UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO
UNIDADE ACADÊMICA DE BELO JARDIM
BACHARELADO EM ENGENHARIA DA COMPUTAÇÃO

PEDRO HENRIQUE DE ALMEIDA SANTOS

SINTONIA ADAPTATIVA DE CONTROLADORES PID: ABORDAGEM VIA APRENDIZAGEM POR REFORÇO PARA PROCESSO INDUSTRIAL SIMULADO

Belo Jardim
2026

PEDRO HENRIQUE DE ALMEIDA SANTOS

SINTONIA ADAPTATIVA DE CONTROLADORES PID: ABORDAGEM VIA APRENDIZAGEM POR REFORÇO PARA PROCESSO INDUSTRIAL SIMULADO

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Engenharia da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharelado em Engenharia da Computação.

Orientador: D.Sc. Moisés Tavares da Silva

Belo Jardim
2026

Dados Internacionais de Catalogação na Publicação
Sistema Integrado de Bibliotecas da UFRPE
Bibliotecário(a): Nubia Matos – CRB-4 2234

S237s Santos, Pedro Henrique de Almeida.
Sintonia adaptativa de controladores PID : abordagem via aprendizagem por reforço para processo industrial simulado / Pedro Henrique de Almeida Santos. – Belo Jardim, 2026.
50 f.; il.

Orientador(a): Moisés Tavares da Silva.

Trabalho de Conclusão de Curso (Graduação) – Universidade Federal Rural de Pernambuco, Unidade Acadêmica de Belo Jardim - UABJ, Bacharelado em Engenharia da Computação, Belo Jardim, BR-PE, 2026.

Inclui referências.

1. Aprendizagem por reforço. 2. Sintonia adaptativa. 3. Controladores PID. 4. Reator CSTR 5. Algoritmo ARS. I. Silva, Moisés Tavares da, orient. II. Título

CDD 621.3

Pedro Henrique de Almeida Santos

SINTONIA ADAPTATIVA DE CONTROLADORES PID: ABORDAGEM VIA APRENDIZAGEM POR REFORÇO PARA PROCESSO INDUSTRIAL SIMULADO

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Engenharia da Computação da Universidade Federal Rural de Pernambuco, como requisito parcial para obtenção do título de Bacharelado em Engenharia da Computação.

Aprovado em 17/07/2026.

BANCA EXAMINADORA

D.Sc. Moisés Tavares da Silva
Universidade Federal Rural de Pernambuco

Prof. D.Sc. Anderson Pinheiro Cavalcanti
Universidade Federal Rural de Pernambuco

Prof. D.Sc. Egydio Tadeu Gomes Ramos
(Examinador Externo)
Universidade Federal de Campina Grande

RESUMO

Este trabalho descreve o desenvolvimento e a implementação de uma estratégia de sintonia adaptativa de controladores PID (Proporcional-Integral-Derivativo) aplicada ao processo industrial simulado de um reator químico contínuo de mistura perfeita (*Continuous Stirred-Tank Reactor* - CSTR). A abordagem utilizou o algoritmo de aprendizagem por reforço denominado busca aleatória aumentada (*Augmented Random Search* - ARS), que se destaca pela baixa complexidade computacional e eficiência na aplicação de sistemas contínuos sem a exigência de modelos matemáticos exatos. O diferencial do trabalho consistiu na proposição de uma nova função de recompensa que integra simultaneamente o desempenho do rastreamento da referência e o esforço da ação de controle, buscando equilibrar a precisão do sistema com a redução do esforço e do desgaste dos atuadores. As simulações foram realizadas em ambiente computacional utilizando a linguagem Python e bibliotecas de código aberto para modelagem e análise dinâmica. Os resultados obtidos demonstraram que a metodologia proposta superou métodos tradicionais e outras abordagens de aprendizagem por reforço da literatura, alcançando melhores índices de desempenho tanto em termos de erro acumulado quanto de variabilidade da entrada. O controlador sintonizado apresentou respostas mais rápidas, com menor sobressinal e estabilidade superior frente a sucessivas mudanças de referência e perturbações de carga. Neste contexto, foi observado que a técnica oferece uma alternativa robusta e eficiente para a sintonia automática de controladores industriais em processos não lineares, validando a eficácia da nova função de recompensa desenvolvida.

Palavras-chave: Aprendizagem por reforço; Sintonia adaptativa; Controladores PID; Reator CSTR; Algoritmo ARS.

ABSTRACT

This work describes the development and implementation of an adaptive tuning strategy for PID (Proportional-Integral-Derivative) controllers applied to the simulated industrial process of a Continuous Stirred-Tank Reactor (CSTR). The approach employed the reinforcement learning algorithm known as Augmented Random Search (ARS), which is notable for its low computational complexity and efficiency in continuous systems, without requiring exact mathematical models. A key contribution of this work is the proposal of a new reward function that simultaneously integrates reference tracking performance and control action effort, aiming to balance system precision with the reduction of actuator effort and wear. Simulations were conducted in a computational environment using Python and open-source libraries for modeling and dynamic analysis. The results demonstrated that the proposed methodology outperformed traditional methods and other reinforcement learning approaches found in the literature, achieving superior performance metrics regarding both accumulated error and input variability. The tuned controller exhibited faster responses, reduced overshoot, and greater stability in the face of successive reference changes and load disturbances. In this context, the technique proved to be a robust and efficient alternative for the automatic tuning of industrial controllers in nonlinear processes, validating the effectiveness of the newly developed reward function.

Keywords: Reinforcement learning; Adaptive tuning; PID controllers; CSTR; ARS algorithm.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 Introdução

O controle de processos industriais é uma área fundamental da engenharia, pois está diretamente relacionado à segurança, à eficiência operacional e à qualidade dos produtos. Entre as estratégias de controle mais utilizadas na indústria, destacam-se os controladores Proporcional-Integral-Derivativo (PID), amplamente empregados devido à sua simplicidade, robustez e facilidade de implementação (BORASE et al., 2021). Estes controladores são aplicados em diversos processos, como o controle de temperatura, nível, vazão e pressão, sendo essenciais para a manutenção de condições operacionais adequadas (CAMPOS; TEIXEIRA, 2006). No entanto, o desempenho de controladores PID depende fortemente de uma sintonia adequada de seus parâmetros; quando imprecisa, essa sintonia pode resultar em oscilações excessivas, elevado erro em regime permanente e respostas lentas, comprometendo a estabilidade e a eficiência do processo. Como consequência, podem ocorrer perdas na qualidade do produto, aumento dos custos operacionais e até danos aos equipamentos (ÅSTRÖM; HÄGGLUND, 2006).

Tradicionalmente, a sintonia de controladores PID é realizada por métodos heurísticos ou empíricos, como Ziegler-Nichols e Cohen-Coon (SOMEFUN; AKINGBADE; DAHUNSI, 2021). Embora amplamente utilizados, esses métodos apresentam limitações significativas quando aplicados a sistemas dinâmicos, não lineares ou sujeitos a variações de parâmetros e perturbações externas (SOMEFUN; AKINGBADE; DAHUNSI, 2021). Diante dessas limitações, tem crescido o interesse por técnicas baseadas em dados e métodos adaptativos capazes de ajustar os parâmetros do controlador de forma automática, reduzindo a dependência de modelos matemáticos precisos do processo (BITARÃES; SILVA; EUZÉBIO, 2022).

Nesse contexto, a aprendizagem por reforço *(Reinforcement Learning* – RL) tem se destacado como uma abordagem promissora para problemas de controle, especialmente aqueles que envolvem tomada de decisão sequencial em ambientes complexos (SUTTON; BARTO, 2018). Diferentemente dos métodos clássicos, o RL permite que um agente aprenda, por meio da interação com o ambiente, quais ações maximizam uma função de recompensa, dispensando o conhecimento prévio detalhado do modelo do sistema (SPIELBERG et al., 2019), (RUSSELL; NORVIG, 2022).

Este trabalho tem como objetivo estudar e aplicar uma abordagem de sintonia adaptativa de controladores PID baseada em aprendizagem por reforço no processo industrial simulado, em particular, o processo CSTR (do inglês *Continuous Stirred Tank Reactor*). A proposta visa avaliar o desempenho do controlador frente a diferentes condições de operação, buscando reduzir a dependência de métodos heurísticos e contribuir para o desenvolvimento de estratégias de controle mais robustas, eficientes e potencialmente aplicáveis a sistemas industriais reais.

1.1 Motivação

A crescente complexidade dos processos industriais modernos tem demandado estratégias de controle capazes de manter elevados níveis de desempenho mesmo diante de não linearidades, incertezas paramétricas e mudanças nas condições de operação. Embora os controladores PID permaneçam amplamente empregados na indústria devido à sua simplicidade e robustez, sua eficiência depende diretamente da qualidade da sintonia de seus parâmetros. Métodos tradicionais de sintonia, apesar de consolidados, apresentam limitações em processos sujeitos a dinâmicas variáveis, tornando necessária a busca por abordagens adaptativas que realizem esse ajuste de forma automática.

Nesse cenário, os algoritmos de aprendizado por reforço surgem como uma alternativa promissora, uma vez que possibilitam o aprendizado de estratégias de controle por meio da interação com o ambiente, dispensando modelos matemáticos precisos do processo. Dentre essas técnicas, o algoritmo ARS (do inglês *Augmented Random Search*) destaca-se por sua simplicidade de implementação, reduzido custo computacional e eficiência na otimização de problemas contínuos, características que favorecem sua aplicação em sistemas de controle industrial (MANIA; GUY; RECHT, 2018).

Entretanto, observa-se que o desempenho do aprendizado por reforço está fortemente associado à definição da função de recompensa utilizada durante o treinamento do agente. Diversos dos trabalhos encontrados na literatura priorizam apenas a minimização do erro de rastreamento, negligenciando aspectos relacionados ao esforço de controle e às variações do sinal de atuação. Essa abordagem pode resultar em controladores que apresentam boa precisão, porém exigem ações de controle excessivamente agressivas, reduzindo a vida útil dos atuadores e aumentando o consumo de energia.

