de superar a capacidade humana. Apesar de seu grande potencial de aplicação nas mais diversas áreas, ainda é escasso o estudo voltado para o controle de processos. O objetivo deste trabalho é aplicar a aprendizagem por reforço, por meio do algoritmo de otimização de política proximal (PPO) em um modelo simulado de um espessador convencional, responsável pela separação sólido-líquido na mineração. O algoritmo desenvolvido atuará no controle da concentração de sólidos na saída e da altura da interface de sedimentos a partir da vazão de saída e da dosagem de floculantes. Foi possível controlar o processo por meio do PPO, porém, seu desempenho se mostrou inferior ao de técnicas de controle convencionais, a não ser no uso mais eficiente de floculante e na imunidade a distúrbios. Conclui-se que há uma necessidade de maior estudo da parametrização dos algoritmos de aprendizagem por reforço de forma a facilitar os ajustes e fazer melhor proveito de sua capacidade de abstração e generalização.

Palavras-chave: Aprendizagem por Reforço, PPO, Espessador, Controle.

Macrotema: Usina; **Linha de Pesquisa:** Análise e Projeto de Sistemas de Controle Avançado; **Tema:** Redução de Variabilidade e Melhoria de Controle; **Área Relacionada da Vale:** Pesquisa.

Abstract

Abstract of Dissertation presented to the Graduate Program on Instrumentation, Control and Automation of Mining Process as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

APPLICATION OF REINFORCEMENT LEARNING TECHNIQUES IN MINING PROCESSES

Jonathan Ribeiro Silva

October/2023

Advisors: Márcio Feliciano Braga
Thiago Antonio Melo Euzébio

Reinforcement learning is a machine learning technique that has been applied with great success, mainly in the context of electronic games and robotics, where it has been able to surpass human capacity and the performance of conventional techniques. Despite its great potential for application in the most diverse areas, studies focused on process control are still scarce. The objective of this work is to apply reinforcement learning, through the proximal policy optimization algorithm (PPO), in a simulated model of a conventional thickener, responsible for solid-liquid separation in mining. The algorithm developed will control the underflow solid volume fraction and the sediment level based on the underflow and the flocculant dosage. The simulations show that it is possible to use reinforcement learning to control the thickener process. However, the performance of PPO does not surpass conventional control techniques, except for more efficient use of flocculant and immunity to disturbances. It is concluded that there is a need for further study of the parameterization of reinforcement learning algorithms in order to facilitate adjustments and make better use of their abstraction and generalization capacity.

Keywords: Reinforcement Learning, PPO, Thickener, Control.

Macrotheme: Usine; **Research Line:** Analysis and Design of Advanced Control Systems; **Theme:** Variability Reduction and Control Improvement; **Related Area of Vale:** Research.

Lista de Figuras

Figura 2.1 Diagrama típico de tratamento de minério. 21
Figura 2.2 Esquema de um espessador contínuo. 23
Figura 2.3 Zonas de sedimentação em um espessador contínuo convencional. 24

Figura 3.1 Esquema geral do processo de decisão de Markov. 29

Figura 4.1 Comparação entre neurônio artificial típico e neurônio biológico. 34
Figura 4.2 Algoritmo de retropropagação para uma rede neural de três camadas. 36

Figura 5.1 Arquitetura Ator-Critico. 40
Figura 5.2 Funções (a) tangente hiperbólica e (b) sua derivada. 41

Figura 6.1 Sinais das malhas de controle de vazão para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow*. 47
Figura 6.2 Vazão no *underflow* ao longo de 24 *h* (a) proposto pelo controlador PPO e (b) sinal amostrado de um espessador real em funcionamento na Vale S.A. 47
Figura 6.3 Sinais das malhas de controle de vazão para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow* e considerando o sinal de controle na recompensa. 48
Figura 6.4 Sinais das malhas de controle de floculante para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow*. 49
Figura 6.5 Sinais das malhas de controle MIMO para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se os *setpoints* de densidade no *underflow* e altura da interface. 50
Figura 6.6 Recompensa do PPO para o controle MIMO variando-se *setpoints* de densidade no *underflow* e altura na interface. 51
Figura 6.7 Sinais das malhas de controle para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow*. 51

Figura 6.8 Sinais das malhas de controle para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de altura da interface. 52
Figura 6.9 Sinais das malhas de controle DMC e PI para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se a densidade da polpa de alimentação. 53
Figura 6.10 Sinais das malhas de controle PPO para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se a densidade da polpa de alimentação. Fonte: O autor. 54

Lista de Tabelas

Tabela 5.1 Parâmetros gerais do PPO utilizados nos experimentos. 44
Tabela 5.2 Parâmetros de treinamento do ator. 44
Tabela 5.3 Parâmetros de treinamento do crítico. 44

Lista de Siglas e Abreviaturas

ADAM	*Adaptive Moment Estimation*
DMC	*Dynamic Matrix Control*
GAE	*Generate Advantage Estimative*
MATLAB	*Matrix Laboratory*
MDP	*Markov Decision Process*
MIMO	*Multiple Input Multiple Output*
MPC	*Model Predictive Control*
PG	*Policy Gradient*
PI	Proporcional e Integral
PID	Proporcional, Integral e Derivativo
PPO	*Proximal Policy Optimization*
RL	*Reinforcement Learning*
SGD	*Stochastic Gradient Descent*
TRPO	*Trust Region Policy Optimization*

Lista de Símbolos

α	Taxa de aprendizagem
β	Coeficiente de maximização
β_1	Taxa de decaimento do primeiro momento
β_2	Taxa de decaimento do segundo momento
δ	Sensibilidade
$\delta(z)$	Função delta de Dirac
$\dot{F}$	Matriz diagonal dos gradientes da função de ativação
$\dot{f}$	Gradiente da função de ativação
ε	Parâmetro de ajuste de corte
γ	Fator de desconto
$\Gamma(z)$	Indica se z é um nível dentro ou fora do espessador
$\hat{A}$	Função de vantagem estimada
λ	Fator de desconto da vantagem
$\mathbb{E}$	Esperança matemática
μ	Média
∇	Gradiente
$\overline{F}$	Função de fluxo convectivo
ϕ	Concentração volumétrica de sólidos
π	Política
ρ	Restrição para área de confiança
σ	Desvio padrão
$\sum$	Somatório
θ	Parâmetros da política
υ_t	Segundo momento
υ_{hs}	Velocidade de sedimentação retardada
a	Ação
A	Área de seção transversal
a_{floc}	Ação de controle para dosagem de floculante
a_{vaz}	Ação de controle para vazão de saída
b	Vetor de limiares

B	Profundidade máxima do espessador
d	Função de compressabilidade
$D(\phi)$	Primitiva da função de compressabilidade
dt	Passo de tempo
dz	Passo de profundidade
e_{alt}	Erro percentual da altura de interface
e_{den}	Erro percentual da densidade de sólidos
f	Função de ativação
g_t	Gradiente da função de custo
H	Altura da zona de clarificação
J	Função de custo
K	Passo atual
k	Velocidade de sedimentação de uma partícula definida pela dosagem de floculante
KL	Divergência de Kullback–Leibler
L^{CLIP}	Função de perda da otimização de política proximal
L^{PG}	Função de custo do gradiente de política
M	Número de camadas
m	Camada
m_t	Primeiro momento
N	Tamanho do conjunto de treinamento
N	Número de passos
Q	Vazão volumétrica
Q^{π}	Função Q dada uma política π
r	Recompensa
$r_t(\theta)$	Razão de probabilidade
s	Estado
S^m	Número de neurônios na camada m
t	Tempo contínuo
T	Passos totais até atualização
u	Ativação interna ou entrada líquida
V^{π}	Função de valor
W	Matriz de pesos
w	Concentracão local de partículas sólidas que carregam a propriedade k
x	Vetor de entradas
y	Vetor de saídas
z	Profundidade

Sumário

1 Introdução **16**

1.1 Estado da arte 17

1.2 Objetivos 18

1.2.1 Objetivos Específicos 19

1.3 Estrutura do Texto 19

2 Tratamento de Minérios **21**

2.1 Espessadores 23

2.1.1 Modelo do Espessador 25

2.2 Conclusão 26

3 Aprendizagem por Reforço **28**

3.1 Processo de Decisão de Markov 28

3.2 Gradiente de Política 30

3.2.1 Otimização da Política por Área de Confiança 31

3.2.2 Otimização de Política Proximal 32

3.3 Conclusão 33

4 Redes Neurais Artificiais **34**

4.1 Treinamento 35

4.2 Retropropagação 35

4.3 Estimativa de Momento Adaptativo — ADAM 37

4.4 Conclusão 38

5 Detalhes da implementação do algoritmo PPO **39**

5.1 Redes Neurais 39

5.2 Função de densidade de probabilidade 41

5.3 Função de recompensa 42

5.4 Estimador da Vantagem 42

5.5 Gradiente da função de custo 43

5.6 Parâmetros utilizados 43

5.7 Conclusão 45

6 Resultados **46**

6.1 Controle pela Vazão de Saída 46

6.2 Controle pela Dosagem de Floculante 49

6.3 Controle por Vazão de Saída e Dosagem de Floculante 49

6.4 Controle com a presença de perturbações 52

6.5 Conclusão 53

7 Conclusão **55**

7.1 Propostas de Trabalhos Futuros 55

Referências Bibliográficas **57**

Apêndices **59**

1. Introdução

A mineração destaca-se como uma importante área da economia, pois tanto atividades agrícolas quanto industriais dependem de algum tipo de minério ou seus derivados (LUZ e LINS, 2018). Devido a essa demanda, torna-se necessário realizar a produção em larga escala, sendo fundamental trabalhar na eficiência dos processos de forma a evitar desperdícios e melhorar a utilização dos recursos disponíveis. Esse último ponto tem sido largamente estudado, pois o minério é um recurso não renovável. Assim, deseja-se aproveitar ao máximo o que é coletado, demandando o uso de tecnologias cada vez mais avançadas e eficientes.

O beneficiamento de minério, trata de todo processo desde a retirada de material das minas até o seu tratamento e preparação do material para comercialização. Essa atividade também está incluída no crescente processo de inovação que vem se desenvolvendo rapidamente nos últimos anos, conhecido como 4ª revolução industrial ou mesmo indústria 4.0. São várias as aplicações, associadas à indústria 4.0 (JIANG, 2017), no setor de mineração e metalurgia buscando inovação e processos mais eficientes (MESQUITA *et al.*, 2016), como exemplo, é possível encontrar implementado e consolidado o uso de robôs em minas, veículos não tripulados e mesmo veículos autônomos. Nesse contexto, uma das principais tecnologias emergentes é a Inteligência Artificial. Com a maior capacidade de processamento e comunicação dos dispositivos atuais, é possível trabalhar com grandes quantidades de dados, tornando muito complexa a tarefa de analisar tais informações. Dessa forma, se faz necessário programar computadores para que sejam capazes de fazer a análise de tais dados, tirar conclusões e até mesmo tomar suas próprias decisões. Esse é o princípio básico do aprendizado de máquina, que permite classificar informações ou mesmo reconhecer padrões não reconhecidos pelo homem através do desenvolvimento e aplicação de modelos estatísticos e matemáticos (MCCOY e AURET, 2019).

Na área de aprendizado de máquina, há grandes expectativas nas técnicas de aprendizado por reforço devido a sua flexibilidade e potencial de alcançar resultados otimizados e até superiores aos resultados de métodos convencionais. A ideia é trabalhar com uma busca empírica, ou seja, tentativa e erro, conhecendo o ambiente em que um determinado sistema está inserido e aprendendo o que é bom ou ruim a partir de suas interações com o ambiente. O método pode ser aplicado em várias áreas como robótica, saúde, finança, jogos eletrônicos, energia, processamento de linguagem natural e outros (LI, 2019; WANG *et al.*, 2020). Como características do método, podem-se citar o fato de que ele não depende do conhecimento completo do sistema em que está inserido e que apresenta ótimo desempenho, bastando uma etapa inicial de treinamento, a qual demanda maior tempo e custo computacional. Apesar de não ser impeditivo, é aconselhável que a etapa de treinamento seja feita com o auxílio de um simulador do processo, para que não seja comprometida a segurança do sistema físico durante a exploração dos estados.

Apesar do grande sucesso, as técnicas de aprendizado por reforço ainda têm sido mais empregadas no ambiente de jogos e entretenimento, e poucas e recentes são as pesquisas na área de processos industriais. O uso tardio das técnicas pode ser justificado devido a alguns desafios

impostos pelo método. Destaca-se, inicialmente, o fato da aprendizagem por reforço se basear em uma representação discreta de um sistema, com tarefas e ações discretas. Porém, em maior parte, os processos industriais são sistemas contínuos. Assim para aplicar a aprendizagem por reforço nessa situação, é necessário o uso de métodos de aproximação de funções que muitas vezes não são triviais e possuem alto custo computacional (NIAN *et al.*, 2020). A seguir, é desejada uma boa representação do sistema por um simulador, permitindo contornar o longo tempo de treinamento inicial e a insegurança operacional de se tomar ações por tentativa e erro diretamente do processo real. Outro desafio associado ao emprego do aprendizagem de máquina em controle é a falta de clareza para definir uma função objetivo, sinais de sensores serem ruidosos e incertos e restrições dinâmicas das ações de controle (HALLÉN *et al.*, 2019). Em virtude disso, é compreensível que a aplicação das técnicas de aprendizado em processos contínuos de fato só começou a ganhar atenção nos últimos anos com a popularização de redes neurais para a representação dos estados e ações.

1.1. Estado da arte

Ao contrário dos controladores clássicos que necessitam de etapas complexas para seu desenvolvimento, como uma cuidadosa análise da dinâmica do sistema e a obtenção de modelos matemáticos abstratos, os algoritmos de Aprendizagem por Reforço (RL, do inglês *Reinforcement Learning*) podem trabalhar sem conhecimentos específicos a respeito do sistema no qual será inserido (SPIELBERG *et al.*, 2017). Essa característica instigou o uso de RL no controle de processos dinâmicos que vem ganhando espaço nos últimos anos. Um levantamento das dificuldades, avanços e principais aplicações nesta área pode ser encontrado em Nian *et al.* (2020) e Shin *et al.* (2019).

Tratando especificamente de beneficiamento mineral, tais estudos são ainda mais escassos, com o trabalho de Hallén *et al.* (2019) tendo destaque ao demonstrar a eficiência da aprendizagem de máquina aplicada na área. No trabalho, foi realizada a integração de um modelo de circuito de moagem simulado, em conjunto com um *framework* de código aberto que implementa diversos algoritmos de RL. Dessa forma, utilizou-se o algoritmo de Otimização de Política Proximal (PPO, do inglês *Proximal Policy Optimization*) para substituir uma malha de controle PID (Proporcional, Integral, Derivativo) presente. Foi demonstrado que o uso de RL permite controlar o sistema com as mesmas restrições de processos existentes. Além disso, por permitir definir um objetivo com alto nível de abstração, foi possível manter o controle com condições operacionais mais rentáveis ao se relacionar a saída do sistema com o custo financeiro. Conclui-se a partir desse trabalho que RL é uma alternativa para técnicas de controle convencionais. Devido à sua característica de caixa preta, isto é, trabalhar com sistemas sem o conhecimento das equações que regem o processo, pode ser uma boa escolha para sistemas que apresentam não linearidades, dinâmicas estocásticas ou alto nível de incertezas (BUŞONIU *et al.*, 2018).