Dessa forma, este trabalho é motivado pela necessidade de investigar uma estratégia de sintonia adaptativa que considere simultaneamente a qualidade do controle e o esforço de atuação do controlador. Além do desenvolvimento da nova função de recompensa, a escolha do processo CSTR (ANTONELLI; ASTOLFI, 2003) como estudo de caso é motivada por sua elevada não linearidade e por representar um problema clássico de controle de processos. A comparação dos resultados obtidos com métodos de sintonia já presentes na literatura permite avaliar quantitativamente os benefícios da abordagem proposta, contribuindo para o avanço da aplicação de algoritmos de aprendizado por reforço em problemas reais de controle industrial.

1.2 Objetivos

1.2.1 Geral

Desenvolver e implementar um algoritmo de aprendizado por reforço baseado no método ARS para a sintonia adaptativa de controladores PID aplicados ao processo CSTR, propondo uma nova função de recompensa capaz de melhorar o desempenho do sistema de

controle.

1.2.2 Específicos

- Realizar uma revisão da literatura sobre algoritmos de aprendizado por reforço aplicados à sintonia de controladores em processos industriais;
- Implementar o algoritmo ARS para a sintonia adaptativa de controladores PID utilizando a linguagem Python;
- Desenvolver e incorporar uma nova função de recompensa baseada nos índices de desempenho IAE e SII para o treinamento do agente de aprendizado por reforço;
- Aplicar o algoritmo ARS ao processo CSTR em diferentes condições de operação;
- Comparar o desempenho do controlador PID proposto com métodos presentes na literatura;
- Avaliar a influência da nova função de recompensa na qualidade da sintonia do controlador PID e na resposta dinâmica do processo CSTR.

1.3 Justificativas

O processo de sintonia de controladores PID continua sendo um dos principais desafios da engenharia de controle, especialmente em processos industriais não lineares e sujeitos a variações dinâmicas, como o CSTR (ANTONELLI; ASTOLFI, 2003). Embora métodos clássicos de sintonia apresentem resultados satisfatórios em condições específicas de operação, seu desempenho pode ser comprometido diante de mudanças nas características do processo, exigindo estratégias capazes de adaptar automaticamente os parâmetros do controlador.

O desempenho de algoritmos de aprendizado por reforço depende diretamente da função de recompensa utilizada durante o treinamento. Em muitos trabalhos, a recompensa considera apenas o erro de rastreamento, desconsiderando aspectos importantes relacionados ao esforço de controle. Dessa forma, torna-se relevante investigar funções de recompensa que incorporem simultaneamente critérios de desempenho e de esforço de atuação, proporcionando uma sintonia mais equilibrada e adequada às necessidades práticas da indústria.

Desta forma, a aplicação do algoritmo ARS ao processo CSTR justifica-se pela elevada não linearidade desse sistema, frequentemente utilizado como referência para avaliação de técnicas avançadas de controle. A comparação dos resultados obtidos com metodologias presentes na literatura, como as propostas por (HEDENGREN, 2021) e (BITARÃES; SILVA; EUZÉBIO, 2022), permite avaliar quantitativamente a eficácia da estratégia desenvolvida, contribuindo para a validação da abordagem proposta.

1.4 Revisão de Literatura

Diversos estudos têm demonstrado a viabilidade da aplicação do aprendizado por reforço no controle de processos industriais (BORASE et al., 2021),(SPIELBERG et al., 2019), (LUZ; CAARLS, 2025),(JESAWADA et al., 2022). Esta aplicação abrange sistemas de nível de tanques (OLIVEIRA et al., 2023; LIMA, 2018; MARIZ, 2024; MATE et al., 2023), reatores químicos (CASSOL et al., 2018; BITARãES, 2022), colunas de destilação (JÚNIOR; EULER et al., 2023) e microgrids (ESMAEILI et al., 2017), apresentando desempenho comparável ou superior aos controladores PID tradicionais (CHEN et al., 2025; REGO; ARAÚJO, 2021; LIMA et al., 2013; LAKHANI; CHOWDHURY; LU, 2021).

A literatura recente também evidencia avanços na integração entre aprendizado por reforço e controladores PID, com o objetivo de unir a interpretabilidade e robustez do PID à capacidade adaptativa do RL (JESAWADA et al., 2022; CHEN et al., 2025). Abordagens híbridas baseadas em estruturas *Ator-Critic* (CHEN et al., 2025; LIMA, 2018; MATE et al., 2023; YU et al., 2025) e métodos de ajuste automático (LAKHANI; CHOWDHURY; LU, 2021), têm sido propostas para a sintonia de controladores PID, mostrando melhorias no erro de rastreamento (BITARãES, 2022; DING et al., 2023), tempo de acomodação (BITARãES, 2022; ESMAEILI et al., 2017; DING et al., 2023) e estabilidade do sistema (LAKHANI; CHOWDHURY; LU, 2021). Entretanto, muitos desses métodos apresentam elevada complexidade computacional, alto custo de treinamento e grande demanda por dados, o que dificulta sua aplicação prática em ambientes industriais (MATE et al., 2023; DING et al., 2023).

Dentre as diversas abordagens de aprendizagem por reforço, destaca-se a necessidade de investigar algoritmos que conciliam bom desempenho com menor complexidade computacional, uma vez que muitas técnicas avançadas demandam elevado custo de treinamento e grandes volumes de dados. Nesse contexto, o algoritmo *Augmented Random Search* (ARS), proposto por Mania, Guy e Recht (2018), apresenta-se como uma alternativa promissora, por sua estrutura simples e eficiência na exploração de espaços de alta dimensionalidade, sendo capaz de obter políticas de controle eficazes sem a necessidade de extensos conjuntos de dados de treinamento. Resultados reportados na literatura indicam que a aplicação do ARS à sintonia de controladores PID em processos industriais simulados, como o CSTR, pode proporcionar reduções significativas no erro de controle e melhorias na estabilidade do sistema quando comparada a métodos tradicionais de sintonia (BITARãES, 2022).

Apesar desses avanços, a aplicação de algoritmos de menor complexidade para a sintonia adaptativa de controladores PID em processos industriais não lineares e multivariáveis ainda representa um desafio relevante, especialmente diante das variações operacionais e das incertezas inerentes a esses sistemas. Dessa forma, torna-se pertinente aprofundar a investigação de estratégias de controle adaptativas que aliam robustez, eficiência e viabilidade computacional.

2 Fundamentação Teórica

2.1 Fundamentos da Aprendizagem por Reforço

Esta fundamentação teórica sobre os fundamentos da aprendizagem por reforço (RL) baseia-se nos conceitos apresentados por Russell e Norvig (2022), que descrevem a aprendizagem por reforço como um paradigma de aprendizado voltado para problemas de tomada de decisão sequencial, nos quais um agente aprende por meio da interação contínua com um ambiente.

A aprendizagem por reforço é uma área da Inteligência Artificial em que um agente aprende a agir em um ambiente por meio de interações sucessivas, recebendo recompensas que indicam a qualidade de suas ações. Diferentemente da aprendizagem supervisionada, o aprendizado ocorre a partir da experiência adquirida durante a interação com o ambiente, e não de exemplos previamente rotulados (RUSSELL; NORVIG, 2022).

Uma distinção importante é que o agente não está apenas resolvendo um problema estático, mas encontra-se inserido em um Processo de Decisão de Markov (*Markov Decision Process* – MDP), frequentemente sem conhecer previamente o modelo de transição dos estados nem a função de recompensa do ambiente (RUSSELL; NORVIG, 2022).

O MDP é o formalismo matemático utilizado para representar problemas de decisão sequencial. Segundo Russell e Norvig (2022), um MDP é definido por cinco elementos principais:

- Estados (S): conjunto de situações possíveis em que o sistema pode se encontrar;
- Ações (A): conjunto de ações disponíveis ao agente em cada estado;
- Modelo de Transição ($P(s'|s,a)$): probabilidade de o sistema evoluir para o estado s' após a execução da ação a no estado s;
- Função de Recompensa ($R(s,a,s')$): retorno imediato recebido após cada transição;
- Fator de Desconto (γ): parâmetro pertencente ao intervalo $0 \leq \gamma < 1$, responsável por ponderar a importância das recompensas futuras em relação às imediatas.

Quando o fator de desconto assume valores próximos de zero, o agente privilegia recompensas imediatas. Em contrapartida, valores próximos de um fazem com que recompensas futuras tenham maior peso no processo de tomada de decisão (RUSSELL; NORVIG, 2022).

2.1.1 Utilidade e Políticas

Uma política, representada por π, especifica a ação que o agente deve executar em cada estado alcançável. O objetivo da aprendizagem por reforço consiste em determinar uma política ótima, denotada por π^*, capaz de maximizar a utilidade esperada ao longo do tempo (RUSSELL; NORVIG, 2022).

Em problemas de horizonte infinito, ou seja, que não possuem um número pré-determinado de iterações, a utilidade de um histórico é normalmente expressa como a soma das recompensas descontadas:

$$U = \sum_{t=0}^{\infty} \gamma^t R_t. \tag{1}$$

A Equação de Bellman estabelece a relação recursiva entre os estados, mostrando que a utilidade de um estado corresponde à recompensa imediata somada à utilidade esperada dos estados futuros, considerando que o agente executa ações ótimas (RUSSELL; NORVIG, 2022).

2.1.2 Funções de Valor e Função Q

A função de utilidade $U(s)$ representa o retorno esperado de um estado quando o agente segue uma política ótima. Por outro lado, a função ação-valor, conhecida como função $Q(s,a)$, representa a utilidade esperada de executar uma ação específica em determinado estado e, posteriormente, seguir uma política ótima. Sua principal vantagem é permitir que o agente selecione diretamente a melhor ação (a^*) por meio da expressão sem necessidade de conhecer explicitamente o modelo de transição do ambiente (RUSSELL; NORVIG, 2022).

$$a^* = \arg\max_a Q(s,a). \tag{2}$$

2.1.3 Tipos de Aprendizado

Os métodos de aprendizagem por reforço podem ser classificados em aprendizado passivo e aprendizado ativo. No aprendizado passivo, o agente segue uma política fixa e busca apenas estimar a utilidade dos estados. Entre as principais técnicas encontram-se a Estimativa de Utilidade Direta, a Programação Dinâmica Adaptativa (PDA) e o Aprendizado por Diferença Temporal (*Temporal Difference* – TD) (RUSSELL; NORVIG, 2022).

No aprendizado ativo, além de estimar as utilidades, o agente deve decidir quais ações executar. Surge, então, o clássico dilema entre exploração, que consiste em experimentar novas ações para adquirir conhecimento sobre o ambiente, e explotação, que utiliza o conhecimento adquirido para maximizar as recompensas esperadas (RUSSELL; NORVIG, 2022).

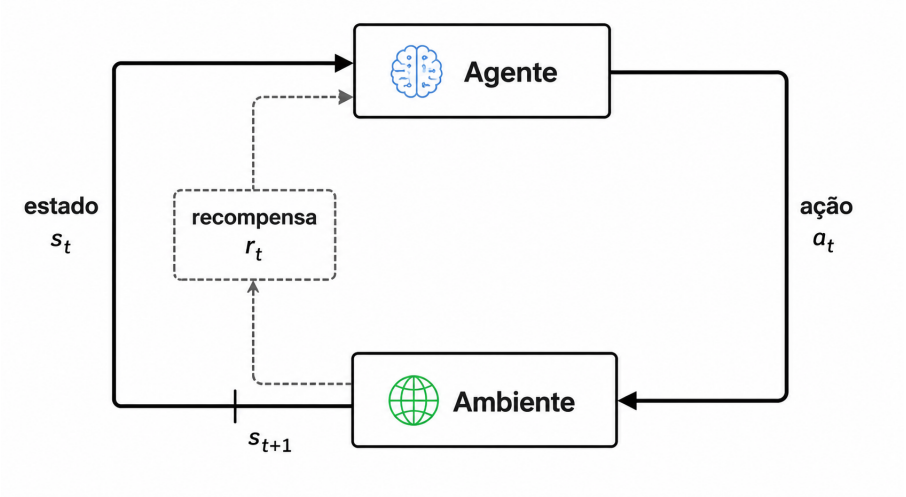
Entre os algoritmos mais difundidos destacam-se o Q-Learning e o *State-Action-Reward-State-Action* (SARSA). O Q-Learning é um algoritmo *off-policy*, pois aprende a política ótima independentemente da estratégia de exploração utilizada. O SARSA, por sua vez, é um algoritmo *on-policy*, aprendendo diretamente a partir da política efetivamente executada pelo agente (RUSSELL; NORVIG, 2022).

2.1.4 Elementos básicos do aprendizado por reforço

A Figura 1 ilustra o ciclo de funcionamento da estrutura básica do RL. Inicialmente, o ambiente fornece um estado ao agente. Em seguida, o agente seleciona uma ação de acordo com sua política. Após a execução da ação, o ambiente atualiza seu estado e retorna uma recompensa juntamente com o novo estado. Esse processo é repetido sucessivamente durante

todas as épocas de treinamento, permitindo que o agente aperfeiçoe continuamente sua política de controle (SUTTON; BARTO et al., 1998). Dessa forma, a aprendizagem por reforço é composta pelos elementos descritos a seguir.

Figura 1 – Estrutura Básica de Aprendizagem por Reforço



Fonte: Autoria própria, 2026.

2.1.4.1 Agente

O agente corresponde ao algoritmo de aprendizagem por reforço responsável por tomar decisões durante o processo de controle. Sua função consiste em observar o estado atual do ambiente, selecionar uma ação e aprender, por meio das recompensas recebidas, quais decisões produzem os melhores resultados ao longo das interações.

2.1.4.2 Ambiente

O ambiente representa o processo físico a ser controlado. Neste trabalho, o ambiente corresponde ao modelo computacional do CSTR utilizado nas simulações. É o ambiente que recebe as ações executadas pelo agente, atualiza a dinâmica do sistema e retorna um novo estado juntamente com uma recompensa.

2.1.4.3 Estados

Os estados representam as informações disponíveis ao agente em cada instante de tempo. Essas informações descrevem a condição atual do processo e são utilizadas como entrada para a tomada de decisão. A escolha adequada das variáveis de estado é fundamental para que o agente consiga identificar corretamente a situação do ambiente.

2.1.4.4 Ações

As ações correspondem aos comandos de controle que podem ser aplicados pelo agente ao ambiente. Em cada estado, o agente escolhe uma ação pertencente ao conjunto de ações disponíveis, buscando conduzir o sistema para condições que proporcionem maior recompensa.

2.1.4.5 Política

A política, representada por π, define a estratégia utilizada pelo agente para selecionar ações em função dos estados observados. Durante o treinamento, essa política é continuamente atualizada com base nas experiências adquiridas, tornando-se progressivamente mais eficiente.

2.1.4.6 Recompensa

A recompensa é um valor numérico fornecido pelo ambiente após a execução de cada ação. Sua função é avaliar o desempenho da decisão tomada pelo agente. Recompensas positivas incentivam comportamentos desejáveis, enquanto recompensas negativas penalizam ações que afastam o sistema do objetivo estabelecido.

2.1.4.7 Iteração

A iteração, representa um único ciclo de interação dentro de um episódio. Em cada iteração, o agente observa o estado atual, escolhe uma ação de acordo com sua política, executa essa ação no ambiente, recebe uma recompensa e observa o novo estado. Assim, um episódio é composto por diversas iterações.

2.1.4.8 Episódio

O episódio, corresponde a uma sequência completa de interações entre o agente e o ambiente, iniciando em um estado inicial e terminando quando uma condição de parada é atingida, como o alcance do objetivo ou um limite máximo de tempo. Durante um episódio, o agente observa o estado do ambiente, seleciona ações, recebe recompensas e transita entre estados.

2.1.4.9 Treinamento

É o processo de aprendizado do agente por meio da interação repetida com o ambiente. Durante o treinamento, o agente alterna entre exploração e explotação, ajustando sua política com base nas recompensas obtidas até encontrar uma estratégia de controle eficiente.

2.1.4.10 Exploração

É a fase em que o agente testa diferentes ações ou estratégias para adquirir conhecimento sobre o ambiente. O objetivo é descobrir soluções que possam proporcionar melhores recompensas, mesmo que inicialmente apresentem desempenho inferior.

2.1.4.11 Explotação

É a fase em que o agente utiliza o conhecimento adquirido durante o treinamento para selecionar as ações que apresentam melhor desempenho conhecido, buscando maximizar a recompensa acumulada.

2.1.5 Modelo de aprendizagem por reforço

Durante o processo de treinamento, o agente realiza uma sequência de interações com o ambiente. Inicialmente, o ambiente fornece um estado atual ao agente, representando a condição do sistema naquele instante. A partir desse estado, o agente seleciona uma ação dentre as ações disponíveis de acordo com sua política de decisão (SUTTON; BARTO et al., 1998).

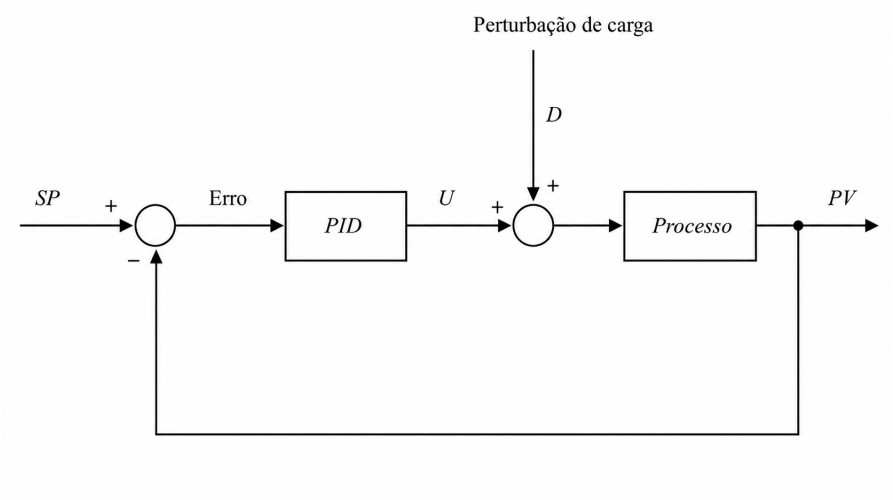
Após a execução da ação, o ambiente responde atualizando seu estado e fornecendo uma recompensa ao agente. Essa recompensa representa uma medida quantitativa da qualidade da ação executada, indicando se ela contribuiu ou não para alcançar o objetivo do controle. Em seguida, o novo estado é novamente enviado ao agente, reiniciando o ciclo de interação.

Esse processo é repetido durante todas as iterações de treinamento, permitindo que o agente aprenda gradativamente quais ações produzem melhores resultados em cada situação encontrada. Ao final do treinamento, espera-se que o agente tenha aprendido uma política capaz de selecionar ações que maximizem a recompensa acumulada e, consequentemente, o desempenho do sistema controlado (RUSSELL; NORVIG, 2022).

2.2 Aspectos dos controladores PID

O controlador PID é o algoritmo de controle mais utilizado na indústria, sendo composto por três ações fundamentais: Proporcional (P), Integral (I) e Derivativa (D) (CAMPOS; TEIXEIRA, 2006). O objetivo principal de um controlador é manter as variáveis de um processo em valores pré-determinados, seguindo etapas que envolvem a medição via sensor, comparação com a referência, cálculo do sinal de correção e envio do ajuste ao atuador (CAMPOS; TEIXEIRA, 2006). Dentre as vantagens do controlador PID, destacam-se a quantidade reduzida de parâmetros para sintonizar, a facilidade de associação entre esses parâmetros e o desempenho, além de sua robustez e ampla disponibilidade em equipamentos industriais (ÅSTRÖM; HÄGGLUND, 2006).