Recentemente, outros autores têm utilizado abordagens diferentes para aplicar a aprendizagem por reforço no controle de processos da mineração. Jiang *et al.* (2018) utilizou uma arquitetura ator-crítico para a otimização de um processo de flotação industrial. Para isso, foi utilizado um controlador PI (Proporcional Integral) multivariável para manter a referência (*set-point*) do nível da polpa e do fluxo de alimentação. O algoritmo de reforço foi responsável por controlar toda a etapa de flotação, a partir da obtenção de um modelo parcialmente desconhecido do processo, em malha fechada. Este controlador geral foi composto por duas redes neurais, com os papéis de ator e crítico, buscando otimizar as condições de operação do processo, por meio da escolha adequada dos *setpoints*. O método proposto se mostrou satisfatório, permitindo otimizar os índices operacionais. Reyes *et al.* (2020) utilizou RL para o controle de um circuito de moagem com moinhos semiautógenos. Foi utilizado um modelo simulado desenvolvido em *MATLAB/Simulink*, no qual o controle foi feito primariamente por um controlador preditivo baseado em modelo (MPC, do inglês *Model Predictive Control*) e, posteriormente, empregando a *toolbox* de RL para otimizar o *setpoint*. Apesar de mostrar a melhora do desempenho do MPC, servindo como um controlador supervisório ótimo, o trabalho não implementou a aprendizagem por reforço como controlador principal. Por fim, em Shipman (2021), utiliza-se o aprendizado por reforço profundo para ajustar os parâmetros de um controlador PI de forma contínua. O algoritmo PPO é empregado para criar um agente genérico que atenda diferentes plantas. Os testes foram realizados em duas malhas de vazão do processo de moagem, em um circuito de cominuição industrial. Foi alcançada a melhora do desempenho da malha que controla a vazão de água de entrada do moinho. No entanto, houve piora no desempenho da malha de vazão de água de alimentação do reservatório de descarga. O trabalho concluiu que, apesar de haver algumas deficiências na aprendizagem por reforço, ainda é um método promissor para ajustar malhas de controle de forma contínua e automática.

1.2. Objetivos

O objetivo deste trabalho, portanto, é aplicar uma técnica avançada de controle por aprendizagem por reforço em um processo real da mineração. Dessa forma, o algoritmo de PPO será utilizado por ser um método voltado para processos contínuos e que apresenta bom desempenho, mesmo sendo mais simples do que algoritmos equivalentes (SCHULMAN *et al.*, 2017). O processo escolhido para ser controlado é o espessamento, etapa responsável por retirar o excesso de água que foi adicionada em processos anteriores, permitindo a reutilização dessa água e facilitando a disposição de rejeitos e o tratamento do minério em etapas posteriores. Os espessadores são equipamentos de grande porte de complexa modelagem, sendo assim, há amplo potencial para desenvolvimento de técnicas que sejam mais eficientes do que as técnicas aplicadas tradicionalmente, as quais demandam grandes simplificações do sistema. Normalmente são desconsideradas interferências importantes como a adição de floculantes, a ação do *rake* (rastelo ou ancinho) e o perfil de sedimentação dos sólidos.

Para a implementação da técnica, será usado o modelo de espessador proposto em Pereira (2021). O modelo representa adequadamente o sistema, pois considera o formato cônico-cilíndrico dos espessadores convencionais levando em conta o uso de floculantes. Por meio de um simulador implementado com o *software MATLAB*, será implementado o algoritmo PPO como controlador do sistema. O objetivo é ajustar os parâmetros do algoritmo e realizar a etapa de treinamento *offline* do agente de aprendizado por reforço. Desta forma, para uma posterior aplicação no sistema real, o algoritmo terá explorado o espaço de estados o suficiente para realizar o controle a partir de um bom estado de funcionamento, sendo necessário esforço apenas para o ajuste e a manutenção de um controle ótimo.

1.2.1. Objetivos Específicos

O presente trabalho busca demonstrar a eficiência de métodos de aprendizagem de reforço para o controle de processos contínuos, abrindo caminho para que mais estudos sejam implementados para outros processos da mineração e mesmo em outras áreas da indústria. Os objetivos específicos são apresentados a seguir:

- Desenvolver uma forma eficaz de controle do espessador a partir do uso de aprendizagem por reforço
- Economizar o uso de floculante
- Melhorar o material de saída
- Aumentar o volume de água recuperada no *overflow*
- Diminuir paradas e situações de transbordo devido a solidificação excessiva de material
- Demonstrar a eficiência da técnica PPO
- Comparar o controle utilizando RL com o de técnicas de controle clássicas

1.3. Estrutura do Texto

Este trabalho começa com uma breve introdução a respeito da mineração e da aprendizagem por reforço, além de apresentar uma contextualização sobre os trabalhos correlatos às duas áreas. Em seguida, definem-se os objetivos gerais e específicos.

Além deste capítulo introdutório, estrutura-se o trabalho da seguinte forma:

- **Capítulo 2**: apresenta-se uma introdução a respeito do tratamento de minérios e seus processos, entrando em mais detalhes no funcionamento e características dos espessadores. Ainda nesse capítulo, é apresentado o modelo do espessador no qual serão aplicadas as técnicas de controle.

- **Capítulo 3**: introduz a aprendizagem por reforço e o processo de decisão de Markov. Dentro desse tema, há a explicação de como funcionam as técnicas de gradiente de política e suas vantagens, apresentando a otimização da política por área de confiança e a otimização de política proximal.
- **Capítulo 4**: dedica-se à definição de redes neurais, sua utilização e formas de treinamento. É apresentado o algoritmo de retropropagação e o método ADAM.
- **Capítulo 5**: detalha a estrutura do algoritmo e as decisões de implementação, destacando as redes neurais utilizadas, a função de densidade de probabilidade, o estimador de vantagem e a função de recompensa utilizada. Contém ainda o desenvolvimento matemático do gradiente da função de custo e por fim apresenta os parâmetros que serão utilizados.
- **Capítulo 6**: apresenta e discute os testes realizados e os resultados encontrados.
- **Capítulo 7**: pontua as considerações finais e as propostas para trabalhos futuros.

2. Tratamento de Minérios

Na mineração, o tratamento de minérios é o conjunto de operações que tem como objetivo alterar a forma, granulometria ou concentração de matérias de uma determinada substância mineral (LUZ e LINS, 2018). Essa substância é uma matéria-prima natural, não renovável, encontrada de forma escassa na natureza. Devido a isso, é necessário que sua extração e utilização aconteça da forma mais eficiente possível, evitando desperdícios. Infelizmente, a indústria de mineração é uma das indústrias que possuem as maiores perdas no processo. Assim, para contornar esses problemas, é fundamental empregar meios sofisticados e equipamentos bem ajustados.

O minério coletado na natureza passa por diversas etapas para separação da ganga, material sem valor econômico. Um fluxograma com as principais etapas normalmente encontradas no tratamento de minérios pode ser visto na Figura 2.1.

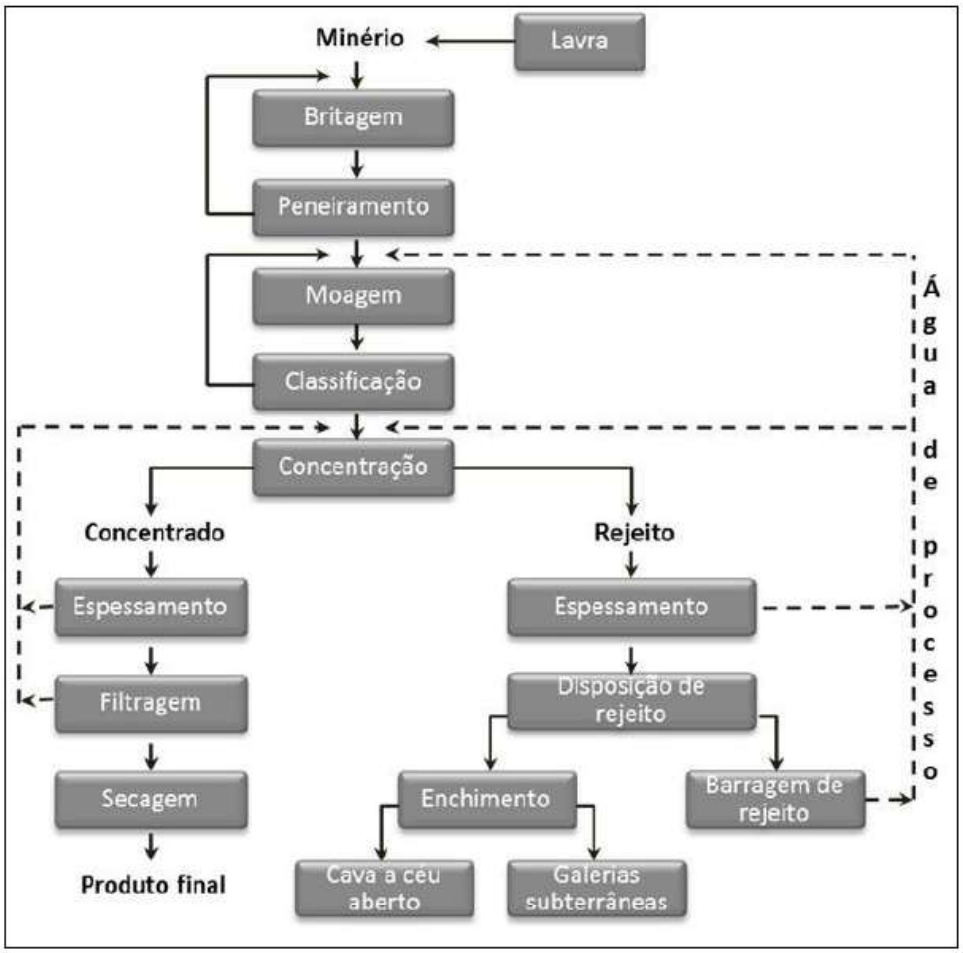


Figura 2.1: Diagrama típico de tratamento de minério.
Fonte: Luz e Lins (2018).

A etapa de lavra é o processo inicial e consiste na retirada do mineral do solo, normalmente por uso de explosivos, perfuratriz, escavadeira ou equipamentos equivalentes. O material coletado é a seguir levado para a britagem, onde o minério tem sua granulometria reduzida.

Usualmente, a britagem se divide em 3 ou 4 etapas, que se diferem pela granulometria do material de entrada e de saída. A ideia é reduzi-lo para ser trabalhado nas etapas subsequentes. Após a britagem, o material é levado ao peneiramento para separação do material que está na granulometria desejada e o restante é levado de volta para a britagem para ser reprocessado.

Para a boa utilização do material coletado, é utilizada mais uma etapa de redução da geometria, a moagem. Até esse momento, o minério é usualmente transportado entre os processos por meio de correias transportadoras. A moagem tem papel semelhante ao realizado pelo britador, ou seja, diminuir a granulometria do material na saída, sendo que a diferença entre eles é devido ao processo usado para isso. O foco são materiais finos, sendo muito importante o bom dimensionamento, pois materiais com grande granulometria não liberam de forma adequada o material útil, e materiais muito pequenos aumentariam os custos e perdas no processo de concentração (LUZ e LINS, 2018). Na mineração, o mais comum é fazer a moagem a úmido com a adição de água, o que implica em uma maior capacidade de processamento, contudo também aumentando o consumo de energia. Destaca-se ainda que, para os próximos processos, o transporte do material pode ser feito por meio de tubulações e calhas, trazendo uma alternativa para o uso de correias transportadoras.

Após uma etapa de classificação, o material que ainda não atingiu a granulometria desejada volta para os moinhos, enquanto que o material com granulometria adequada passa para a concentração. O meio mais utilizado de concentração é a flotação. Nesse processo, o minério é concentrado e separado dos outros materiais por meio da formação de espuma com a ajuda de reagentes. O minério concentrado e o rejeito do processo são então levados para a etapa de sedimentação. Nesse ponto, o material já possui a geometria desejada, porém, devido à adição de água para os últimos processos, são necessárias algumas etapas para a retirada do excesso de líquidos.

A primeira etapa responsável por separar o material sólido do líquido é realizada pelos espessadores. Esses equipamentos fazem uso de sedimentação por gravidade, onde o material sólido, mais denso, desce e se concentra no fundo do espessador e, então, é retirado do equipamento pela abertura de uma descarga. A água transborda em canaletas na parte superior a medida que a vazão de entrada é maior do que a vazão de saída. Com o intuito de acelerar o processo de sedimentação, empregam-se floculantes que têm a função de reunir o material sólido aumentando a sua densidade. Tanto o minério concentrado quanto o rejeito passam por essa etapa, o que ocorre de forma paralela.

O material concentrado ainda passa por mais duas etapas de retirada de umidade, que são a filtragem e a secagem. Ao fim do processo, o produto final estará pronto para ser armazenado, transportado ou utilizado. Enquanto que o rejeito, seguindo o fluxograma (Figura 2.1), após o espessamento, passa diretamente para a disposição em barragens, cavas a céu aberto ou galerias subterrâneas. Como o rejeito não tem valor comercial, ele não passa pela filtragem e secagem que são processos custosos. Atualmente, essa realidade tem se alterado, com projetos de filtragem de rejeitos para armazenamento a seco e recuperação de material (BRAGA *et al.*, 2019;

DE ÁVILA FONSECA *et al.*, 2019). Esses projetos acompanham a crescente preocupação com as questões ambientais e de segurança, permitindo a descaracterização de barragens existentes e evitando a formação de novas barragens. É importante ressaltar que, em alguns processos, incluindo o espessamento, há a reutilização da água retirada do material úmido, sendo reaproveitada para os processos de moagem e de concentração.

Uma vez que o foco deste trabalho envolve o controle de espessadores, ênfase maior é dada a esse equipamento nas seções seguintes.

2.1. Espessadores

A etapa de separação sólido-líquido é muito importante na indústria mineral, pois geralmente utiliza água em algum de seus processos. Os espessadores, principal equipamento que realiza a tarefa, é um equipamento de alto porte e alto custo. Dessa forma, qualquer melhoria traz um grande impacto na produção ou nos seus custos. Independente de qual material será separado, o espessamento tem dois objetivos básicos que são: (i) a produção de um sólido desaguado e; (ii) a produção de um líquido clarificado (LUZ e LINS, 2018).

As características da polpa que será tratada no espessador possuem grande influência no processo, como exemplo, pode-se citar o efeito de dispersão sofrido por partículas muito finas, dificultado a sedimentação. Nesses casos, é muito importante o uso de floculantes ou coagulantes para forçar a união das partículas e diminuir o efeito de dispersão. Outro problema importante, muito comum em polpas de minerais mais densos, é a aglomeração no fundo do espessador. Para evitá-lo, os espessadores comuns possuem um braço, conhecido como *rake* ou ancinho, que gira continuamente no interior do equipamento, carreando os sólidos aglomerados em direção a descarga de saída, como visto na Figura 2.2.

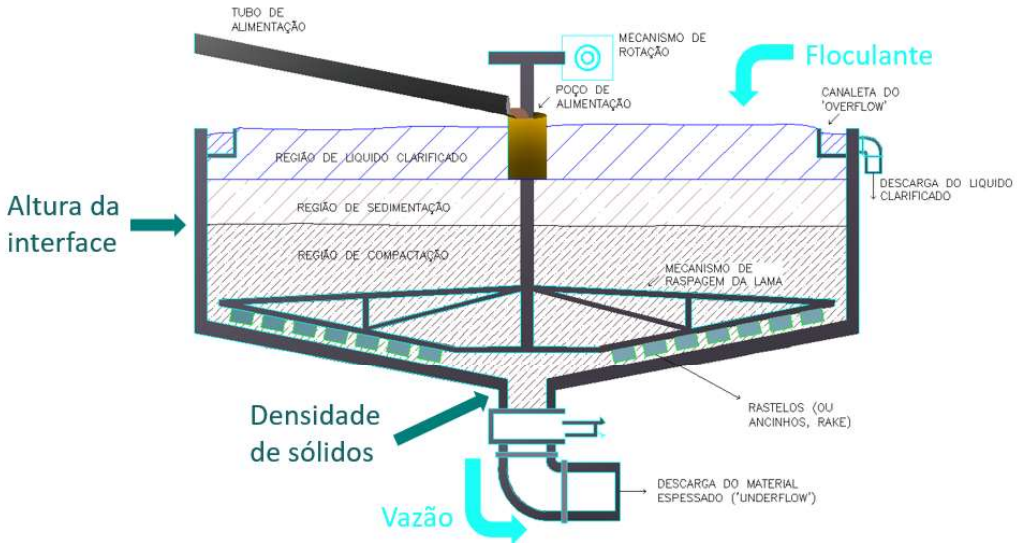


Figura 2.2: Esquema de um espessador contínuo.
Fonte: O autor.