Figura 2 – Sistema de controle em malha fechada com perturbação de carga.



Fonte : adaptado de (CAMPOS; TEIXEIRA, 2006)

O funcionamento de um sistema em malha fechada, também conhecido como controle por retroalimentação, ilustrado na Figura 2, baseia-se no monitoramento constante da variável que se deseja controlar para garantir que ela permaneça no objetivo estabelecido (CAMPOS; TEIXEIRA, 2006). Diferente da malha aberta, onde a ação de controle não depende do resultado atual. De acordo com Campos e Teixeira (2006), na malha fechada o processo segue um ciclo dinâmico descrito pelas etapas abaixo:

- Medição e Comparação: Um sensor realiza a leitura da variável de processo (PV) e envia essa informação ao controlador. Este componente compara o valor medido com o valor de referência desejado, chamado de setpoint (SP). A diferença calculada entre esses dois valores é o que define o erro do sistema.
- Ação de Correção: Com base no erro identificado, o controlador utiliza um algoritmo (como o PID) para calcular e enviar um sinal de ajuste ao elemento final de controle, como uma válvula ou um motor. O objetivo dessa intervenção é manipular o processo para eliminar o desvio e aproximar a variável controlada do valor desejado.
- Compensação de Perturbações: Uma das principais capacidades do sistema de malha fechada é a sua habilidade de reagir automaticamente a perturbações externas ou mudanças nas condições internas do sistema, corrigindo a trajetória do processo sem a necessidade de intervenção manual constante. Esse comportamento é ilustrado na Figura 2, em que a perturbação de carga atua diretamente entre o controlador e o processo, sendo compensada pela ação do controlador por meio da realimentação.

Esses sistemas podem operar em dois modos principais: o regulatório, que busca manter a variável fixa em um ponto mesmo diante de interferências, e o servo, onde a variável controlada deve acompanhar uma referência que muda ao longo do tempo, seguindo uma trajetória específica (CAMPOS; TEIXEIRA, 2006). Neste trabalho é usada a equação do controlador PID paralelo, onde as ações proporcional, integral e derivativa são:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)\, d\tau + K_d \frac{de(t)}{dt} \tag{3}$$

onde:

- $u(t)$ é a variável de controle (sinal de controle ou variável manipulada);
- $e(t)$ é o erro de controle, correspondente à diferença entre o valor de referência e a variável de processo;
- K_p é o ganho proporcional do controlador;
- K_i é o ganho integral do controlador;
- K_d é o ganho derivativo do controlador;

Embora seja mais eficaz para manter a precisão, o controle em malha fechada pode introduzir instabilidades ou oscilações se os parâmetros do controlador PID não estiverem devidamente ajustados (sintonizados) (CAMPOS; TEIXEIRA, 2006). Abaixo, cada uma das ações que compõem esse controlador são descritas:

- Ação Proporcional: Esta componente gera um sinal de correção que é diretamente proporcional ao erro atual do sistema, ou seja, à diferença entre o valor desejado e o valor medido. O parâmetro principal aqui é o ganho proporcional (K_p), que determina a intensidade da resposta; ganhos mais elevados resultam em ações mais rápidas, mas podem levar à instabilidade. Uma característica importante dessa ação é que, na maioria dos processos industriais, ela não consegue eliminar completamente o desvio em regime permanente, mantendo um erro residual conhecido como "*off-set*" (CAMPOS; TEIXEIRA, 2006).
- Ação Integral: O objetivo fundamental desta ação é eliminar o erro de regime permanente que a ação proporcional não consegue resolver. Ela atua baseada no acúmulo dos erros ao longo do tempo (a integral do erro), fazendo com que o controlador continue ajustando a saída enquanto houver qualquer diferença entre a variável de processo e o valor de referência. O ajuste é feito através do ganho integral (K_i), e embora garanta que o processo atinja exatamente o ponto desejado, uma ação integral muito agressiva pode tornar o sistema mais oscilatório (CAMPOS; TEIXEIRA, 2006).
- Ação Derivativa: Esta parte do controlador atua de forma antecipatória, observando a velocidade com que o erro está mudando (a derivada do erro). Ao perceber uma tendência de aumento ou redução rápida do desvio, a ação derivativa tenta frear essa mudança antes mesmo que ela se concretize, o que ajuda a estabilizar o processo e a reduzir oscilações e sobressinais. É particularmente útil em processos lentos para acelerar a resposta, mas deve ser usada com cautela em sistemas com muito ruído na medição, pois pode amplificar esses distúrbios (CAMPOS; TEIXEIRA, 2006).

2.3 Índices de desempenho da malha fechada

A quantificação do desempenho de sistemas de controle em malha fechada é feita frequentemente através de índices de desempenho que integram a função do erro no tempo, permitindo uma análise numérica objetiva da qualidade do controle (CAMPOS; TEIXEIRA, 2006). Os índices mais utilizados na prática industrial baseiam-se na trajetória da variável controlada em relação ao seu valor de referência ao longo de uma janela de tempo. Abaixo seguem as expressões matemáticas para alguns desses critérios:

- Integral do Erro Absoluto (IAE): calcula a integral do valor absoluto do erro ao longo do tempo, atribuindo o mesmo peso para erros positivos e negativos. Esse índice fornece uma medida do erro acumulado do sistema, sendo amplamente utilizado para avaliar o desempenho global do controlador. O IAE é dado por:

$$IAE = \int_0^\infty |e(t)|\, dt \tag{4}$$

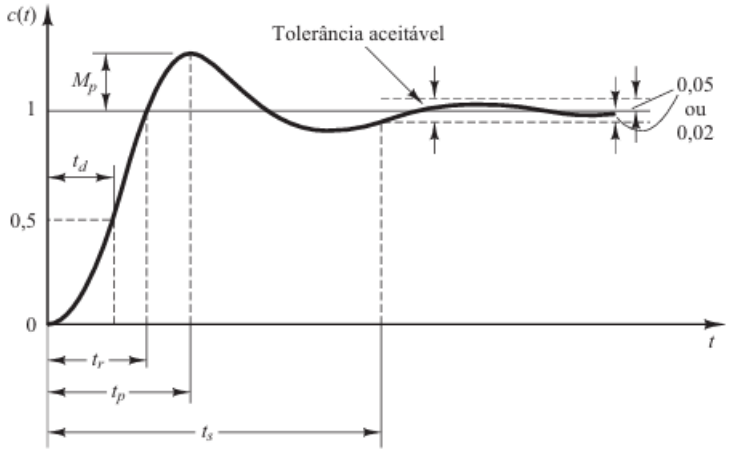
- Soma dos incrementos absolutos da entrada (SII): é uma métrica de desempenho utilizada para avaliar a variabilidade e o esforço da ação de controle sobre o processo (CAMPOS; TEIXEIRA, 2006). Diferente de índices como IAE ou ITAE, que focam no desvio da variável controlada em relação ao setpoint, o SII concentra-se na movimentação do atuador (como a abertura e fechamento de uma válvula) (CAMPOS; TEIXEIRA, 2006). O SII mede o quanto a variável manipulada se desloca para manter o controle, sendo um indicador direto do desgaste mecânico dos equipamentos e do consumo de energia do sistema (CAMPOS; TEIXEIRA, 2006). Um controle bem sintonizado busca um equilíbrio entre a rápida correção do erro (baixo IAE/ITAE) e a suavidade na movimentação da entrada (baixo SII). O SII é dado por:

$$\mathrm{SII} = \int_{t_i}^{T} \left| \frac{du(t)}{dt} \right| dt \tag{5}$$

2.4 Parâmetros da resposta do sistema em regime transitório

Os parâmetros de resposta em regime transitório constituem importantes indicadores do desempenho de sistemas de controle, permitindo avaliar características como rapidez, estabilidade e precisão. Esses parâmetros são obtidos a partir da resposta do sistema a um degrau unitário (OGATA, 2010) e estão ilustrados na Figura 3. Os principais indicadores são descritos a seguir.

Figura 3 – Curva de resposta em degrau unitário e parâmetros.



Fonte: adaptado de (OGATA, 2010).

- Sobressinal (*Maximum Overshoot*): Representa o valor máximo que a resposta atinge acima do valor de regime permanente como ilustrado na Figura 3, geralmente expresso em porcentagem (OGATA, 2010). Ele é um indicador direto da estabilidade relativa do sistema; valores muito altos indicam um sistema pouco amortecido e propenso a oscilações excessivas. Este índice é frequentemente calculado em termos percentuais através da seguinte equação:

$$M_p = \frac{y_{\max} - y_f}{y_f} \times 100\% \tag{6}$$

  em que $y_{\max}$ representa o valor máximo da resposta, y_f o valor final em regime permanente.
- Tempo de Subida (*Rise Time*): Define o tempo necessário para que a saída do sistema mude de um valor inicial baixo para a vizinhança do novo valor desejado. Sistemas subamortecidos, é comum medir o tempo para ir de 0% a 100% do valor final.
- Tempo de Acomodação (*Settling Time*): representa o tempo necessário para que a resposta entre e permaneça dentro de uma faixa de tolerância em torno do valor final, normalmente de $\pm 2\%$ ou $\pm 5\%$, conforme ilustrado na Figura 3. Esse parâmetro indica a rapidez com que o sistema elimina os efeitos transitórios e atinge o regime permanente (OGATA, 2010).
- Erro em Regime Permanente (*Steady-State Error*): Refere-se à diferença residual que permanece entre o valor de referência (*setpoint*) e a variável controlada após o desaparecimento de todos os efeitos transitórios, conforme ilustrado na Figura 3. A capacidade de um sistema em zerar este erro depende diretamente do tipo de sistema e da presença de ação integral no controlador (OGATA, 2010). O erro em regime permanente é dado por:

$$e_{ss} = \lim_{t \to \infty} |SP(t) - PV(t)| \tag{7}$$

2.5 Algoritmo Augmented Random Search (ARS)

O algoritmo ARS, proposto por (MANIA; GUY; RECHT, 2018), é um método de aprendizagem por reforço baseado em otimização por busca aleatória, desenvolvido com o objetivo de reduzir a complexidade computacional presente em algoritmos tradicionais de aprendizado por reforço profundo. O método é baseado em políticas lineares em treinamento sem o uso de redes neurais (RAJESWARAN et al., 2017) e técnicas de otimização sem o uso de derivadas (SALIMANS et al., 2017). O algoritmo ARS utiliza uma política linear e atualiza diretamente seus parâmetros a partir da avaliação de perturbações aleatórias, dispensando o cálculo de gradientes e reduzindo significativamente o custo computacional do treinamento.

O desenvolvimento do ARS surgiu como uma resposta à crescente complexidade dos algoritmos de aprendizado por reforço, que frequentemente demandam elevado poder computacional, grande quantidade de dados de treinamento e arquiteturas de implementação sofisticadas. Segundo Mania, Guy e Recht (2018), mesmo utilizando uma estrutura bastante simples, o algoritmo é capaz de alcançar desempenho comparável a métodos mais complexos em diversos problemas clássicos de controle contínuo.

A principal ideia do ARS consiste em explorar diferentes direções aleatórias no espaço dos parâmetros da política. Em cada iteração são geradas perturbações aleatórias positivas e negativas, as quais são avaliadas por meio da função de recompensa do ambiente. Em seguida, apenas as direções que apresentam melhor desempenho são utilizadas para atualizar os parâmetros da política, tornando o processo de aprendizagem mais eficiente e reduzindo a influência de perturbações pouco relevantes.

Outra característica importante do algoritmo é a normalização das observações do ambiente durante o treinamento. A cada interação são atualizadas a média e a covariância dos estados observados, permitindo que as entradas da política permaneçam aproximadamente normalizadas. Essa estratégia melhora a estabilidade do treinamento e reduz problemas associados às diferentes escalas das variáveis de estado.

Neste trabalho, foi utilizada uma política linear composta por uma matriz de pesos $M \in \mathbb{R}^{p \times n}$, em que n representa o número de entradas e p o número de saídas do controlador PI. A ação de controle é calculada por:

$$u = M \, \mathrm{diag}(\Sigma)^{-\frac{1}{2}}(x - \mu),$$

em que x representa o vetor de estados do ambiente. Enquanto μ e Σ correspondem, respectivamente, à média e à covariância das observações coletadas durante o treinamento.

O procedimento completo do algoritmo empregado neste trabalho é apresentado no Algoritmo 1. Inicialmente são definidos os hiperparâmetros do treinamento, tais como taxa de aprendizagem (α), ruído de exploração (ν), número de direções exploradas (N), número de melhores direções utilizadas na atualização (b) e número máximo de iterações.

- **Taxa de aprendizagem (α):** controla a intensidade da atualização dos parâmetros da política a cada iteração do treinamento.
- **Ruído de exploração (ν):** define a magnitude das perturbações aleatórias aplicadas à política durante a exploração, permitindo avaliar diferentes estratégias de controle.
- **Número de direções exploradas (N):** quantidade de perturbações aleatórias geradas e avaliadas em cada iteração do algoritmo.
- **Número de melhores direções (b):** quantidade das direções com melhor desempenho selecionadas para atualizar a política.
- **Número máximo de iterações:** limite de ciclos de treinamento executados pelo algoritmo antes do encerramento do processo de aprendizado.

Em seguida, para cada episódio, são geradas perturbações aleatórias sobre a matriz de pesos da política, sendo avaliadas duas políticas para cada direção de exploração: uma utilizando perturbação positiva e outra negativa. Para cada política perturbada, a ação de controle é calculada por meio da política linear

$$\pi_{j,k,\pm}(x) = (M_j \pm \nu\delta_k)\ \mathrm{diag}(\Sigma_j)^{-\frac{1}{2}}(x - \mu_j), \tag{8}$$

em que x representa o vetor de estados normalizado do processo e $\pi_{j,k,\pm}(x)$ corresponde à ação aplicada ao ambiente. Após a aplicação dessa ação, é obtida a recompensa associada à respectiva política, denotada por $r(\pi_{j,k,\pm})$.

Após a avaliação de todas as direções, estas são ordenadas de acordo com a recompensa máxima obtida. Apenas as b melhores direções participam da atualização dos parâmetros da política, dada por:

$$M_{j+1} = M_j + \frac{\alpha}{b\sigma_R} \sum_{k=1}^{b} \left[r(\pi_{j,(k),+}) - r(\pi_{j,(k),-}) \right] \delta_k,$$

em que σ_R representa o desvio padrão das recompensas utilizadas na atualização. Esse procedimento reduz a influência de recompensas com grande variabilidade e melhora a estabilidade do

processo de otimização.

Algoritmo 1: ARS

Hiperparâmetros:

$n \leftarrow$ número de entradas

$p \leftarrow$ número de saídas

$\alpha \leftarrow$ taxa de aprendizagem

$\nu \leftarrow$ ruído de exploração

$steps \leftarrow$ número de iterações

$N \leftarrow$ número de direções por direção de ajuste

$b \leftarrow$ número de direções com os melhores desempenhos

Inicializa:

$M_0 \leftarrow 0 \in \mathbb{R}^{p \times n}$ matriz de pesos da perceptron

$\mu_0 \leftarrow 0 \in \mathbb{R}^n$ média das entradas

$j \leftarrow 0$

for $j \leftarrow 0$ **to** $steps$ **do**

Crie as matrizes δ_1,$\delta_2, \ldots, \delta_N$ em $\mathbb{R}^{p \times n}$ preenchidas com valores uniformes entre 0 e 1.

Crie dois vetores $r(\pi_{j,k,-})$ e $r(\pi_{j,k,+})$ em $\mathbb{R}^N$ vazios.

for $k \leftarrow 0$ **to** N **do**

$r(\pi_{j,k,-}) \leftarrow$ recompensa de $\pi_{j,k,-}(x) = (M_j - \nu\delta_k)\,\mathrm{diag}(\Sigma_j)^{-\frac{1}{2}}(x - \mu_j)$.

$r(\pi_{j,k,+}) \leftarrow$ recompensa de $\pi_{j,k,+}(x) = (M_j + \nu\delta_k)\,\mathrm{diag}(\Sigma_j)^{-\frac{1}{2}}(x - \mu_j)$.

end

Ordena as direções δ_k pela máxima recompensa em $\{r(\pi_{j,k,+}),r(\pi_{j,k,-})\}$, sendo então $\overline{\pi}_{j,(k),+}$ e $\overline{\pi}_{j,(k),-}$ as políticas respectivas.

Atualiza

$$M_{j+1} = M_j + \frac{\alpha}{b\sigma_R} \sum_{k=1}^{b} \left[r(\overline{\pi}_{j,(k),+}) - r(\overline{\pi}_{j,(k),-}) \right] \delta_k,$$

onde σ_R é o desvio padrão das recompensas usadas na etapa de atualização.

Configura μ_{j+1},Σ_{j+1} para ser a média e covariância dos estados $2NH(j+1)$ encontrados desde o início do treinamento.

$j \leftarrow j + 1$;

end

3 Metodologia

Este capítulo apresenta a estratégia de controle proposta. Inicialmente, é realizada a descrição do processo CSTR, destacando suas características dinâmicas e os desafios inerentes ao seu controle. Em seguida, é apresentada a estrutura de controle desenvolvida para o processo CSTR, fundamentada em uma estratégia baseada em aprendizagem por reforço. Para esta estrutura, é abordado a arquitetura, os componentes envolvidos e a forma como o agente interage com o ambiente para determinar ações de controle.

O desenvolvimento computacional foi realizado utilizando a linguagem de programação Python versão 3.12.13 na plataforma Google Colaboratory. A implementação do modelo do processo foi baseada em um código disponibilizado pelo repositório APMonitor[1], que descreve a simulação dinâmica CSTR. A partir desse código-base foram realizadas adaptações para atender aos objetivos da pesquisa, incluindo modificações na lógica de controle, implementação de diferentes pontos de operação (*setpoints*) e organização da simulação em malha fechada.

As principais bibliotecas utilizadas foram:

- NumPy (2.0.2), para operações matemáticas e manipulação de dados;
- Matplotlib (3.10.0), para geração de gráficos e visualização dos resultados;
- Gymnasium (OpenAI Gym 1.3.0), utilizada para estruturar o ambiente de simulação visando sua futura integração com algoritmos de aprendizagem por reforço.

Os dados utilizados no estudo foram provenientes exclusivamente das simulações computacionais realizadas durante os experimentos, sendo armazenados em arquivos estruturados para posterior análise.

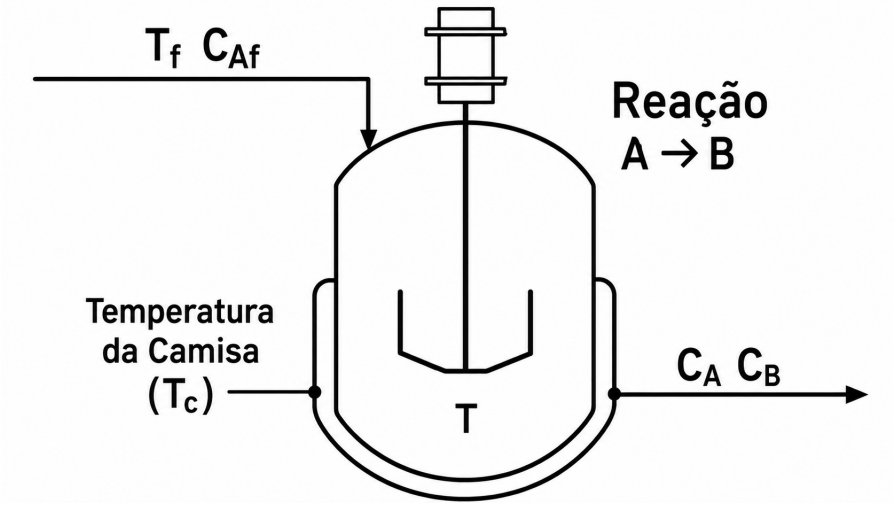
3.1 Descrição do Processo CSTR

O processo estudado consiste em um reator contínuo de mistura perfeita (CSTR), amplamente utilizado na indústria química devido à sua importância em processos de reação exotérmica (ANTONELLI; ASTOLFI, 2003). O modelo representa a conversão irreversível do reagente A em produto B, sendo controlada a temperatura do reator por meio da temperatura da camisa de resfriamento.