O líquido clarificado é retirada do processo ao transbordar para uma canaleta localizada na parte superior do tanque, esse fluxo de saída é chamado de *overflow*, efluente ou sobrenadante. O material sedimentado, ou *underflow*, é retirado pela descarga, muitas vezes com o apoio de uma bomba na tubulação de saída. É importante ressaltar que o acúmulo de material sólido em excesso no fundo de espessadores pode levar ao travamento do *rake*. Sistemas atuais possuem como forma de segurança a leitura do torque do *rake*. Caso o torque se torne muito alto, os braços são automaticamente erguidos facilitando a raspagem. Para melhor direcionamento do material sólido até a saída do sistema, os espessadores convencionais possuem a parte inferior, na região de compactação, em formato cônico. Dessa forma, o material é puxado para o centro, onde há o escoamento do *underflow* pela própria força da gravidade, além de contar com o auxílio do *rake* no processo, facilitando o posterior bombeamento.

Durante o processo de sedimentação a separação do material no espessador forma zonas distintas como visto na Figura 2.3.

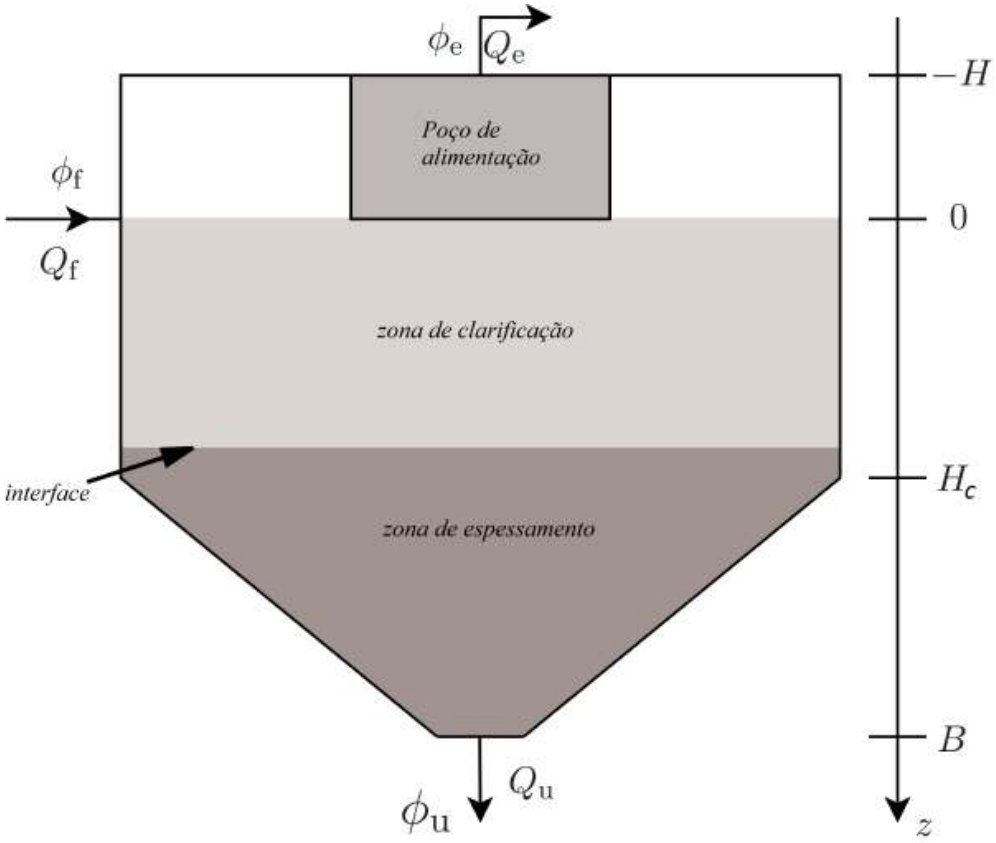


Figura 2.3: Zonas de sedimentação em um espessador contínuo convencional.
Fonte: Pereira *et al.* (2020).

A zona de espessamento é a área onde há alta concentração de sólidos em sedimentação. A compressão do material expulsa a água presente criando a zona de clarificação na parte superior do tanque. O fluxo do material é indicado pelas setas onde Q representa a vazão e ϕ a concentração volumétrica de sólidos. Os índices f, e e u são respectivamente a entrada, *overflow* e *underflow*. O ponto de referência inicial, onde a altura é considerada 0, é o local onde

há entrada do material. A altura é crescente em direção a saída, sendo $-H < z \leq 0$ a área onde subtrai-se o diâmetro do poço de alimentação e há presença predominante de água. A profundidade máxima do espessador é indicada por B e H_c indica o ponto onde o espessador deixa de ser cilíndrico e passa a ser cônico. A altura da interface é o ponto de separação entre a zona de clarificação e a zona de espessamento. Quanto maior a altura de interface, maior a zona de clarificação, portanto maior é a separação sólido-líquida. Em relação aos custos diretos do processo, destaca-se o valor do insumo utilizado como floculante. A água que pode ser recuperada, representa uma economia não só financeira como também ambiental. Já a variação das vazões ocorre sem grandes demandas, pois não apresenta grande esforço para as bombas utilizadas. Energeticamente, destaca-se o esforço para movimentação do *rake*, mas esse não será um ponto de estudo do trabalho. Os maiores custos do processo seriam indiretos, incluindo benefícios para o tratamento do minério e do rejeito desaguados nas etapas posteriores ao espessamento, além do alto custo que seria demandado em caso de solidificação excessiva do material ou de um possível transbordo de sólidos.

2.1.1. Modelo do Espessador

O presente trabalho será desenvolvido com base em um espessador convencional cilíndrico-cônico. O modelo utilizado é o desenvolvido em Pereira (2021). O trabalho partiu da modelagem de um sistema de espessamento considerando a adição do floculante e acrescentou a variação de área que parte do formato cilíndrico para o cônico. Com isso, chegou-se a um modelo que se aproxima da geometria real do espessador levando em consideração seus principais processos e este foi implementado em um simulador.

Para realizar a modelagem utilizando a adição de floculantes, deve-se considerar que algumas partículas sofrerão uma variação de velocidade k de acordo com o tempo. Essa propriedade varia de 0, quando a velocidade é nula, até 1, situação em que a velocidade é a maior possível. A concentração local das partículas sólidas sob efeito do floculante será denotada por w. Considerando d como a função de compressabilidade dos sólidos em alta concentração, $D(\phi)$ é sua função primitiva definido como

$$D(\phi) = \int_0^{\phi} d(s)ds. \tag{2.1}$$

O modelo do espessador representado pelo sistema de equações diferenciais parciais é formado por (2.2) e (2.3) e expressa a concentração total de sólidos ϕ e a concentração w em função da profundidade z do espessador e do tempo T, considerando a variação da área A. O índice f representa as variáveis da alimentação, $\Gamma(z)$ indica se z é um nível dentro ($\gamma(z) = 1$) ou

fora ($\gamma(z) = 0$) do espessador e $\delta(z)$ é a função delta de Dirac.

$$\frac{\partial(A(z)\phi)}{\partial t} + \frac{\partial}{\partial z}A(z)\overline{F}(\phi,w/\phi,z,t) = \frac{\partial}{\partial z}\left(A(z)\gamma(z)\frac{w}{\phi}\frac{\partial D(\phi)}{\partial z}\right) + Q_f(t)\phi_f(t)\delta(z), \quad (2.2)$$

$$\frac{\partial(A(z)w)}{\partial t} + \frac{\partial}{\partial z}\left(\frac{w}{\phi}A(z)\overline{F}(\phi,k,z,t)\right) = \frac{\partial}{\partial z}\left(A(z)\gamma(z)\frac{w^2}{\phi^2}\frac{\partial D(\phi)}{\partial z}\right) + Q_f(t)w_f(t)\delta(z), \quad (2.3)$$

em que $\overline{F}(\phi,k,z,t)$ é a função de fluxo convectivo que varia com a profundidade do tanque:

$$\overline{F}(\phi,k,z,t) = \begin{cases} Q(t)\phi/A(z), & \text{para} \quad z < -H, \\ Q(t)\phi/A(z) + k\upsilon_{hs}(\phi)\phi, & \text{para} \quad -H < z < B, \\ Q(t)\phi/A(z), & \text{para} \quad z > B. \end{cases} \quad (2.4)$$

A variável υ_{hs} é a velocidade de sedimentação retardada das partículas. A vazão volumétrica $Q(t)$ é dada em função da vazão no efluente $Q_e(t)$ e da vazão do *underflow* $Q_u(t)$, de modo que

$$Q(t) = \begin{cases} -Q_e(t), & \text{para} \quad z < 0, \\ Q_u(t), & \text{para} \quad z > 0. \end{cases} \quad (2.5)$$

O simulador de espessadores criado a partir desta modelagem foi validado a partir de dados reais de um espessador existente em planta da Vale S.A. em Carajás - PA. A implementação em MATLAB foi realizada utilizando um procedimento numérico para fazer a discretização no espaço. Maiores detalhes da implementação e a validação do modelo podem ser encontrados no trabalho de Pereira (2021). Apesar de ser uma boa representação do espessador, algumas aproximações naturalmente foram feitas durante a modelagem. Para o equacionamento realizado, foram considerados uniformes a sedimentação e concentração de sólidos, a suspensão como ideal e a influência do trabalho do *rake* foi descartada. Devido a isso, a construção de um controlador baseado no modelo não é perfeito. Isso indica a necessidade de aplicação de alguma forma de controle adaptativo para contornar as aproximações ou, como será feito, a aplicação de aprendizagem de máquina para construir um controlador que, empiricamente, decidirá as ações de controle ideais para o sistema mesmo sem o conhecimento perfeito de sua dinâmica. Com a abordagem de aprendizagem por reforço, o modelo simulado será uma boa base de treinamento e permitirá encontrar um controlador que idealmente possa contornar as imperfeições do modelo ao ser executado *online*.

2.2. Conclusão

Neste capítulo foi realizada uma introdução a área de tratamento de minério com um maior foco na etapa de espessamento. É apresentado também o modelo matemático de um espessador convencional que foi utilizado para gerar o simulador no qual será implementada

a técnica de aprendizagem por reforço. A partir do modelo matemático, é possível entender a influência das variáveis controladas, densidade de sólidos e altura da interface, e as variáveis de controle, vazão no *underflow* e dosagem de floculante, que serão utilizadas.

3. Aprendizagem por Reforço

A inteligência artificial é uma área que vem se desenvolvendo de forma crescente nos últimos anos como solução para vários problemas de tecnologia. Nesse contexto, destaca-se o aprendizado de máquina. A capacidade de uma máquina reconhecer padrões a partir de grande quantidade de dados é uma importante ferramenta de análise e pode ser usada nas mais diversas áreas, como saúde (YU *et al.*, 2021), energia (PERERA e KAMALARUBAN, 2021), robótica (DALAL *et al.*, 2021), economia (CHARPENTIER *et al.*, 2021), veículos autônomos (KIRAN *et al.*, 2021), entre outras.

No entanto, hoje não basta o computador conseguir realizar tarefas de inteligência. Com a tecnologia atual, os computadores são capazes de aprender por conta própria e até superar o conhecimento humano. Esse é o papel da Aprendizagem por Reforço, um subtópico da aprendizagem por máquina.

Em Sutton e Barto (2018), é feita a comparação da aprendizagem por reforço com a forma que as pessoas aprendem. Inicialmente, o ser humano não tem nenhum conhecimento do ambiente e chega ao mundo sem uma orientação, aprendendo por conta própria a ouvir, sentir e interpretar os sinais que recebe. Desde o nascimento, tem-se a interação com o ambiente ao nosso redor, e as respostas que recebe o leva aos primeiros aprendizados. E assim segue-se por toda a vida, toma ações nos ambientes em que está inserido e espera a recompensa, seja positiva ou negativa, e, a partir dessa experiência adquirida, busca maximizar as respostas positivas. Vez ou outra ainda arrisca novas experiências, para buscar recompensas ainda maiores do que as que está acostumado, ou então para descartar essas opções.

A aprendizagem por reforço, como técnica de inteligência artificial, nada mais é do que uma forma de realizar computacionalmente o aprendizado a partir de interações, por tentativa e erro. Nos próximos tópicos será detalhado a base utilizada para implementação do PPO. Inicialmente será abordado o processo de decisão de Markov que define a estrutura do algoritmo. Em sequência há um tópico a respeito do gradiente de política, que explica como que o algoritmo atualiza e adquire conhecimento. Ainda sobre o gradiente de política, é dado foco a otimização da política por área de confiança que traz um conceito de limitação dos gradientes de forma a alcançar atualizações estáveis. Por fim, será tratado a otimização de política proximal, técnica que aprimora os conceitos existentes buscando simplificar a sua aplicação, sendo essa a ferramenta que será utilizada.

3.1. Processo de Decisão de Markov

Para se trabalhar com RL, os sistemas são descritos como processos de decisão de Markov (MDP, do inglês *Markov Decision Process*). O MDP é uma formulação matemática para tomada de decisões sequenciais (HALLÉN *et al.*, 2019). Consiste em um agente, que é o responsável por tomar as decisões, e o ambiente, que é tudo que é externo ao agente e que não é

controlado. A Figura 3.1 contém uma representação esquemática do MDP.

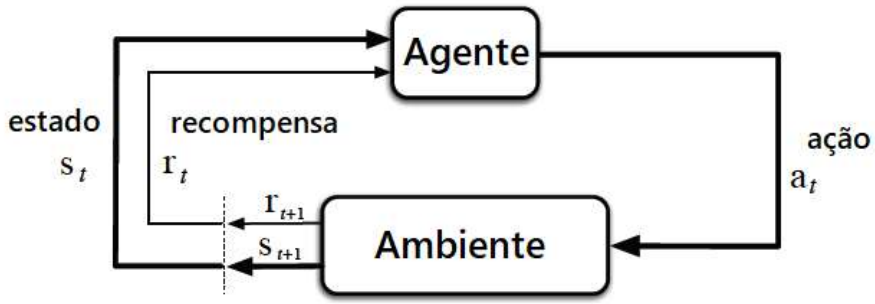


Figura 3.1: Esquema geral do processo de decisão de Markov.
Fonte: Sutton e Barto (2018).

O agente, inicialmente em um estado s_t, toma uma ação a_t que o leva a um novo estado s_{t+1} e gera uma recompensa r_{t+1}. Para cada estado, há um conjunto de ações possíveis que o agente pode tomar para tentar alterar o sistema. O objetivo com o MDP é fazer uma escolha adequada de ações para maximizar as recompensas geradas ao longo do tempo (SUTTON e BARTO, 2018). Tais recompensas são definidas de acordo com o sistema trabalhado e o propósito desejado. Para aplicações em controle de processos, uma boa escolha é construir a função de recompensa a partir do erro e das ações de controle. Outros parâmetros de desempenho como máximo sobressinal e tempo de estabilização também podem ser utilizados. Até mesmo parâmetros indiretos e de alto nível podem ser adicionados, como o custo financeiro envolvido, contanto que seja possível construir uma representação matemática. Quem determina qual ação será tomada a partir de um determinado estado é a política, representada por π. Para sistemas contínuos, π normalmente é definida como uma distribuição de probabilidade que varia para cada ação no estado atual.

Os estados podem ser entendidos como o retorno de um conjunto de sensores, no caso do espessador, a leitura de densidade na saída do espessador e o medidor de nível da interface. As ações seriam os sinais de controle dos atuadores, que seriam as bombas que variam a vazão de saída do espessador e a vazão de floculante que é adicionado no tanque para dosagem. A recompensa representa alguma formulação matemática, ou mesmo o retorno de um conjunto de sensoriamento, algo que seja capaz de medir a qualidade do processo. Como será detalhado posteriormente, nesse trabalho a recompensa utilizada terá como base o erro das variáveis controladas em relação as referências desejadas.