A escolha do CSTR deve-se ao seu comportamento dinâmico não linear, tornando-o um ambiente adequado para avaliação de técnicas avançadas de controle. A Figura 4 apresenta uma representação simplificada do processo CSTR.

[1] APMonitor. *Stirred Reactor*. Disponível em: <https://github.com/APMonitor/pdc/blob/master/app_cstr_pid.py>. Acesso em: 28 jun. 2026.

Figura 4 – Representação simplificada do processo CSTR



Fonte: Adaptado de (ANTONELLI; ASTOLFI, 2003).

O modelo matemático utilizado para representar a dinâmica do processo é descrito pelas Equações (9) e (10). Os parâmetros utilizados na simulação são apresentados na Tabela 1.

Tabela 1 – Parâmetros do processo CSTR.

Parâmetro	**Notação**	**Valor**
Vazão do processo	q	$100\,\mathrm{m}^3/\mathrm{s}$
Concentração da alimentação	C_{AF}	$0{,}877\,\mathrm{mol}/\mathrm{m}^3$
Temperatura de alimentação	T_f	$350\,\mathrm{K}$
Temperatura de entrada do refrigerante	T_{cf}	$350\,\mathrm{K}$
Volume do reator	V	$100\,\mathrm{m}^3$
Coeficiente global de transferência de calor	UA	$7 \times 10^5\,\mathrm{cal}/(\mathrm{min}\cdot\mathrm{K})$
Fator pré-exponencial	k_0	$7{,}2 \times 10^{10}\,\mathrm{min}^{-1}$
Energia de ativação	E/R	$8750\,\mathrm{K}$
Calor de reação	ΔH	$-5 \times 10^4\,\mathrm{J}/\mathrm{mol}$
Densidade dos líquidos	ρ,ρ_c	$1 \times 10^3\,\mathrm{kg}/\mathrm{m}^3$
Calor específico	C_p,C_{pc}	$0{,}239\,\mathrm{J}/(\mathrm{kg}\cdot\mathrm{K})$
Limite inferior da temperatura da camisa	$T_{c,\min}$	$250\,\mathrm{K}$
Limite superior da temperatura da camisa	$T_{c,\max}$	$350\,\mathrm{K}$

Fonte: Adaptado de (HEDENGREN, 2021).

$$\dot{C_A} = \frac{q}{V}(C_{AF} - C_A) - k_0 C_A e^{-\frac{E}{RT}} \tag{9}$$

$$\dot{T} = \frac{q}{V}(T_f - T) - \frac{\Delta H k_0}{\rho C_p} C_A e^{-\frac{E}{RT}} + \frac{\rho_c C_{pc}}{\rho C_p V} q_c \left[1 - e^{-\left(\frac{hA}{\rho_c C_{pc} q_c}\right)}\right](T_{cf} - T) \tag{10}$$

As Equações (9) e (10) representam, respectivamente, os balanços de massa do reagente A e de energia do reator contínuo de tanque agitado. Esses balanços descrevem a

evolução temporal da concentração do reagente e da temperatura do reator em função das condições de alimentação, da cinética da reação química e da troca de calor com a camisa de resfriamento (BITARÃES; SILVA; EUZÉBIO, 2022).

Na Equação (9), o primeiro termo corresponde à contribuição do escoamento de entrada e saída do reator, enquanto o segundo termo representa o consumo do reagente A devido à reação química de primeira ordem, cuja constante cinética é descrita pela equação de Arrhenius. Já a Equação (10) descreve a dinâmica da temperatura do reator por meio de três contribuições principais. O primeiro termo representa a troca de energia decorrente da alimentação do fluido no reator. O segundo termo corresponde ao calor liberado pela reação exotérmica, sendo proporcional à velocidade da reação. Por fim, o terceiro termo representa a transferência de calor entre o reator e a camisa de resfriamento, responsável pelo controle da temperatura do processo.

Neste trabalho, a variável de processo (PV) corresponde à temperatura do reator (T), enquanto a variável manipulada (MV) corresponde à temperatura da camisa de resfriamento (T_{cf}), utilizada para regular a temperatura do reator (WAYNE, 2020).

3.2 Estrutura de controle para o processo CSTR baseada em aprendizagem por reforço

Após a implementação do modelo do CSTR, foi aplicado um controlador PID convencional para controlar a temperatura do reator por meio da temperatura da camisa de resfriamento. O objetivo foi avaliar o comportamento dinâmico do sistema em distintas regiões de operação. Para isto, foi usada a equação do controlador PID paralelo, conforme apresentado na Equação (11). Durante as simulações, foram registradas a temperatura do reator, a temperatura da camisa de resfriamento e o sinal de controle, possibilitando a análise do desempenho do controlador.

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau)d\tau + K_d \frac{de(t)}{dt} \tag{11}$$

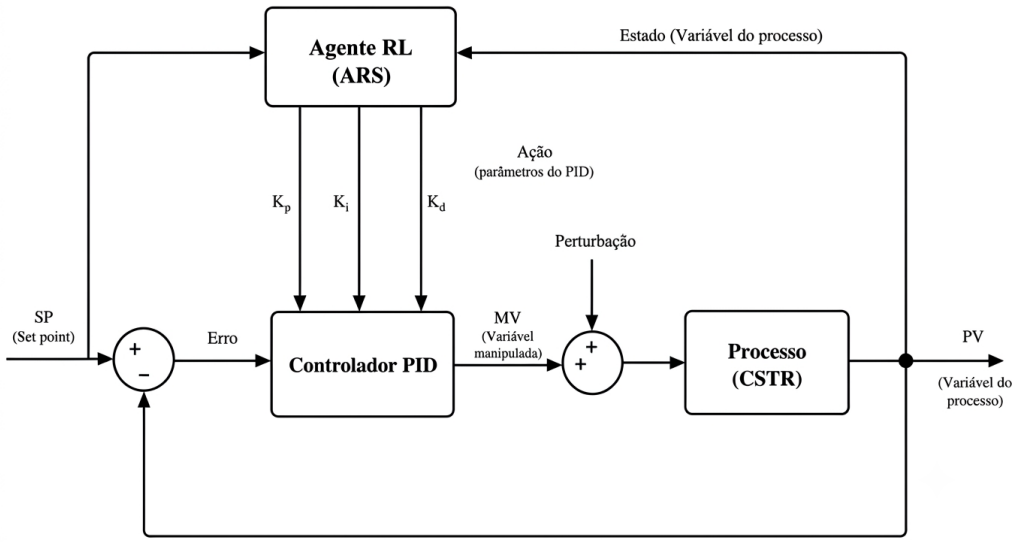
Neste trabalho, o algoritmo ARS é empregado como uma estratégia de aprendizado por reforço para ajustar o controlador PID (ou sua variante PI). Em vez de depender de cálculos matemáticos fixos ou modelos exatos do processo, o ARS define os ganhos do controlador através de tentativas e ajustes sucessivos fundamentados no desempenho real do sistema (BITARãES, 2022). Assim, a estratégia de controle proposta baseia-se na integração do algoritmo ARS ao controlador PI, permitindo a sintonia adaptativa dos parâmetros K_p e K_i durante a operação do processo. Embora a estrutura do algoritmo tenha sido implementada de forma a possibilitar a sintonia de um controlador PID, incluindo o ganho derivativo K_d.

A principal razão para a escolha do controlador PI é porque as respostas obtidas com o controlador PID apresentaram comportamento semelhantes. Dessa forma, optou-se por utilizar apenas o controlador PI, reduzindo o número de parâmetros a serem ajustados pelo algoritmo

de aprendizado e, consequentemente, simplificando o processo de sintonia sem comprometer o desempenho do sistema. Além disso, a ação derivativa é conhecida por apresentar maior sensibilidade a ruídos de medição.

A estrutura básica da aprendizagem por reforço é apresentada na Figura 5. Nessa abordagem, o agente de aprendizagem interage continuamente com o ambiente representado pelo modelo do CSTR. A cada episódio de treinamento, o algoritmo observa o estado atual do sistema, modifica os parâmetros do controlador e recebe uma recompensa calculada a partir do desempenho obtido. O objetivo consiste em maximizar essa recompensa, reduzindo o erro de controle e o esforço da ação de controle.

Figura 5 – Diagrama de controle PID com agente RL.



Fonte: Adaptado de (SHUPRAJHAA; SUJIT; SRINIVASAN, 2022).

Para o treinamento da sintonia do controlador PI utilizando o algoritmo ARS, foi implementada no modelo do CSTR uma função de recompensa baseada no desempenho da resposta do sistema. Essa função considera simultaneamente o erro de controle e o esforço da ação de controle, dado por:

$$R = \frac{1}{\alpha \cdot IAE + \beta \cdot SII}, \tag{12}$$

em que IAE representa a Integral do Erro Absoluto, SII corresponde à Soma dos Incrementos Absolutos da Variável Manipulada e α, β são coeficientes de ponderação responsáveis por ajustar a contribuição de cada índice na função de recompensa.

Dessa forma, o algoritmo ARS busca maximizar a recompensa, o que implica minimizar simultaneamente os valores de IAE e SII. Como consequência, ao final do treinamento são obtidos os parâmetros do controlador PI (K_p e K_i), proporcionando um compromisso entre o desempenho da malha fechada e a suavidade da ação de controle.

Os coeficientes de ponderação α e β permite ajustar a importância relativa de cada métrica durante o processo de treinamento. Nos experimentos realizados, os parâmetros de α e β foram variados no intervalo de 0,1 a 1,0.