É natural pensar que dado dois estados diferentes, os possíveis novos estados alcançáveis a partir deles podem ser diferentes. Desta forma, o resultado final, após uma sequência de ações, não está diretamente ligado à maior recompensa no estado inicial. É necessário levar em consideração o caminho que foi percorrido e os estados que se tornaram disponíveis. Usando o controle de processos como exemplo, um sobressinal temporário pode levar a uma estabilização mais rápida do sistema. Desta forma, para a tomada de decisão e a depender do caso, é importante considerar não só a recompensa imediata, mas também as recompensas fu-

turas (RAVICHANDIRAN, 2018). Em MDP, isso é feito por meio da função de valor dada por

$$V^{\pi}(s) = \mathbb{E}_{\pi}\left[\sum_{K=0}^{\infty} \gamma^{K} r_{t+1+K} \middle| s_t = s\right]. \tag{3.1}$$

A função de valor define o quão bom é estar no estado s em um sistema regido pela política π. Considerando $\mathbb{E}$ a esperança matemática, V^{π} leva em consideração o valor esperado do somatório das recompensas futuras r_{t+1}, caso se mantenha a política atual. K é a contagem de passos que o algoritmo repete a tomada de ações e leitura das recompensas. γ é um fator de desconto entre 0 e 1, responsável por diminuir as parcelas de recompensa conforme o tempo passa, visto que o erro da estimativa dos valores tende a aumentar. Esse fator também é fundamental em sistemas ilimitados no tempo pois garante que $V^{\pi}(s)$ não seja infinito fazendo as recompensas tenderem a zero conforme K cresce (NIAN *et al.*, 2020).

Ocasionalmente, a função de valor pode ser relativa ao par estado-ação e não somente ao estado atual. Nesse caso, ela é conhecida como função de ação-valor e representa o quão bom é tomar a ação a partindo do estado s por meio de uma política π

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}\left[\sum_{K=0}^{\infty} \gamma^{K} r_{t+1+K} \middle| s_t = s, a_t = a\right]. \tag{3.2}$$

Na função de ação-valor, o peso de cada ação é enviesado pelo estado em que a ação é tomada, assim, ações ruins realizadas em bons estados podem retornar maior valor do que ações ótimas realizadas em estados ruins. Para contornar esse problema, subtrai-se o valor dos estados do par estado-ação

$$A(s,a) = Q^{\pi}(s,a) - V^{\pi}(s). \tag{3.3}$$

A função (3.3) é chamada de vantagem pois representa quão vantajoso é escolher uma determinada ação em relação as outras possíveis ações naquele estado. Valores positivos indicam ações que levam a estados que possuem maiores recompensas, enquanto valores negativos indicam ações que levam a estados que alcançam recompensas menores.

3.2. Gradiente de Política

Os métodos tradicionais de RL trabalham atualizando o valor das ações, nesses casos a política depende de V^{π}, pois as ações são determinadas baseado no valor que receberam. Por outro lado, os métodos de gradiente de política (PG, do inglês *Policy Gradient*) focam em atualizar a própria política, e não dependem de V^{π} para definir qual ação será tomada, portanto são independentes. Esses métodos ainda podem ter uma função de valor ou de ação-valor, mas, nesse caso, esta função não é o objetivo de maximização, ela é utilizada apenas para guiar

a atualização dos parâmetros da política, julgando se é uma política boa ou ruim (SUTTON e BARTO, 2018).

A grande vantagem dos métodos PG é que podem trabalhar com ações e estados contínuos, para isso são utilizadas redes neurais artificiais (RAVICHANDIRAN, 2018). Na atualização da política, a rede recebe como entrada os estados, enquanto as saídas são as probabilidades de cada ação possível para esse estado. As probabilidades são usadas para escolher qual ação o sistema irá executar, mesmo que não seja a ação ideal, a recompensa é armazenada e forma, após vários passos, um conjunto de dados de treinamento. Com base nesses dados, os pesos da rede neural são atualizados a partir do gradiente do resultado fazendo com que as ações com maior recompensa terão maior probabilidade naquele estado e, de forma análoga, as ações com menor recompensa terão menor probabilidade.

3.2.1. Otimização da Política por Área de Confiança

A aprendizagem por reforço ocorre por tentativa e erro, isso quer dizer que, em alguns momentos, o agente irá explorar ações ruins para o sistema e que trarão recompensas negativas. Devido a isso, foram criadas estratégias de otimização da política com restrição. A ideia é limitar a atualização da política para que ela não vá para uma área insegura onde a probabilidade de tomar ações ruins seria muito alta (RAVICHANDIRAN, 2018).

O método de otimização da política por área de confiança (TRPO, do inglês *Trust Region Policy Optimization*) impõe uma restrição ρ, de quanto a política nova pode se distanciar da política antiga. Se a probabilidade de acontecer algo indesejado for menor do que esse limite ρ, o sistema está seguro. O método TRPO busca manter o sistema nessa região de segurança maximizando a função objetivo

$$\hat{\mathbb{E}}_t \left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{anterior}}(a_t|s_t)} \hat{A}_t \right], \tag{3.4}$$

sujeito a restrição

$$\hat{\mathbb{E}}_t \left[KL[\pi_{\theta_{anterior}}(.|s_t), \pi_\theta(.|s_t)] \right] \leq \delta, \tag{3.5}$$

em que *KL* é a divergência de Kullback–Leibler e retorna o quanto duas distribuições de probabilidade são diferentes. O termo $\pi_\theta(a_t|s_t)$ é a probabilidade da ação a_t ser escolhida no estado s_t quando a política for definida pelos parâmetros *theta*. Note que (3.5) é a inequação responsável por restringir a nova política π_θ, mantendo-a próxima da política anterior $\pi_{\theta_{anterior}}$. Substituindo essa restrição por uma penalidade em (3.4), o objetivo do TRPO passa a ser a resolução do problema de otimização do coeficiente β que maximize

$$\hat{\mathbb{E}}_t \left[\frac{\pi_{\theta(a_t|s_t)}}{\pi_{\theta_{anterior}}(a_t|s_t)} \hat{A}_t - \beta KL[\pi_{\theta_{anterior}}(.|s_t), \pi_\theta(.|s_t)] \right]. \tag{3.6}$$

Com essa limitação na atualização da política, evita-se tomar passos muito grandes. Caso a política nova se afastasse da antiga, ela mudaria completamente as ações tomadas e o

comportamento do sistema, o que poderia levá-lo para uma região insegura. O TRPO auxilia para que a evolução do aprendizado ocorra de forma progressiva e estável.

3.2.2. Otimização de Política Proximal

O problema do método TRPO é a dificuldade em definir um valor para β que seja satisfatório em todo o tempo. Além disso, trabalhar com (3.4) e (3.5) computacionalmente teria um custo muito alto. O trabalho em Schulman *et al.* (2017) criou uma alternativa mais simples substituindo a restrição (3.5) por uma penalidade na própria função objetivo. Tal solução se mostrou tão eficiente quanto a TRPO e é chamada de Otimização de Política Proximal. A função de custo do PPO é

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t) \right], \tag{3.7}$$

em que $\hat{A}$ é uma estimativa de (3.3), ε é o limite definido para região de confiança. $clip$ e min são funções. $\hat{\mathbb{E}}_t$ indica que (3.7) se trata da expectativa de um valor estimado, ou seja, é necessária uma previsão, de quais serão as recompensas que poderão ser alcançadas no futuro. Por fim, o termo $r_t(\theta)$ se refere a razão de probabilidade dada por

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}. \tag{3.8}$$

Para melhor entendimento, (3.7) deve ser analisada por partes. Primeiramente, a razão $r_t(\theta)$ relaciona a política atual com a política antes da atualização dos parâmetros e indica o quanto a probabilidade de se tomar a ação a_t no estado s_t aumentou após a política ser atualizada. A razão $r(t)$ nunca retornará um valor negativo pois a probabilidade deve ser maior igual a zero. Sendo assim, para valores maiores do que 1, a razão indica que a probabilidade aumentou. Já valores entre 0 e 1 indicam que, na nova política, é menos provável escolher a ação a_t do que era na política anterior. Portanto, o produto $r_t(\theta)\hat{A}_t$ visa maximizar a probabilidade de se tomar ações vantajosas, e diminuir a probabilidade de se tomar ações que levem a valores negativos de vantagem.

O segundo termo de (3.7), $clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}$ é semelhante ao primeiro termo, com exceção que a razão passa pela função de corte, $clip$. Para explicar essa função, considere uma equação X qualquer e x_{inf} e x_{sup}, que são, respectivamente, os valores de limite inferior e superior, a função $clip$ possui o seguinte comportamento

$$clip(X, x_{inf}, x_{sup}) = \max\left(\min(X, x_{sup}), x_{inf}\right). \tag{3.9}$$

Essa função irá limitar $r(t)$ em um valor próximo de 1 de acordo com uma região de confiança definida por ε, com isso, limita-se o quanto a política pode aumentar ou diminuir a probabilidade de escolher a ação a_t em cada atualização.

Por fim, a função *min* garante um comportamento pessimista, limitando as atualizações que avançam em direção ao ponto de máximo da função, visto que se tratam de valores estimados e não de valores exatos, evitando que haja uma mudança muito grande na escolha das ações, que poderia gerar resultados inesperados. Do contrário, são permitidas atualizações negativas sem restrição. Caso se aumente muito a probabilidade de se tomar uma ação que gere uma vantagem negativa, a próxima atualização não será limitada, permitindo se desfazer integralmente do último passo de otimização que havia sido feito.

Dessa forma, a política se mantém próxima de uma região de confiança onde há uma boa estimativa dos valores. Limitando a atualização, é possível manter a tendência de crescimento ou decrescimento da probabilidade de certa ação, ou seja, mantém-se a direção dada pelo gradiente, mas atualizando-a de forma gradual.

3.3. Conclusão

Esse capítulo apresentou a base teórica para a utilização da aprendizagem por reforço e introduziu os métodos de gradiente de política. Foi detalhada a técnica de otimização da política por área de confiança que limita os passos de atualização ao redor de uma área segura e, posteriormente, foi apresentada a evolução da técnica, chamada de otimização de política proximal, que facilita a aplicação ao fazer a limitação diretamente na função objetivo.

4. Redes Neurais Artificiais

O crescimento dos estudos relacionados a inteligência artificial estão relacionados a popularização das redes neurais artificiais. Em aprendizado por reforço elas são especialmente importantes quando se trabalha com sistemas contínuos e são a principal maneira de se representar os estimadores e a política.

Nos seres vivos, o neurônio é a célula especializada em transmitir e processar as informações. O conjunto de neurônios é chamado de rede neural. Uma rede neural artificial é uma estrutura computacional baseada nesse sistema, formada por neurônios artificiais que são inspirados nos neurônios biológicos conforme mostrado na Figura 4.1.

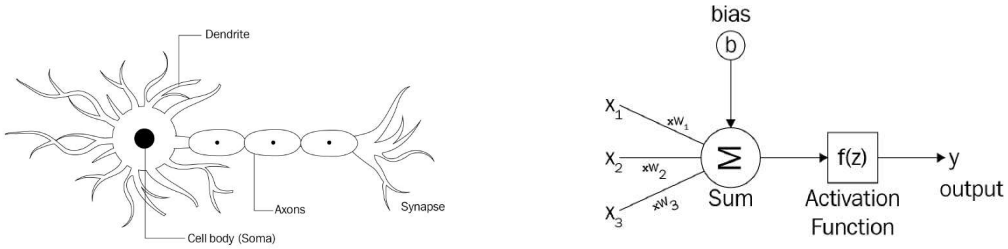


Figura 4.1: Comparação entre neurônio artificial típico e neurônio biológico.
Fonte: Ravichandiran (2018).

O neurônio artificial faz a soma de todos os sinais de entrada ponderados pelo respectivo peso da conexão. O valor resultante determina se o neurônio irá produzir um sinal de saída ou não. O limiar tem o papel de aumentar ou diminuir a influência do valor de entrada para essa ativação interna do neurônio. Já função de ativação altera o valor da soma das entradas e também permite adicionar não linearidades no sistema.

O aprendizado de um neurônio ocorre com a adaptação dos pesos sinápticos de forma que a saída apresente os valores desejados a partir de determinados valores de entrada. O verdadeiro poder de se utilizar modelos de neurônios é alcançado pela combinação de vários deles, formandos as redes neurais em que os pesos de cada neurônio são treinados usando as dependências de uns com os outros (AGGARWAL *et al.*, 2018). Dessa forma, é possível representar funções e comportamentos altamente complexos quanto mais neurônios forem utilizados. Para uma implementação em sistema contínuo, no qual se tem uma infinidade de possíveis valores para entrada e para a saída, usar redes neurais permite uma representação fiel do sistema sem a necessidade de aproximações. A estrutura de uma rede neural é feita por camadas, sendo diferenciadas nos neurônios da camada de entrada, neurônios da camada de saída e os neurônios das camadas intermediárias, também chamadas de ocultas. Esses últimos não são obrigatórios, e as redes neurais que possuem uma ou mais camadas de neurônios intermediários são nomeadas como redes neurais profundas.

4.1. Treinamento

Para treinar redes neurais, é necessário ter um conjunto de dados separado para esta finalidade. Então, aplica-se algum algoritmo de atualização dos pesos em busca do correto mapeamento entre as entradas e saídas no conjunto de dados. Esse processo deve acontecer repetidas épocas para um treinamento adequado. Sendo que a época é cada vez que todo o conjunto de dados de treinamento é apresentado à rede. O principal método de se atualizar os pesos da rede neural é por meio do gradiente de descida

$$\theta_{t+1} = \theta_t - \alpha\widehat{\nabla J(\theta_t)}. \tag{4.1}$$

Note que a equação (4.1) busca atualizar os parâmetros θ a partir do gradiente de uma função objetivo, também conhecida como função de custo. Essa equação permite atualizar os parâmetros, que no caso são os pesos da rede neural, de forma a buscar a minimização de $J(\theta_t)$, visto que a atualização ocorre no sentido contrário ao gradiente, o qual indica o ponto de maior variação da função. Essa busca é feita de passo em passo, cujo tamanho do passo é definido pela taxa de aprendizagem α, valor que varia de 0 a 1. Valores maiores levam a um aprendizado rápido mas não garantem que seja alcançado o resultado ideal. Por sua vez, valores muito pequenos permitem se aproximar mais do valor mínimo da função objetivo ao custo de um aprendizado mais lento.

A função objetivo deve representar o valor desejado na saída num determinado tempo t a partir de um dado sinal de entrada no mesmo instante. Essa função muitas vezes é estocástica, formada por subconjuntos aleatórios dos dados de treinamento, também chamados de lotes. Nesses casos, a equação (4.1) passa a ser conhecida como SGD (gradiente de descida estocástico, do inglês *Stochastic Gradient Descent*).

4.2. Retropropagação

Aplicar SGD em uma rede neural complexa não é trivial pois os valores de saída de uma camada dependem indiretamente de cada parâmetro e função de ativação das camadas anteriores. A maneira mais comum de se encontrar os gradientes utilizados em (4.1) é pelo algoritmo de retropropagação (do inglês *backpropagation*).

Conforme Figura 4.2, o algoritmo possui três fases bem definidas.

No primeiro momento, há a propagação positiva dos sinais de entrada x, os quais caminham até a saída y sem alteração da matriz de pesos W ou do vetor de limiares b. Os valores de ativação interna u, antes de passarem pela função de ativação f, são armazenados juntamente com os valores de y para serem utilizados posteriormente. Como ilustrado na Figura 4.2, esse processo se repete para todas camadas da rede neural até a saída final y^3.

Em seguida, o valor de saída é retropropagado utilizando o gradiente das funções de

custo $\dot{F}$ em relação às ativações internas e gerando os valores de sensibilidades δ. A retropropagação parte da última camada e se repete até a camada inicial.

Por fim os valores de sensibilidade serão utilizados para atualizar cada peso e limiar da rede neural. Este processo deve ser realizado para cada padrão de treinamento disponível por um número de épocas pré-determinado ou até que seja alcançado algum outro requisito de parada, como a minimização do erro.

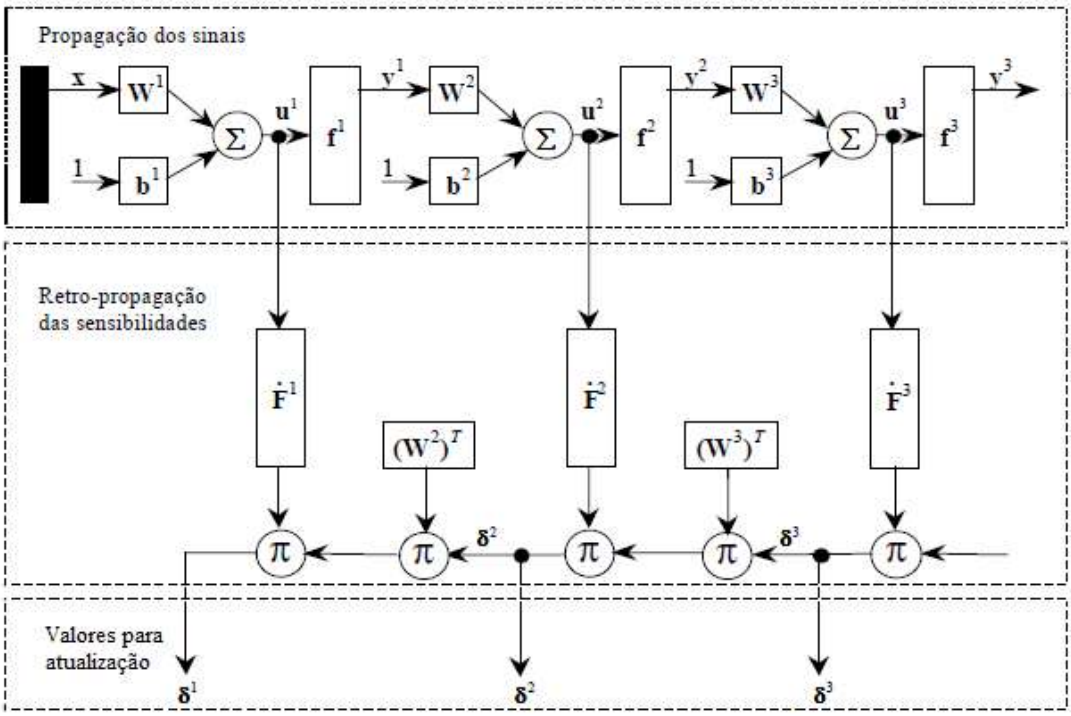


Figura 4.2: Algoritmo de retropropagação para uma rede neural de três camadas.
Fonte: O Autor.

Este procedimento é detalhado no algoritmo 1 considerando um conjunto com N padrões de treinamento e uma rede neural de M camadas.

Algorithm 1 Backpropagation

for $i \leftarrow 1$ até N **do**

$y^0 \leftarrow x_i$

$y_i^{m+1} \leftarrow f^{m+1}(W^{m+1} y_i^m + b^{m+1}),\ m = 0,1,\cdots,M-1,$

$\delta_i^M \leftarrow \frac{\partial J}{\partial u_i^M}$

$\delta_i^m \leftarrow \dot{F}^m(u_i^m)(W^{m+1})\delta_i^{m+1},\ m = M-1,\cdots,2,1,$

$W^m \leftarrow W^m - \alpha \delta_i^m (y_i^{m-1})^T,\ m = 1,\cdots,M,$

$b^m \leftarrow b^m - \alpha \delta_i^m,\ m = 1,\cdots,M,$

end for

Para iniciar a retropropagação, o primeiro valor de sensibilidade é calculado a partir do gradiente da função de custo J em relação a ativação interna da última camada

$$\delta^M = \frac{\partial J}{\partial u^M}. \tag{4.2}$$

As demais sensibilidades são calculadas por meio do gradiente da função de ativação em relação a cada neurônio S^m da camada m, organizados em formato matricial como se segue

$$\dot{F}^m(u^m) = \begin{bmatrix} \dot{f}^m(u_1^m) & 0 & \cdots & 0 \\ 0 & \dot{f}^m(u_2^m) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \dot{f}^m(u_{S^m}^m) \end{bmatrix}, \tag{4.3}$$

em que

$$\dot{f}^m(u^m) = \frac{\partial f^m(u^m)}{\partial u^m}. \tag{4.4}$$

4.3. Estimativa de Momento Adaptativo — ADAM

Em otimizações estocásticas, os parâmetros são atualizados a partir de diferentes subconjuntos dos dados de treinamento. A aplicação direta da (4.1) em conjunto com a retropropagação não possui boa eficiência, nesses casos, por apresentar um mesmo passo de atualização para todas as entradas. Neste contexto, o algoritmo ADAM (do inglês *adaptive moment estimation*) é uma extensão do SGD que computa taxas de aprendizado individuais para cada padrão, que são adaptadas a partir do primeiro e segundo momentos dos respectivos gradientes (KINGMA e BA, 2014). O método apresenta uma convergência rápida, trabalha bem com gradientes esparsos e configurações não estacionárias, e a magnitude das suas atualizações não é afetada pelas variações na escala do gradiente.

O algoritmo utiliza médias móveis exponenciais do gradiente para realizar a atualização dos parâmetros. As médias móveis são estimadas pelo primeiro momento (média) e segundo momento (variância) e controlados por uma taxa de decaimento β_1 e β_2. Estas taxas podem ser constantes, tipicamente valendo 0,9 e 0,999, respectivamente. Ambos momentos, respectivamente m_t e υ_t, são iniciados como vetores de 0, e então são atualizados a partir dos gradientes da função de custo estocástica

$$m_t = \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t, \tag{4.5}$$

$$\upsilon_t = \beta_2 \cdot \upsilon_{t-1} + (1-\beta_2) \cdot g_t^2, \tag{4.6}$$

em que

$$g_t = \nabla_\theta J(\theta_{t-1}). \tag{4.7}$$

Em seguida, é necessário corrigir o viés dos estimadores de momento

$$\widehat{m}_t = \frac{m_t}{1-\beta_1^t}, \tag{4.8}$$

$$\widehat{\upsilon}_t = \frac{\upsilon_t}{1-\beta_2^t}, \tag{4.9}$$

Por fim, os pesos da rede neural são atualizados

$$\theta_t = \theta_{t-1} - \alpha \frac{\widehat{m}_t}{\sqrt{\widehat{\upsilon}_t} + \varepsilon}. \tag{4.10}$$

4.4. Conclusão

Nesse capítulo foi introduzido o conceito de redes neurais artificiais, que são essenciais para uma boa representação das variáveis do processo de forma contínua. Neste trabalho as redes neurais serão responsáveis por mapear os estados observados, que consistem das variáveis controladas, em ações, que serão as variáveis de controle. Foi apresentado o método de treinamento das redes neurais e o detalhamento da técnica de retropropagação. Por fim foi apresentado o método de estimativa de momento adaptativo que permite um treinamento mais eficiente da rede neural ao fazer uso da média móvel dos gradientes. Essas informações foram utilizadas para fazer a implementação das redes neurais e seu treinamento no MATLAB.

5. Detalhes da implementação do algoritmo PPO

Para a implementação do controle de espessador empregando aprendizado por reforço, define-se que as observações são as entradas do sistema, representadas pelas variáveis controladas que serão a altura da zona de clarificação e a densidade no *underflow*, conforme Figura 2.2. As variáveis manipuladas são as ações geradas pela política do PPO e consistem da vazão de saída no *underflow* e a dosagem de floculante. O controle da vazão permite manter o processo de sedimentação no espessador e liberar a polpa à medida que ela atinge o nível de solidificação desejada. Por outro lado, o uso de floculante tem como função acelerar o processo de sedimentação de sólidos por meio da formação de partículas maiores. Como esse insumo tem um custo, o ideal é reduzir a necessidade de sua utilização, logo deseja-se que a ação de controle que representa o gasto de floculante seja a menor possível.

Apesar de grande parte dos trabalhos na área de aprendizado por reforço utilizarem *Python* como linguagem de programação, optou-se por se trabalhar com *MATLAB*. A escolha visa fazer uma melhor utilização do simulador desenvolvido em Pereira (2021), evitando conversões e adaptações.

Como ressaltado no trabalho de Andrychowicz *et al.* (2020), detalhes e decisões de implementação possuem um grande peso no sucesso ou não do algoritmo de aprendizado por reforço, ao passo que são pouco explorados na teoria. Desta forma, para a programação do PPO, foram utilizadas como referência as produções em *Python* do *OpenAI* que possuem algoritmos de código aberto consolidados, como os apresentados em Dhariwal *et al.* (2017) e Achiam (2018).

5.1. Redes Neurais

O controle foi feito considerando tanto as observações quanto as ações como variáveis contínuas, portanto, optou-se utilizar duas redes neurais numa estrutura ator-crítico conforme a Figura 5.1. Nessa estrutura, o ator tem o papel de definir a política de escolha das ações que serão aplicadas, e o crítico avalia quão bom é o ator guiando seu treinamento. A entrada de ambas as redes são as observações que representam o estado do ambiente. A saída do crítico é uma estimativa do valor que será alcançado no próximo estado, e é essencial para treinar a rede neural do ator. A rede neural do crítico é treinada a partir do erro de estimativa entre o valor gerado e a recompensa retornada do ambiente.

Tanto a rede crítico quanto o ator possuem estruturas equivalentes, consistindo de 2 camadas internas, sendo que o crítico possui 64 neurônios na primeira camada e 32 na segunda, enquanto o ator possui 32 neurônios em ambas camadas. As duas redes possuem 2 entradas, onde serão observadas a densidade do material na saída do espessador e a altura da interface do líquido clarificado. O número de neurônios e de camadas foi definido de forma empírica, utilizando como referência inicial os trabalhos de Hallén *et al.* (2019) e Andrychowicz *et al.*

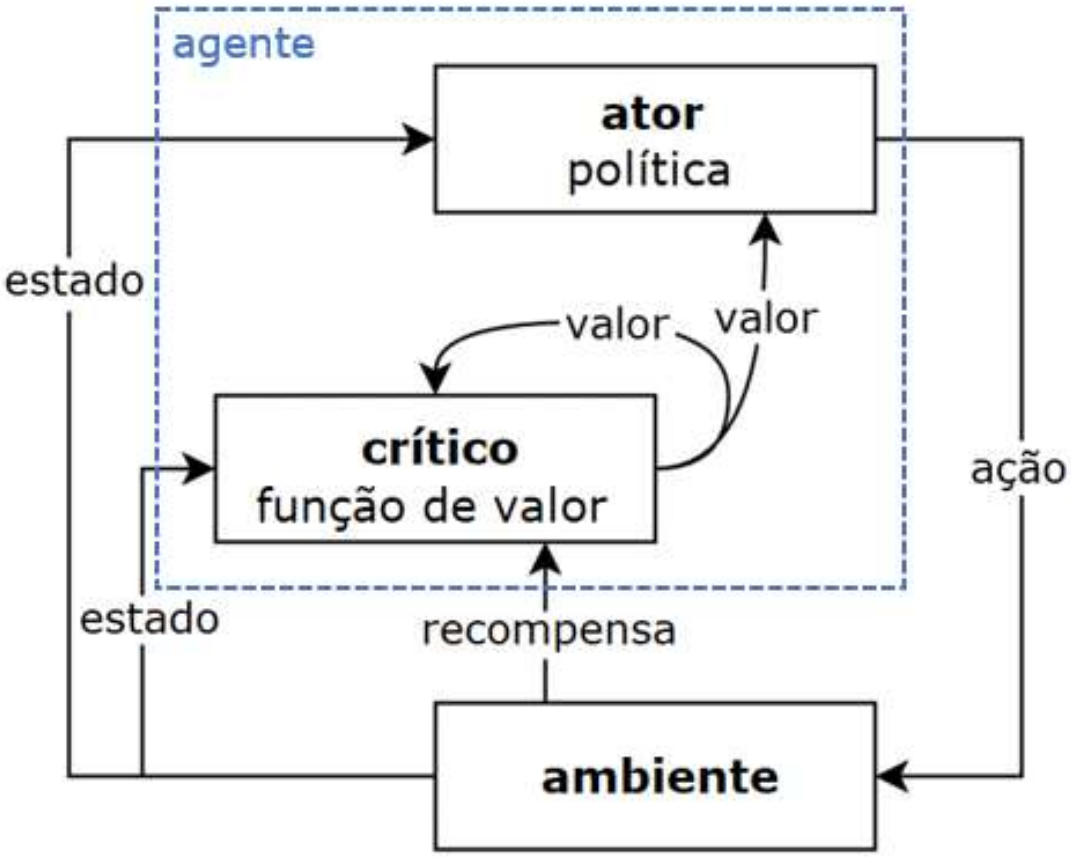


Figura 5.1: Arquitetura Ator-Critico.
Fonte: O autor.

(2020).

O crítico terá uma única saída para representar o valor geral do sistema. Enquanto o ator terá 2 saídas, que serão a base do sinal de controle. Cada saída representa a média μ de uma distribuição estatística normal. Para formar essa distribuição normal será utilizado ainda um desvio padrão σ predefinido. A partir dessa distribuição, será feita uma amostragem de qual próxima ação a ser aplicada, permitindo assim a exploração do espaço de estados. Quanto maior o desvio padrão, maior será a exploração do sistema, uma característica que é mais desejada no início do treinamento e que deve diminuir ao longo do processo. As ações definidas como sinal de controle serão a vazão de saída do espessador e a dosagem de floculante adicionada.

Seguindo a recomendação apresentada em Andrychowicz *et al.* (2020), a última camada das redes neurais foi inicializada com valores 100 vezes menores do que nas demais. Todas as camadas utilizarão a função tangente hiperbólica

$$y(u_k) = \tanh(pu_k) = \frac{\varepsilon^{pu_k} - \varepsilon^{-pu_k}}{\varepsilon^{pu_k} + \varepsilon^{-pu_k}}, \tag{5.1}$$

como função de ativação para permitir não linearidades. Note que a derivada de (5.1) é dada

por

$$\frac{\partial y}{\partial u_k} = p(1-y^2) > 0. \tag{5.2}$$

A função tangente hiperbólica limita a saída da rede neural entre -1 e 1, como pode ser visto na Figura 5.2. Essa é uma característica importante tendo em vista que as ações de controle também deverão ser limitadas, bastando escalar para os limites corretos.

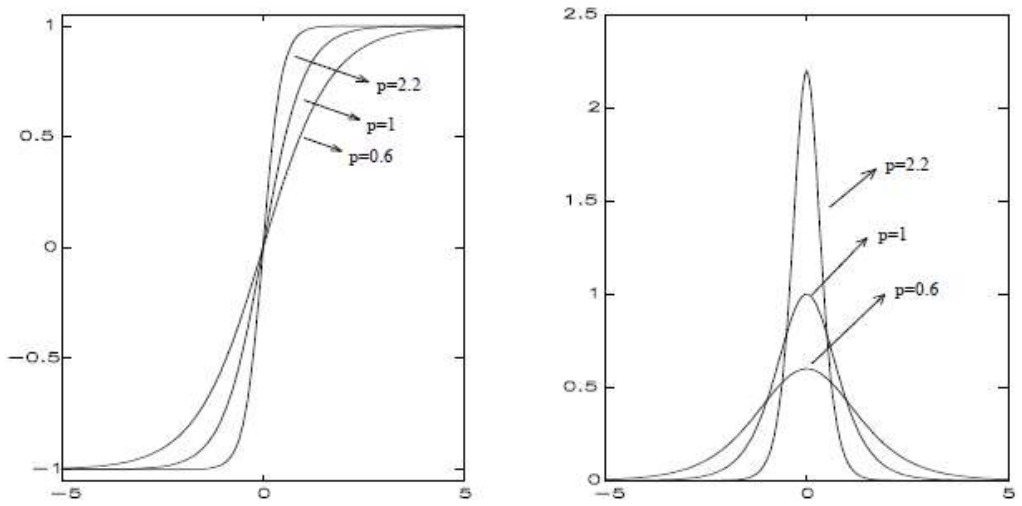


Figura 5.2: Funções (a) tangente hiperbólica e (b) sua derivada.
Fonte: Aggarwal *et al.* (2018).