No sistema proposto para o controle do reator CSTR, o algoritmo ARS atua em um nível hierárquico superior ao controlador PI, sendo responsável pela determinação dos parâmetros de sintonia utilizados pelo controlador durante a operação do processo (BITARãES, 2022). Conforme ilustrado na Figura 5, o agente de aprendizado interage continuamente com o ambiente, observando o estado do sistema, avaliando o desempenho obtido e atualizando sua política de decisão ao longo das iterações. Neste contexto, observe que:

- **Estados e ações:** O estado fornecido ao agente é composto pelas informações da variável do processo (PV), correspondente à temperatura do reator, e da variável manipulada (MV), correspondente ao sinal de controle aplicado ao processo, obtidas a partir da interação com o ambiente de simulação. A partir dessas informações, a política do algoritmo ARS calcula a ação correspondente, representada pelos parâmetros do controlador. Embora a implementação permita a sintonia de um controlador PID, incluindo os ganhos K_p, K_i e K_d, neste trabalho optou-se pela utilização de um controlador PI, de modo que apenas os ganhos K_p e K_i foram efetivamente ajustados pelo algoritmo.
- **Mecanismo de recompensa:** Após a aplicação dos parâmetros calculados pelo agente, o ambiente executa a simulação do processo e retorna uma recompensa que quantifica a qualidade da sintonia obtida. Diferentemente de abordagens baseadas apenas no erro de rastreamento, a função de recompensa empregada neste trabalho considera simultaneamente a Integral do Erro Absoluto (IAE) e a Soma dos Incrementos Absolutos da Variável Manipulada (SII). Dessa forma, soluções que reduzem o erro de controle e, ao mesmo tempo, evitam variações excessivas na variável manipulada recebem maiores recompensas.
- **Busca aleatória otimizada:** O processo de aprendizado é realizado por meio da exploração de perturbações aleatórias aplicadas aos parâmetros da política. Em cada iteração, o algoritmo avalia direções positivas e negativas dessas perturbações, executando simulações para cada uma delas. Ao final da avaliação, apenas as direções que produziram os maiores valores de recompensa são utilizadas para atualizar os pesos da política, conduzindo progressivamente a sintonia do controlador para regiões de melhor desempenho.
- **Estabilização do aprendizado:** Para aumentar a robustez do processo de treinamento, os estados observados são normalizados continuamente antes de serem utilizados pela política de decisão. Além disso, a atualização dos parâmetros é ponderada pelo desvio padrão das recompensas obtidas em cada iteração, reduzindo oscilações excessivas durante o treinamento e favorecendo uma convergência mais estável.

Como em todos os algoritmos de aprendizagem por reforço, a eficiência da técnica ARS depende diretamente da escolha adequada de seus hiperparâmetros. Esses parâmetros controlam aspectos como a intensidade da exploração, o número de direções avaliadas, a taxa

de aprendizado e o nível de perturbação aplicado à política, influenciando tanto a velocidade de convergência quanto a qualidade da sintonia obtida (SUTTON; BARTO et al., 1998). Portanto, a eficácia da estratégia de controle desenvolvida depende da configuração precisa de seus hiperparâmetros, presentes na Tabela 2 que regulam desde a intensidade da exploração até a velocidade com que o algoritmo converge para a solução ideal (MANIA; GUY; RECHT, 2018).

Tabela 2 – Hiperparâmetros utilizados no treinamento do algoritmo ARS.

Hiperparâmetro	**Valor**	**Descrição**
Número de iterações	3000	Número total de atualizações da política.
Comprimento do episódio	4	Número de episódios executados por iteração.
Taxa de aprendizagem	0,3	Fator de atualização da política.
Número de direções	16	Quantidade de perturbações aleatórias avaliadas por iteração.
Melhores direções	8	Número de direções utilizadas para atualizar a política.
Ruído	0,2	Intensidade das perturbações aplicadas à política.
Semente aleatória	4	Valor utilizado para garantir a reprodutibilidade dos experimentos.

Fonte: Autoria própria, 2026.

Os hiperparâmetros apresentados na Tabela 2 foram definidos empiricamente a partir de diversos testes preliminares, buscando um equilíbrio entre desempenho, estabilidade do treinamento e custo computacional.

- O número de iterações foi fixado em 3000, valor considerado suficiente para que a política convergisse para soluções estáveis. Durante os testes observou-se que um número menor de iterações resultava em controladores com maior variabilidade, enquanto aumentos adicionais produziam ganhos pouco significativos em relação ao tempo computacional empregado.
- O comprimento do episódio foi definido igual a quatro com base em testes preliminares realizados durante a aplicação do algoritmo ARS. Quando apenas um episódio era utilizado, observou-se pouca ou nenhuma evolução da função de recompensa, indicando que a avaliação realizada em um único intervalo não fornecia informações suficientemente discriminantes para orientar a atualização da política. Com quatro episódios consecutivos, a recompensa acumulada passou a representar de maneira mais abrangente o desempenho dos parâmetros propostos pelo ARS, aumentando a diferença entre as avaliações das perturbações positivas e negativas. Essa maior diferenciação favoreceu atualizações mais efetivas dos pesos da política e, consequentemente, uma evolução mais rápida dos ganhos do controlador. Ressalta-se que os diferentes pontos de operação foram treinados separadamente; portanto, o número de episódios não corresponde ao treinamento conjunto desses pontos, mas à quantidade de avaliações realizadas dentro do treinamento individual de cada cenário.

- A taxa de aprendizagem foi estabelecida em 0,3, permitindo que os parâmetros da política fossem atualizados de forma suficientemente rápida, sem provocar oscilações excessivas durante o treinamento. Valores superiores ocasionaram instabilidades na convergência, enquanto valores menores aumentaram significativamente o tempo necessário para o aprendizado.
- Foram adotadas 16 direções aleatórias por iteração, das quais apenas as oito melhores foram utilizadas para atualizar a política. Essa configuração segue a filosofia do algoritmo ARS, na qual diversas perturbações são avaliadas, mas somente aquelas que proporcionam maior recompensa contribuem para a atualização dos parâmetros. A utilização das melhores direções reduz a influência de perturbações pouco informativas, tornando o processo de aprendizado mais eficiente.
- A intensidade do ruído foi fixada em 0,2, produzindo perturbações suficientemente grandes para promover a exploração do espaço de soluções, porém sem comprometer a estabilidade do treinamento. Valores menores dificultaram a exploração de novas regiões da política, enquanto valores maiores produziram atualizações excessivamente agressivas.
- A semente aleatória foi definida como 4 com o objetivo de garantir a reprodutibilidade dos experimentos. Dessa forma, diferentes execuções do algoritmo utilizam a mesma sequência de números pseudoaleatórios, permitindo a reprodução dos resultados apresentados neste trabalho.

4 Resultados e Discussões

Neste capitulo, são apresentados os resultados obtidos, com os controladores PI para o processo CSTR utilizando o algoritmo ARS. O treinamento foi realizado com 3 mil iterações sendo cada iteração composta por 4 episódios. Além disso, em cada iteração foram geradas 16 direções de perturbação, sendo avaliadas nas direções positivas e negativas de perturbação, totalizando 32 explorações. Portanto, 96 mil execuções completas do processo.

Ademais, as 8 melhores direções são selecionadas para a atualização da política. Por fim, foi usado uma taxa de aprendizagem de 0,3, uma intensidade de ruído de 0,2 e uma semente aleatória igual a 4.

4.1 Treinamento para pontos de operações iguais

Com o objetivo de validar os resultados obtidos, foi realizada uma comparação entre diferentes métodos de sintonia do controlador PI. Como referência (*benchmark*), foi utilizada a sintonia proposta por (HEDENGREN, 2021) para o processo CSTR operando no intervalo de temperatura de 300 K a 320 K, com uma perturbação de carga de 10 K aplicada à variável manipulada, correspondente à temperatura da camisa de resfriamento (T_{cf}), no instante de 1,5 minutos.

Também foi considerada a estratégia de controle proposta por (BITARÃES; SILVA; EUZÉBIO, 2022), que emprega o algoritmo ARS utilizando uma função de recompensa baseada apenas no inverso do somatório do erro. Dessa forma, foi possível comparar o desempenho da função de recompensa proposta neste trabalho com abordagens já apresentadas na literatura. A Tabela 3 apresenta os parâmetros de sintonia utilizados pelos diferentes métodos para o mesmo ponto de operação.

Tabela 3 – Parâmetros de sintonia do controlador para o degrau de 300 K para 320 K.

	Referência	K_p	K_i
(HEDENGREN, 2021)	300–320	4,62	40,44
(BITARÃES; SILVA; EUZÉBIO, 2022)	300–320	50,39	52,40
Proposta	300–320	8,29	45,34

Fonte: Autoria Própria, 2026

A função de recompensa utilizada neste trabalho foi definida conforme a Equação (12), sendo composta por uma combinação ponderada entre os índices IAE e SII. Inicialmente, buscou-se avaliar o impacto do parâmetro α, responsável por ponderar a contribuição do índice IAE, diretamente relacionado ao erro de rastreamento da variável controlada. Para isso, o parâmetro β foi mantido fixo em 0,6, enquanto α foi variado nos valores 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8; 0,9 e 1,0. Para cada valor de α, o algoritmo ARS foi executado por 3000

iterações, sendo selecionada a maior recompensa obtida durante o respectivo treinamento. Os ganhos do controlador PI obtidos para cada valor de α são apresentados na Tabela 4.

Tabela 4 – Ganhos do controlador PI obtidos para diferentes valores de α, considerando $\beta = 0{,}6$.

α	β	K_p	K_i
0,1	0,6	1,40	4,59
0,2	0,6	8,31	44,68
0,3	0,6	8,31	44,68
0,4	0,6	2,79	8,47
0,5	0,6	8,31	44,68
0,6	0,6	8,31	44,68
0,7	0,6	8,31	44,68
0,8	0,6	8,31	44,36
0,9	0,6	8,30	44,62
1,0	0,6	8,29	45,34

Fonte: Autoria Própria, 2026.