5.2. Função de densidade de probabilidade

Cada ação de controle terá sua própria distribuição normal que irá definir por amostragem a ação utilizada para o próximo estado do sistema. A característica aleatória da distribuição estatística permite explorar o espaço de ações para que o algoritmo aprenda quais são as ações mais ou menos vantajosas a partir de um determinado estado do sistema.

O desvio padrão será definido por uma relação direta com os erros em regime permanente, isto é,

$$\sigma = erro_{densidade} + erro_{altura} + 0{,}0001. \tag{5.3}$$

O objetivo é diminuir o desvio padrão à medida que o sistema se aproxima do sinal de referência desejado. Com isso, a amostragem da política irá tender ao valor de média gerado pela rede neural. Essa é uma forma de diminuir a exploração levando a sinais de controles que se aproximem dos valores ótimos. A parcela numérica foi acrescentada ao desvio padrão para evitar que seja zerado em algum momento, o que acarretaria erros na representação da distribuição padrão.

A política será representada pela função de densidade de probabilidade

$$\pi(a|s,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e} \left(-\frac{(a-\mu(s,\theta))^2}{2\sigma^2} \right). \tag{5.4}$$

Essa função indica quão provável é uma determinada ação a ser amostrada partindo de uma distribuição normal. Como a média é definida por uma rede neural, ela depende dos estados de entrada e dos parâmetros desta rede. Enquanto a política em si, depende da média e, portanto, dos estados e pesos da rede neural, e também depende da ação. Conforme (3.8), o controlador buscará aumentar a probabilidade das ações que geraram altas recompensas e diminuir a probabilidade de ações que geraram recompensas mais baixas. Essa função de densidade é calculada separadamente para as duas ações de controle.

5.3. Função de recompensa

A proposta para a função de recompensa a ser maximizada é

$$R = -C_1|e_{den}| - C_2|e_{alt}| - C_3|a_{vaz}| - C_4|a_{floc}|, \tag{5.5}$$

em que e_{den} e e_{alt} são, respectivamente, o erro percentual da densidade de sólidos e da altura de interface em relação às referências desejadas. Os termos a_{vaz} e a_{floc} são decorrentes da amplitude dos sinais de controle enviados para a vazão de descarga, e a amplitude do sinal de controle da adição de floculantes. O valor de a_{vaz} é a diferença entre a amplitude da vazão de descarga atual no instante t e a amplitude da vazão no instante $t-1$. Por outro lado, a_{floc} é a dosagem de floculante no próprio instante t, impactando diretamente a função de recompensa. As constantes C_1 a C_4 são pesos adicionados para permitir alterar a influência de cada variável no resultado, sendo variados ao longo dos testes. Todas as variáveis de 5.5 serão sempre positivas, dessa forma, a função de recompensa é maximizada quando todas as variáveis forem zero, indicando o caso em que não há erro em regime permanente e nenhuma ação de controle precisa ser tomada.

Antes da etapa de treinamento, o vetor com as recompensas é normalizado a partir do seu valor médio e desvio padrão

$$R = \frac{R - \mu_R}{\sigma_R}. \tag{5.6}$$

5.4. Estimador da Vantagem

Para o cálculo da vantagem, foi empregada a versão de estimador de vantagens generalizadas (GAE, do inglês *Generate Advantage Estimative*) utilizada por Schulman *et al.* (2017) e definida como

$$\hat{A}_t = \delta_t + (\gamma\lambda)\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \tag{5.7}$$

em que

$$\delta_t = r_t + \gamma V(s_{t+1}) - V(s_t). \tag{5.8}$$

A constante λ é um fator de desconto responsável por ajustar o viés da estimativa definido empiricamente, tendo valor entre 0 e 1. T é o número total de passos que serão utilizados, definindo a quantidade de transições disponíveis para atualização.

5.5. Gradiente da função de custo

Para a utilização do algoritmo de retropropagação, será necessário utilizar o gradiente da função de custo. A partir de (3.7) e (3.8), precisamos encontrar a derivada da seguinte equação

$$J(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}\hat{A}_t. \tag{5.9}$$

Durante a etapa de treinamento do ator, as vantagens permanecerão constantes. Além disso, sabendo que $\nabla \ln x = \frac{\nabla x}{x}$, temos

$$\frac{\partial J(\theta)}{\partial \theta} = \frac{\partial \ln \pi_\theta(a_t|s_t)}{\partial \theta} \tag{5.10}$$

em que a política π_θ é definida pela função de densidade de probabilidade (5.4), e seu logaritmo natural é

$$\ln\pi(a|s,\theta) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(a-\mu(s,\theta)^2). \tag{5.11}$$

Considerando que o treinamento ocorre a partir de repetidas épocas no mesmo conjunto de dados, ou seja, várias passagens do conjunto formado por (a_t,s_t,s_{t+1}), a média $\mu(s,\theta)$ é a única variável dependente dos parâmetros, então pode-se buscar maximizar a seguinte função em relação às ativações internas u

$$\frac{\partial \ln\pi_\theta}{\partial \theta} = \frac{1}{\sigma^2}(a-\mu)\frac{\partial \mu}{\partial \theta} = 0. \tag{5.12}$$

Como as médias são as saídas da rede neural, y^M, o gradiente da função de custo utilizado na retropropagação será

$$\frac{\partial J}{\partial \theta} \equiv \frac{\partial J}{\partial u^M} = \frac{1}{\sigma^2}(a-y^M)\frac{\partial y^M}{\partial u^M}. \tag{5.13}$$

5.6. Parâmetros utilizados

A partir das recomendações de Andrychowicz *et al.* (2020), foram realizados diversos experimentos para definir o conjunto de parâmetros que permitisse alcançar os melhores resultados gerais com um bom desempenho. Os parâmetros gerais utilizados no algoritmo são apresentados na Tabela 5.1, enquanto a Tabela 5.2 e a Tabela 5.3 apresentam os parâmetros

específicos para o ator e para o crítico, respectivamente. Os valores apresentados foram a base para os experimentos realizados.

Tabela 5.1: Parâmetros gerais do PPO utilizados nos experimentos.

Parâmetro	**Valor**
Número de passos T	80
Quantidade de mini lotes	4
Ajuste de corte ε	0,25
Fator de desconto da vantagem λ	0,9
Fator de desconto γ	0,99
Função de recompensa	$-5e_{den} - 2e_{alt} - 0.5a_{floc}$

Fonte: O autor.

Tabela 5.2: Parâmetros de treinamento do ator.

Parâmetro	**Valor**
Número de neurônios nas camadas ocultas	64 x 32
Épocas de treinamento	30
Taxa de aprendizado α	0,002
Taxa de decaimento β_1	0,9
Taxa de decaimento β_2	0,999

Fonte: O autor.

Tabela 5.3: Parâmetros de treinamento do crítico.

Parâmetro	**Valor**
Número de neurônios nas camadas ocultas	32 x 32
Épocas de treinamento	300
Taxa de aprendizado α	0,001
Taxa de decaimento β_1	0,9
Taxa de decaimento β_2	0,999

Fonte: O autor.

Para a função de recompensa 5.5, foi utilizado um peso maior para o erro de densidade, constante C_1, visto que geralmente esse é o principal foco de controle em espessadores. As ações de controle serão um objetivo secundário no trabalho e por isso foram feitos testes com constantes menores. Notou-se que a parcela referente a variação da vazão de saída não trazia benefícios relevantes para o controle enquanto dificultava a estabilização do processo, portanto

definiu-se $C_3 = 0$. Dessa forma, a função de recompensa utilizada como referência para os experimentos foi

$$R = -5|e_{den}| - 2|e_{alt}| - 0{,}5|a_{floc}|. \tag{5.14}$$

5.7. Conclusão

Nesse capítulo foram descritos os pontos da implementação prática do algoritmo que demandaram alguma decisão de projeto. Enquanto essas decisões tem um alto impacto no desempenho e comportamento do agente de aprendizado por reforço, as literaturas atuais possuem pouco detalhamento em quais práticas devem ser adotadas. Sendo assim, esse capítulo foi importante para explicitar qual a estrutura da rede neural utilizada, a função de densidade de probabilidade que gera as ações contínuas e qual a função de recompensa e o estimador da vantagem escolhidos. Foi também apresentado o gradiente da função de custo do PPO, necessário na etapa de retropropagação, e os parâmetros que serviram de base para os testes apresentados no capítulo de resultados.

6. Resultados

Para a aplicação das técnicas apresentadas utilizou-se o simulador de espessadores desenvolvido por Pereira (2021, Capítulo 7). Este simulador, desenvolvido em *MATLAB*, foi adaptado de forma a permitir a integração com o algoritmo de aprendizado por reforço. O PPO foi aplicado a um modelo de espessador convencional cilindro-cônico com 35 m de diâmetro máximo e 7 m de altura total. A altura da região cônica é 3,8 m com inclinação de 14°. O poço de alimentação possui 2 m. A alimentação do espessador foi mantida constante com vazão de 800 ton/h exceto quando indicado. Considera-se a densidade da polpa de 1,64 g/cm^3 e a densidade dos sólidos de 4,7 g/cm^3. Foi utilizado no simulador um passo espacial $dz = 0{,}1\ m$ e um passo de tempo $dt = 11{,}74\ s$.

Devido a complexidade de um sistema com múltiplas entradas e múltiplas saídas, decidiu-se testar inicialmente o sistema com apenas uma ação de controle, mantendo-se a segunda ação fixa, em busca de estabilizar os sinais de densidade e altura da interface. Nesses casos, a função de recompensa não considerou a parcela referente a ação que foi fixada.

O algoritmo de PPO, coleta dados a cada T passos da simulação, que equivalem a aproximadamente 29 minutos. Essa coleta começa após um período de 40 horas, intervalo em que o sistema entra em regime de operação sendo controlado em malha aberta com valor fixo de dosagem de floculante em 10 g/ton e de vazão de saída em 300 ton/h. No momento em que ocorre a coleta, os novos sinais de controle que serão aplicados ao sistema são gerados a partir da política atual do PPO. Como uma nova ação de controle só é gerada após esse tempo de 29 minutos, optou-se por variar a ação linearmente ao longo do tempo até alcançar o novo valor de controle.

6.1. Controle pela Vazão de Saída

O controle pela vazão de saída foi implementado mantendo-se a concentração de floculante adicionada fixa em 10 g/ton. Como referência foi utilizado 2,16 m para altura da interface e variou-se o *setpoint* de densidade de 2,0375 g/cm^3 para 2,128 g/cm^3 em $t = 40\ h$ e 2,073 g/cm^3 em $t = 350\ h$. O resultado para 600 h de simulação pode ser visto na Figura 6.1. As linhas horizontais indicam os momentos em que houveram treinamento do sistema e o inicio da coleta de um novo conjunto de dados.

Analisando as variáveis controladas, a amplitude tanto da densidade quanto da altura diminuem ao longo do tempo, mas com 600 horas de simulação não se estabilizam. Para o *setpoint* de 2,073$^g/cm^3$ de densidade a referência é alcançada em três momentos, mas como nestes momentos ainda há erro relevante na altura de interface, o agente continua a exploração levando a atualizações da política que distanciam o sistema da referência. O comportamento se repete até o final das 600 horas iniciais. Apesar do sinal de controle possuir uma grande variação em amplitude, este é um comportamento esperado ao longo de 600 horas. A Figura

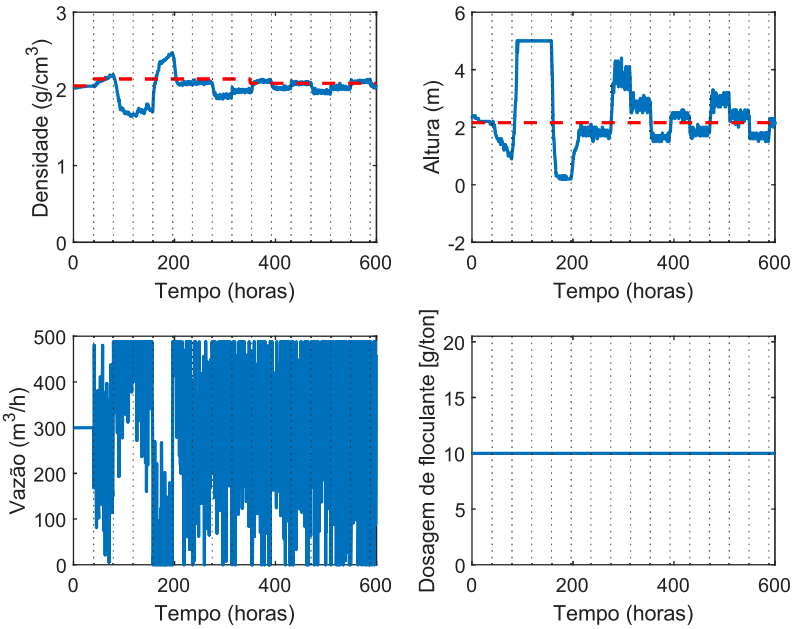
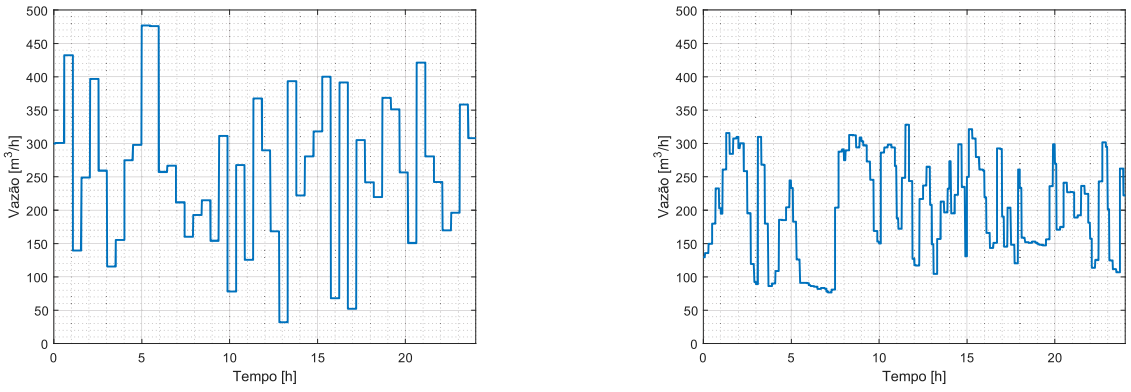


Figura 6.1: Sinais das malhas de controle de vazão para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow*.
Fonte: O autor.

6.2 mostra os dados para as primeiras 24 horas em que a malha é fechada na simulação em comparação com o sinal de vazão obtido de uma base de dados reais do espessador.



(a) (b)

Figura 6.2: Vazão no *underflow* ao longo de 24 *h* (a) proposto pelo controlador PPO e (b) sinal amostrado de um espessador real em funcionamento na Vale S.A.
Fonte: O autor.

Pela Figura 6.2, nota-se como o sinal proposto pelo PPO é semelhante ao aplicado no sistema real, quando comparado o mesmo intervalo de tempo, tanto para a variação de amplitude quanto para a frequência das variações. Enquanto no sistema real as variações ocorrem como degraus, no PPO as variações foram aplicadas como rampa em busca de suavizar o sinal de

controle. A parcela $C_3 a_{vaz}$, na função de recompensa (5.5), é outra forma de tentar melhorar o sinal de controle, sendo assim, foi feita nova simulação mantendo os parâmetros que geraram a Figura 6.1 mas utilizando como recompensa 6.1. O resultado pode ser visto na Figura 6.3.

$$R = -5e_{den} - 2e_{alt} - 0.5a_{vaz} \tag{6.1}$$

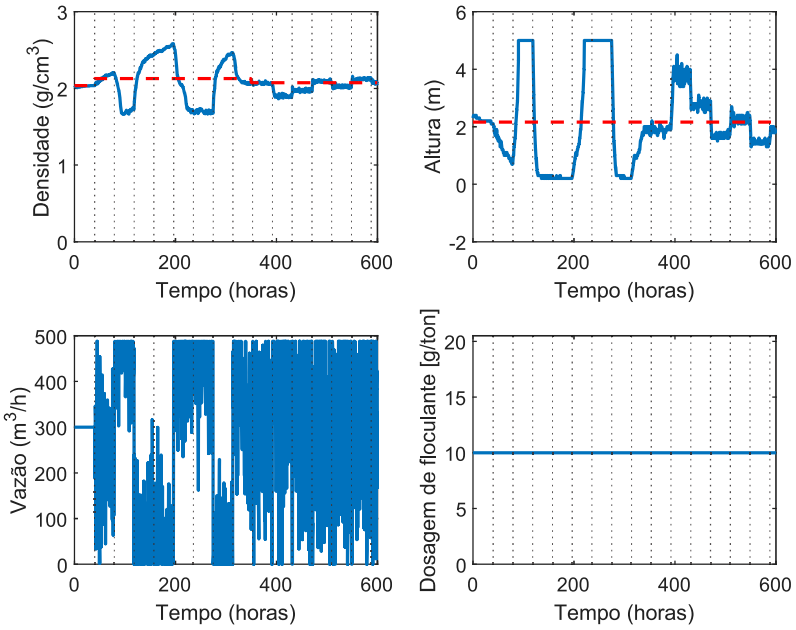


Figura 6.3: Sinais das malhas de controle de vazão para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow* e considerando o sinal de controle na recompensa.
Fonte: O autor.

Nota-se que na primeira metade do teste, o sinal de controle da vazão de saída realmente teve menor variação, no entanto as variáveis controladas oscilaram mais. O controlador só conseguiu levar a densidade e a altura de interface em direção a referência quando aumentou a variação do sinal de controle, levando a um comportamento muito semelhante ao teste anterior, ou seja, o custo de estabilizar o sinal de controle não compensa comparado ao objetivo principal do sistema que é alcançar as referências. É possível que para um tempo maior de simulação, devido a característica de tentativa e erro do PPO, o agente encontrasse uma forma de atender ambas as restrições da recompensa, ou seja, seguir a referência e suavizar o sinal de controle. Como para o período trabalhado essa alteração na recompensa não apresentou melhoras, pelo contrário, atrasou o aprendizado do sistema, foi mantido para os demais testes o valor de C_3 igual a 0.

6.2. Controle pela Dosagem de Floculante

O controle pela dosagem de floculante foi implementado mantendo-se a vazão de saída fixa em 300 m^3/h, o resultado para 600 h de simulação, pode ser visto na Figura 6.4. Como referência foi utilizado 2,16 m para altura da interface e variou-se o *setpoint* de densidade de 2,0375 g/cm^3 para 2,128 g/cm^3 em $t = 40$ h e 2,073 g/cm^3 em $t = 350$ h.

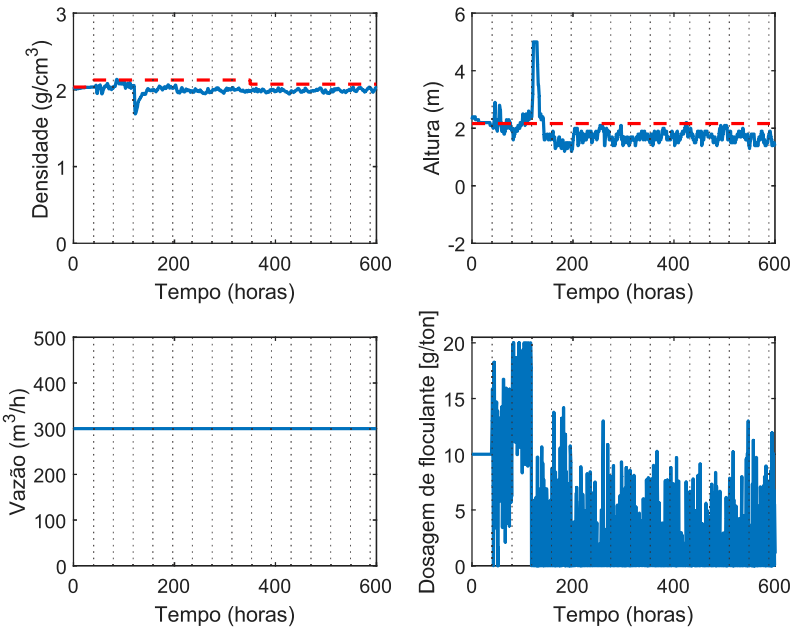


Figura 6.4: Sinais das malhas de controle de floculante para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow*.
Fonte: O autor.

Tanto a densidade quanto a altura de interface se estabilizaram com um erro significativo em relação as referências. Além disso, a variação de *setpoint* não gerou mudanças na política que gere o comportamento do sistema. De fato, notou-se a partir dos testes realizados que o PPO não responde a mudanças pequenas nas obervações, o que pode estar ligado aos parâmetros de treinamento das redes neurais, ou mesmo às suas estrutura. Vale ressaltar que o controle foi feito com uma dosagem de floculante abaixo do valor padrão de 10 g/ton, o que é um bom sinal, pois é um reagente de alto custo para indústria.

6.3. Controle por Vazão de Saída e Dosagem de Floculante

A Figura 6.5 apresenta o controle MIMO (do inglês *Multiple Input, Multiple Output*) do sistema utilizando como ação de controle tanto a vazão da saída quanto a dosagem de floculante. As referências foram alteradas após as 40 horas inciais de regime em malha aberta para 2,5 m de altura da interface e 2,184 g/cm^3 de densidade.

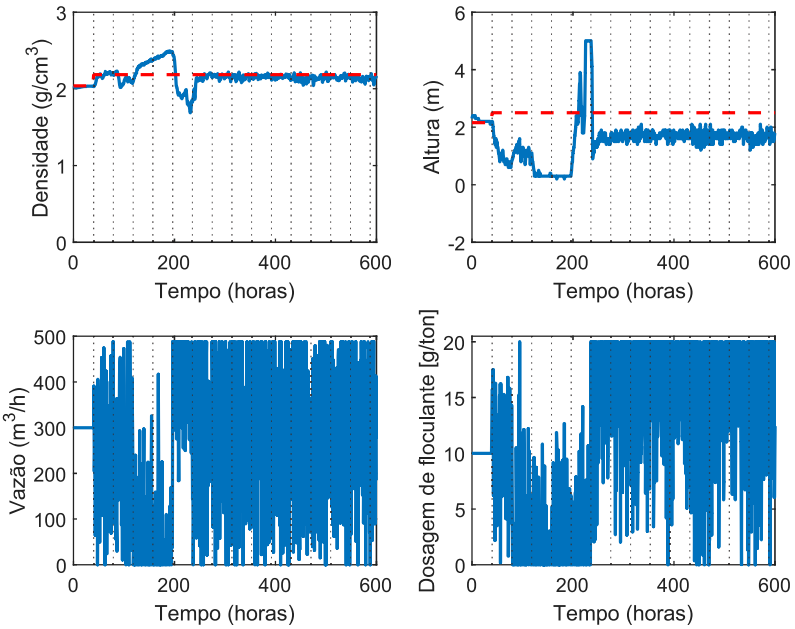


Figura 6.5: Sinais das malhas de controle para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se os *setpoints* de densidade no *underflow* e altura da interface.
Fonte: O autor.

A Figura 6.6 apresenta o valor da recompensa para o controle MIMO. Nota-se que após 250 horas o sistema alcança as maiores recompensas e, mesmo variando os sinais de controle, não consegue alcançar consistentemente valores maiores do que -1.

Tanto a densidade quanto a altura se estabilizam após as 250 horas, no entanto, enquanto foi possível alcançar a referência de densidade, houve um grande erro em relação a alturas de interface. Uma possível explicação está relacionada às características físicas do sistema. Para alcançar densidades mais altas na descarga do espessador, o material deve permanecer por um certo período dentro do tanque para ocorrer a sedimentação, que pode ser natural ou acelerada pelo acréscimo de floculante. Essa situação leva a um aumento da zona de sedimentação e consequente diminuição da altura de interface do líquido clarificado. De forma equivalente, há uma limitação física quando se deseja valores de densidade muito baixos. Caso a referência esteja próxima ao valor de densidade da entrada, o material poderá ser levado direto a saída com pouco tempo por já ter a porcentagem de sólidos desejadas, não chegando a formar uma zona de sedimentação extensa e, portanto, a altura de interface irá tender ao valor máximo. Lembrando que pela descrição do espessador, a altura máxima é de 5 m, que indica o fundo do equipamento.

Em um segundo momento, foi realizado testes variando-se apenas uma das referências. Na Figura 6.7 manteve-se fixo a altura em 2,16 m e variou-se a densidade para 2,184 g/cm^3 em 40 horas e então para 2,073 g/cm^3 após 350 horas. Na Figura 6.8 manteve-se a referência de densidade em 2,0375 g/cm^3 e variou-se a referência de altura da interface para 2,4 m em 40 horas e então 2 m após 350 horas.

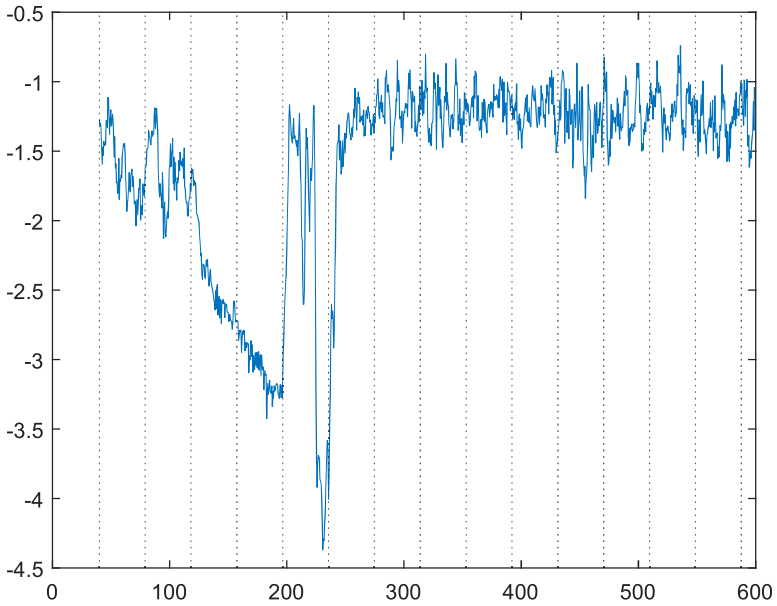


Figura 6.6: Recompensa do PPO para o controle MIMO variando-se *setpoints* de densidade no *underflow* e altura na interface.
Fonte: O autor.

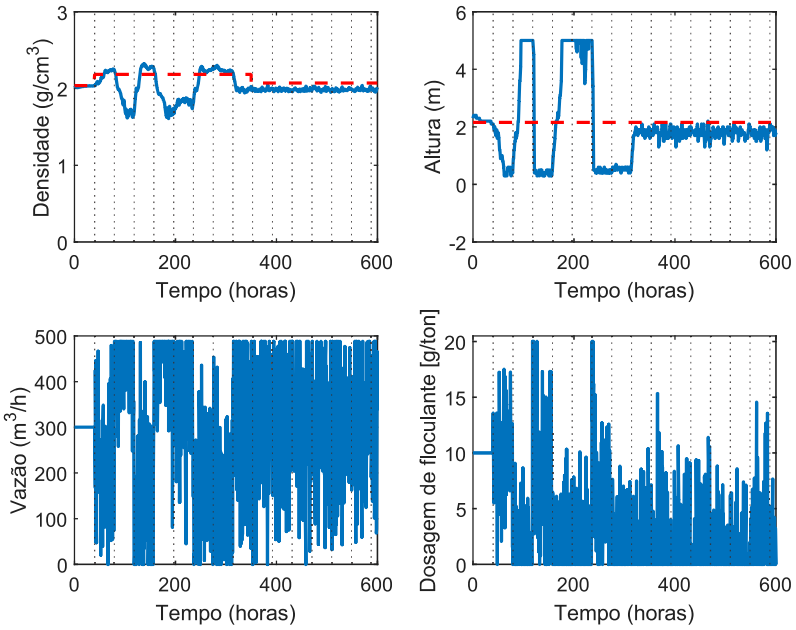


Figura 6.7: Sinais das malhas de controle para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de densidade no *underflow*.
Fonte: O autor.

É possível notar a dificuldade que o agente desenvolvido apresenta de alcançar as referências de forma otimizada na Figura 6.7. Ao alcançar valores que se aproximam da re-

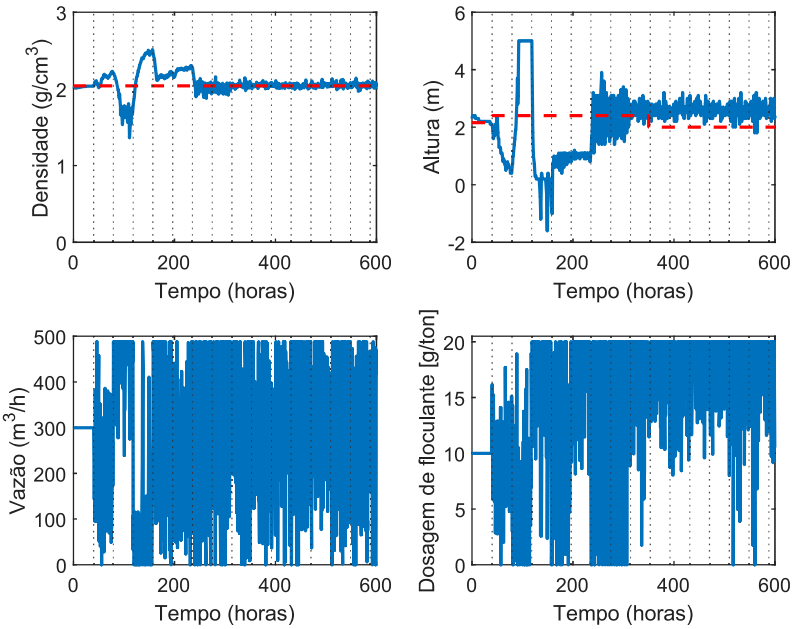


Figura 6.8: Sinais das malhas de controle para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se o *setpoint* de altura da interface. Fonte: O autor.

ferência da densidade, a altura apresenta grande erro, até que, após 300 horas de simulação, ele consegue aproximar a referência de altura. No entanto, na próxima atualização há a alteração da referência de densidade diminuindo o erro dessa variável e, consequentemente, aumentando a recompensa do sistema. Dessa forma, o sistema estabiliza em um ótimo local em que, apesar de alcançar valores próximos da referência para ambas as variáveis, apresenta considerável erro em regime permanente. No teste de variação da referência de altura, Figura 6.8, o agente conseguiu atingir ambas as referências com cerca de 240 horas, mas apresentando grande oscilação, principalmente da altura de interface. Outro problema foi que o sistema não variou o comportamento para alcançar a referência de altura após 350 horas, passando a ter um aumento do erro em regime permanente. Esse problema pode ter ocorrido pela variação de referência ter representado pouca diferença na função de recompensa ou mesmo devido a um *overfitting* do ator, como é chamado o supertreinamento de uma rede neural se especializando em um conjunto de dados específicos.

6.4. Controle com a presença de perturbações

Por fim, realizou-se o teste mantendo as referências fixas em 2,0375 g/cm^3 de densidade no *underflow* e 2,16 m de altura da interface, e aplicando uma perturbação no sistema ao variar a concentração de sólidos na alimentação de 0,173 para 0,1743 em 40 hroas e para 0,1718 em 350 horas. A mesma configuração foi utilizada em testes com técnicas estabelecidas de controle, no

caso controladores PI e DMC desenvolvido no trabalho de Pereira *et al.* (2020), Figura 6.9, e em testes utilizando PPO, Figura 6.10. Observa-se que a variação no material de alimentação apresentou grande impacto nos controladores DMC e PI, sendo que esse último demonstrou grande oscilação, sem conseguir estabilizar no último intervalo. Por outro lado, o controlador PPO conseguiu alcançar ambas as referências com comportamento semelhante ao demonstrado nos testes anteriores, se mostrando pouco sensível a esse tipo de perturbação. O fato do controle ainda estar oscilatório também influência, se tornando natural que mudanças e ruídos de pouca amplitude sejam ignorados, mas espera-se que com um maior ajuste do controle, o sistema continue robusto visto que o distúrbio não impacta diretamente na função de recompensa.

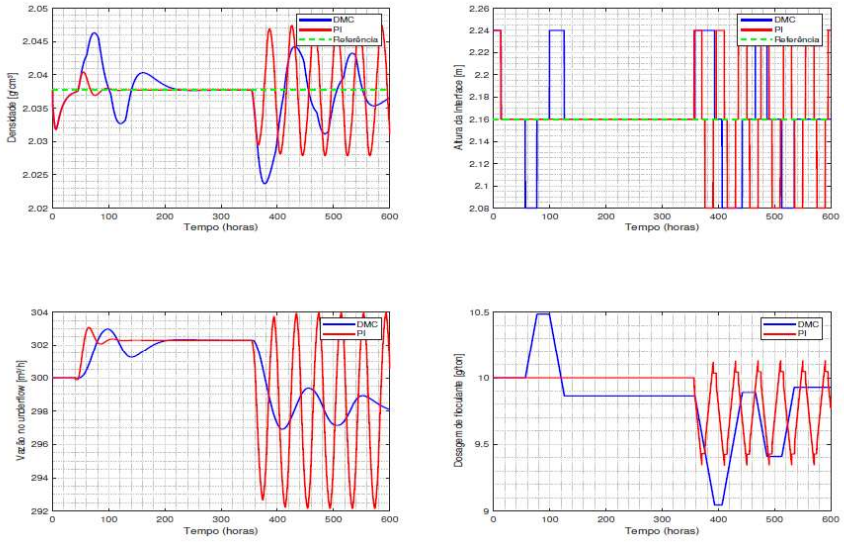


Figura 6.9: Sinais das malhas de controle DMC e PI para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se a densidade da polpa de alimentação.
Fonte: Adaptado de Pereira (2021).

6.5. Conclusão

Para o espessador estudado foi possível perceber que os melhores resultados foram alcançados ao se realizar o controle com múltiplas entradas, ou seja, vazão de saída e dosagem de floculantes, pois permitiu seguir as referências ao mesmo tempo que gerou economia de floculante. Importante destacar o fato de que a redução de gasto de floculante é um objetivo secundário que leva a uma redução de custo do processo, mas que não deve se sobrepor ao objetivo de regular a densidade adequadamente. Apesar do controlador utilizando PPO ter conseguido alcançar e se manter próximo dos valores de referência, o tempo gasto, o erro no regime permanente e a oscilação da variável controlada não são adequados para aplicações típicas de controle, sendo que hoje ainda é mais vantajoso utilizar técnicas já consolidadas como DMC ou o clássico PID. A aplicação de aprendizagem por reforço, em especial o PPO, na área de

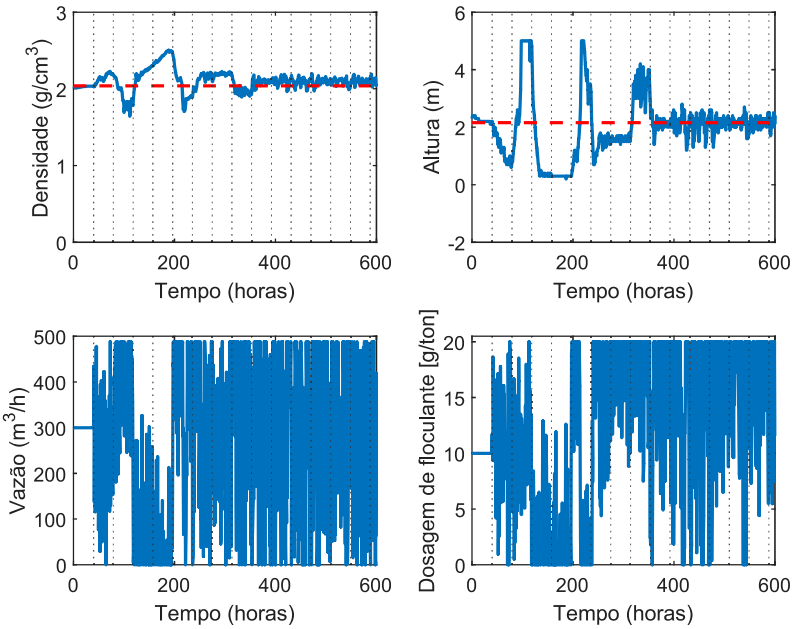


Figura 6.10: Sinais das malhas de controle PPO para (a) densidade na descarga (b) altura da interface (c) vazão na descarga e (d) dosagem de floculante, variando-se a densidade da polpa de alimentação.

Fonte: O autor.

controle de processos, ainda depende de um melhor entendimento dos seus parâmetros e de sua implementação, no entanto esta é uma área que possui muito espaço para evoluir, e sua característica de imunidade a distúrbios, alta abstração e generalização tornam esta uma técnica promissora para o futuro. Dada a liberdade para se trabalhar com tais algoritmos, uma melhor parametrização e até abordagens diferentes para a função de recompensa ou desenvolvimento do código poderiam levar a resultados melhores. Outro ponto a se considerar, é sua utilização para definir os parâmetros ótimos de um controlador convencional como o PID ao invés de ser utilizado diretamente como controlador principal de um sistema.

7. Conclusão

Neste trabalho buscou-se desenvolver uma nova alternativa, ainda pouco explorada, de controle para processos de mineração. O crescente avanço das técnicas de aprendizagem por reforço indicam que tais técnicas são capazes de suprir as necessidades de controle e, talvez, até superar as aplicações clássicas. Com essa motivação, foi utilizada a técnica de PPO para realizar o controle de um espessador convencional.

A partir de um simulador do processo, esperava-se controlar a densidade e a altura da interface de sólidos de forma a obter maior eficiência do processo de sedimentação e separação sólido-líquido, além de melhorar a utilização dos floculantes. No entanto, apesar do controlador utilizando PPO ter sido capaz de manter as variáveis controladas próximas do valor de referência, o erro em regime permanente e a alta oscilação fazem com que os resultados encontrados não sejam adequados para aplicações reais. Sendo assim, não foi possível dar prosseguimento a aplicação do agente treinado no espessador.

O controlador baseado em aprendizagem por reforço é uma aplicação em que se tem muita liberdade e, atualmente, ainda não existe uma técnica que guie a escolha adequada dos parâmetros. Para processos contínuos que fazem uso de redes neurais, há ainda mais escolhas a serem feitas como a estrutura e a forma de treinamento dessas redes. Mesmo assim, foi possível ver boas características como a insensibilidade na presença de distúrbios e a possibilidade de alcançar economia no uso de floculantes, se comparado com a média necessária em aplicações utilizando PID e DMC. Devido a liberdade na definição da recompensa do sistema é possível extrapolar uma aplicação comum de controle levando a ganhos financeiros além dos próprios ganhos de processo. Futuramente, com uma melhor definição dos parâmetros, ou mesmo com novas abordagens como a utilização de outras técnicas de aprendizagem por reforço, ou ainda a utilização para otimizar um controlador PID ou equivalente, espera-se alcançar todo o potencial que a inteligência artificial tem a oferecer.

7.1. Propostas de Trabalhos Futuros

Utilizar vários atores rodando em paralelo de forma que a contribuição de cada um deles seja somada no fim de cada atualização do agente. Com isso é possível otimizar o tempo da simulação, sendo inclusive uma estratégia prevista na proposta inicial do PPO.

Desenvolver o agente em linguagem *Python*, onde há vários *frameworks* dedicados a aprendizagem por reforço já consolidados, evitando dificuldades com detalhes secundários de implementação do código que não são claros. Caso necessário, é mais interessante adaptar o simulador para rodar em *Python*, ou criar uma interface que permita a comunicação do *MATLAB* com essa linguagem.

Implementar uma técnica de *earlystopping* durante o treinamento das redes neurais, interrompendo o aprendizado em momento adequado de forma a evitar *overfitting* e poupar tempo.

Essa estratégia poderia ser decisiva nos casos em que um comportamento próximo do ideal foi alcançado, mas a atualização seguinte leva o sistema novamente a recompensas ruins.

Encontrar os melhores parâmetros a partir de um algoritmo de busca. Apesar de alguns dos parâmetros terem pouca influência no comportamento final do agente, outros como a estrutura das redes neurais e a função de recompensa utilizada, são pontos chave para o bom funcionamento do sistema.

Aplicar o controlador por PPO já treinado em espessadores reais, após alcançar resultados satisfatórios durante o treinamento. Seria importante testar a capacidade de adaptação do agente em diversos espessadores reais que são diferentes entre si, e também diferente dos modelos ideais e simplificados que são os simuladores utilizados para treinamento, testando assim a capacidade de de generalização de um controlador desenvolvido a partir de aprendizagem por reforço.

Referências Bibliográficas

ACHIAM, J. "Spinning Up in Deep Reinforcement Learning". 2018. Disponível em: https://spinningup.openai.com/en/latest/. Acesso em: 13 de setembro de 2023.

AGGARWAL, C. C., OTHERS. "Neural networks and deep learning", *Springer*, v. 10, n. 978, pp. 3, 2018.

ANDRYCHOWICZ, M., RAICHUK, A., STAŃCZYK, P., et al.. "What matters in on-policy reinforcement learning? A large-scale empirical study", *arXiv preprint arXiv:2006.05990*, 2020.

BRAGA, F., MONTENEGRO, L., RUSSO, J., et al.. "Avaliações De Alternativas Para Otimização Da Disposição De Rejeitos Do Minério De Ferro", *XXVIII Encontro Nacional de Tratamento de Minérios e Metalurgia Extrativa*, 2019.

BUŞONIU, L., DE BRUIN, T., TOLIĆ, D., et al.. "Reinforcement learning for control: Performance, stability, and deep approximators", *Annual Reviews in Control*, v. 46, pp. 8–28, 2018.

CHARPENTIER, A., ELIE, R., REMLINGER, C. "Reinforcement learning in economics and finance", *Computational Economics*, pp. 1–38, 2021.

DALAL, M., PATHAK, D., SALAKHUTDINOV, R. R. "Accelerating robotic reinforcement learning via parameterized action primitives", *Advances in Neural Information Processing Systems*, v. 34, 2021.

DE ÁVILA FONSECA, H. D. G., ALEXANDRINO, J. S., FERREIRA, T. E. D. "Metodologias de disposição de rejeitos de minério de ferro para substituir as barragens de rejeitos", *Profiscientia*, , n. 12, pp. 54–72, 2019.

DHARIWAL, P., HESSE, C., KLIMOV, O., et al.. "OpenAI Baselines". 2017. Disponível em: https://github.com/openai/baselines. Acesso em: 13 de setembro de 2023.

HALLÉN, M., ÅSTRAND, M., SIKSTRÖM, J., et al.. "Reinforcement Learning for Grinding Circuit Control in Mineral Processing". Em: *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 488–495. IEEE, 2019. doi: 10.1109/ETFA.2019.8869212.

JIANG, J.-R. "An improved Cyber-Physical Systems architecture for Industry 4.0 smart factories". Em: *2017 International Conference on Applied System Innovation (ICASI)*, pp. 918–920, 2017. doi: 10.1109/ICASI.2017.7988589.

JIANG, Y., FAN, J., CHAI, T., et al.. "Dual-rate operational optimal control for flotation industrial process with unknown operational model", *IEEE Transactions on Industrial Electronics*, v. 66, n. 6, pp. 4587–4599, 2018.

KINGMA, D. P., BA, J. "Adam: A Method for Stochastic Optimization", *CoRR*, v. abs/1412.6980, 2014. Disponível em: <https://api.semanticscholar.org/CorpusID:6628106>.

KIRAN, B. R., SOBH, I., TALPAERT, V., et al.. "Deep Reinforcement Learning for Autonomous Driving: A Survey", *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–18, 2021. doi: 10.1109/TITS.2021.3054625.

LI, Y. "Reinforcement learning applications", *arXiv preprint arXiv:1908.06973*, 2019.

LUZ, A. A. B. D., LINS, F. A. F. "Introdução ao tratamento de minérios". CETEM/MCTIC, 2018.

MCCOY, J., AURET, L. "Machine learning applications in minerals processing: A review", *Minerals Engineering*, v. 132, pp. 95–109, 2019. ISSN: 0892-6875. doi: https://doi.org/10.1016/j.mineng.2018.12.004. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0892687518305430>.

MESQUITA, P. P. D., DE CARVALHO, P. S. L., OGANDO, L. D. "Desenvolvimento e inovação em mineração e metais", *BNDES Setorial*, , n. 43, pp. 359–361, mar. 2016.

NIAN, R., LIU, J., HUANG, B. "A review on reinforcement learning: Introduction and applications in industrial process control", *Computers & Chemical Engineering*, v. 139, pp. 106886, 2020.

PEREIRA, A. M. F. *Simulação Dinâmica e Controle de Nível da Interface e Concentração de Sólidos na Descarga de Espessadores Sujeitos à Adição de Floculante*. Tese de Mestrado, Universidade Federal de Ouro Preto (UFOP), Ouro Preto, MG, 2021.

PEREIRA, A. M., MARTINS, W. T., MOREIRA, V. S., et al.. "Aplicação de Controle PI e DMC Multivariável em Espessadores de Minério de Ferro". Em: *Congresso Brasileiro de Automática-CBA*, 2020.

PERERA, A., KAMALARUBAN, P. "Applications of reinforcement learning in energy systems", *Renewable and Sustainable Energy Reviews*, v. 137, pp. 110618, 2021. doi: 10.1016/j.rser.2020.110618.

RAVICHANDIRAN, S. *Hands-on reinforcement learning with Python: master reinforcement and deep reinforcement learning using OpenAI gym and tensorFlow*. Packt Publishing Ltd, 2018.

REYES, F., HILDEN, M., YAHYAEI, M., et al.. "Reinforcement learning control of a SAG mill grinding circuit: first impressions and implications for process control". Em: *XXX International Mineral Processing Congress, Cape Town, South Africa*, 2020.

SCHULMAN, J., WOLSKI, F., DHARIWAL, P., et al.. "Proximal Policy Optimization Algorithms", *CoRR*, v. abs/1707.06347, 2017. Disponível em: <http://arxiv.org/abs/1707.06347>.

SHIN, J., BADGWELL, T. A., LIU, K.-H., et al.. "Reinforcement Learning–Overview of recent progress and implications for process control", *Computers & Chemical Engineering*, v. 127, pp. 282–294, 2019.

SHIPMAN, W. J. "Learning to Tune a Class of Controllers with Deep Reinforcement Learning", *Minerals*, v. 11, n. 9, pp. 989, 2021.

SPIELBERG, S., GOPALUNI, R., LOEWEN, P. "Deep reinforcement learning approaches for process control". Em: *2017 6th international symposium on advanced control of industrial processes (AdCONIP)*, pp. 201–206. IEEE, 2017.

SUTTON, R. S., BARTO, A. G. *Reinforcement learning: An introduction*. London, England, MIT press, 2018.

WANG, H.-N., LIU, N., ZHANG, Y.-Y., et al.. "ADeep reinforcement learning: a survey", *Frontiers of Information Technology Electronic Engineering*, v. 21, n. 12, pp. 1726–1744, 2020. doi: 10.1631/FITEE.1900533.

YU, C., LIU, J., NEMATI, S., et al.. "Reinforcement Learning in Healthcare: A Survey", *ACM Computing Surveys*, v. 55, n. 1, nov 2021. ISSN: 0360-0300. doi: 10.1145/3477600.