Com o objetivo de investigar a influência do termo associado à suavidade da ação de controle, foi fixado o parâmetro $\alpha = 1$, mantendo constante a importância atribuída ao IAE durante todo o treinamento. Em seguida, foram realizados treinamentos independentes variando o parâmetro β nos valores 0,1; 0,2; 0,3; 0,4; 0,5; 0,6; 0,7; 0,8 e 0,9.

Para cada valor de β, o algoritmo ARS foi executado por 3000 iterações, sendo selecionada a maior recompensa obtida durante o respectivo treinamento. Posteriormente, para cada melhor solução encontrada, foram analisados os índices de desempenho IAE e SII, permitindo avaliar o compromisso entre a qualidade do rastreamento da referência e a suavidade da ação de controle. Os ganhos do controlador PI obtidos para cada valor de β são apresentados na Tabela 5.

Conforme apresentado na Tabela 4, observou-se que, para o ponto de operação considerado (300 K para 320 K), variações do parâmetro α não produziram melhorias significativas no desempenho do controlador, indicando que a influência desse parâmetro foi pouco expressiva para esse cenário específico. Dessa forma, optou-se por manter $\alpha = 1$, concentrando a análise na influência do parâmetro β.

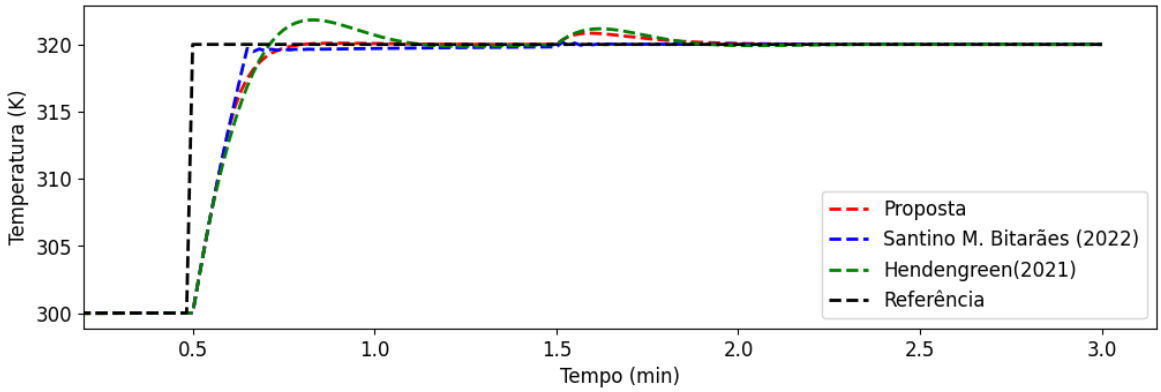
Tabela 5 – Ganhos do controlador PI obtidos para diferentes valores de β, considerando $\alpha = 1$.

α	β	K_p	K_i
1,0	0,1	8,38	46,39
1,0	0,2	8,40	44,75
1,0	0,3	8,38	44,80
1,0	0,4	8,42	44,39
1,0	0,5	8,31	44,80
1,0	0,6	8,29	45,34
1,0	0,7	8,31	44,23
1,0	0,8	8,30	44,62
1,0	0,9	8,30	44,17

Fonte: Autoria Própria, 2026.

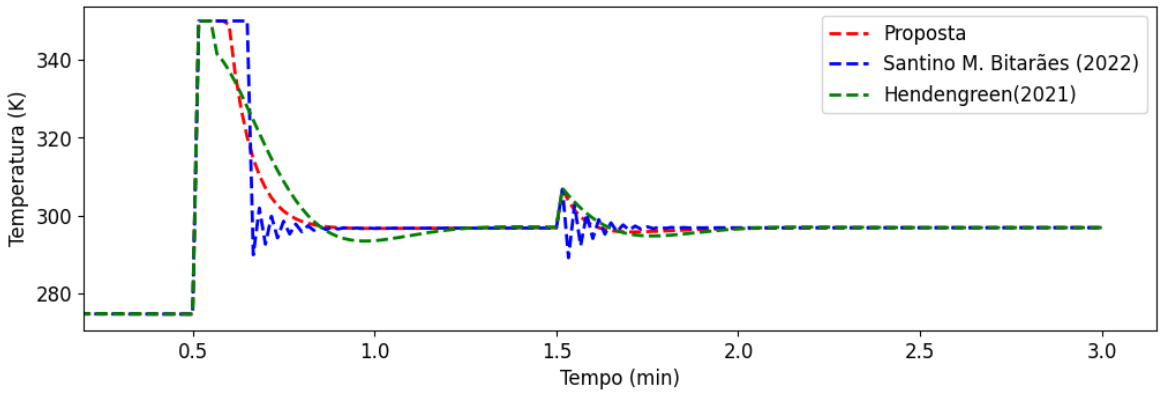
A comparação entre os diferentes treinamentos, apresentada na Tabela 5, mostrou que o caso correspondente a $\beta = 0{,}6$ apresentou o melhor equilíbrio entre os índices IAE e SII. Embora alguns valores de β tenham produzido recompensas semelhantes, a análise conjunta dos indicadores de desempenho demonstrou que $\beta = 0{,}6$ proporcionou uma redução das variações da variável manipulada sem comprometer significativamente o erro de rastreamento da referência. Em razão desse compromisso entre desempenho e esforço de controle, os resultados apresentados nas Figuras 6 e 7 referem-se ao controlador obtido com $\alpha = 1$ e $\beta = 0{,}6$.

Figura 6 – Curvas da saída do processo (T) para degrau de 300 K para 320 K.



Fonte: Autoria Própria, 2026.

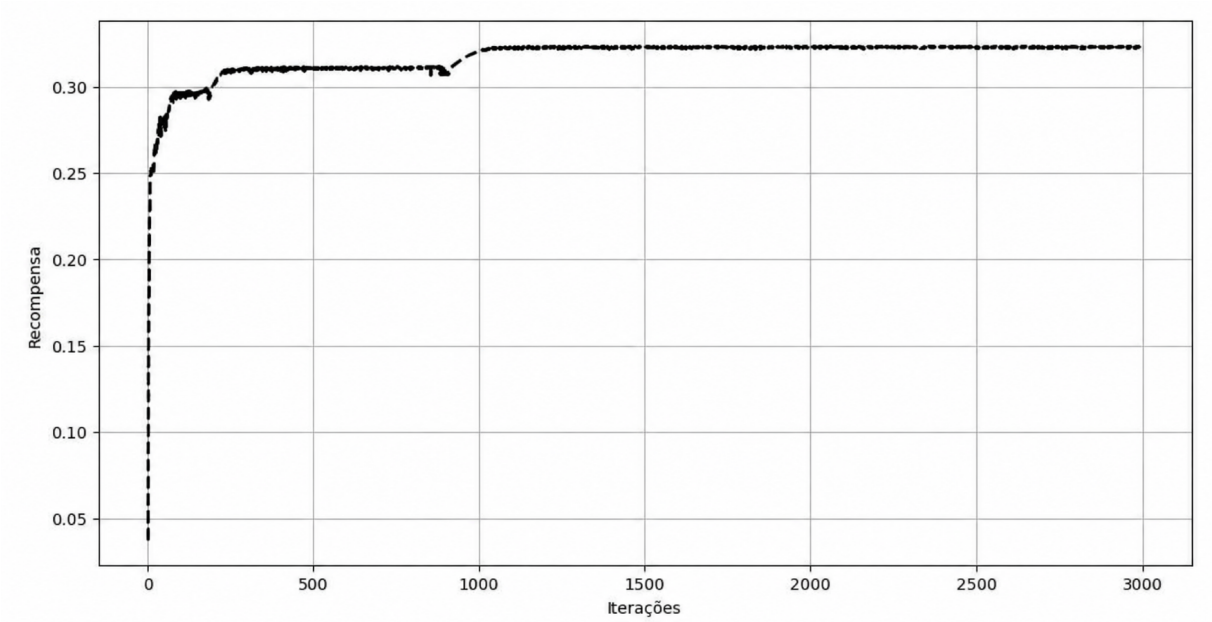
Figura 7 – Curvas da entrada do processo (T_c) para degrau de 300 K para 320 K.



Fonte: Autoria Própria, 2026.

Como apresentado na Figura 8, observa-se que, para o ponto de operação de 300 K para 320 K, a recompensa acumulada tende à convergência a partir de aproximadamente 1000 iterações, indicando a estabilização do processo de aprendizagem do algoritmo ARS. Esse comportamento evidencia que, após esse número de iterações, os ganhos do controlador passam a sofrer pequenas variações, caracterizando a convergência da política de controle. O tempo total de execução do treinamento foi de aproximadamente 22 minutos.

Figura 8 – Curva da recompensa do algoritmo ARS para o degrau de 300 K para 320 K.



Fonte: Autoria própria, 2026.

As Figuras 6 e 7 exibem a saída do processo (T) e a ação de controle T_c alcançadas com os valores de K_p e K_i apresentados na Tabela 3. De acordo com o resultado apresentado na Figura 7, os instantes de saturação da ação de controle ocorreram 0,05 minuto, 0,15 minuto segundos e 0,083 minuto, respectivamente para (HEDENGREN, 2021), (BITARÃES; SILVA; EUZÉBIO, 2022) e ARS com nova função de recompensa proposta neste trabalho.

Observa-se que a sintonia desenvolvida neste trabalho apresentou um tempo de saturação do atuador inferior ao método de (BITARÃES; SILVA; EUZÉBIO, 2022), embora superior ao controlador de (HEDENGREN, 2021). Esse comportamento está relacionado à função de recompensa adotada durante o treinamento do algoritmo ARS, cujo objetivo foi minimizar simultaneamente o erro de rastreamento (IAE) e a variação da ação de controle (SII). Dessa forma, a nova função de recompensa proporcionou um compromisso entre rapidez de resposta e esforço de controle, reduzindo o tempo de saturação em relação ao método anterior baseado exclusivamente na minimização do erro.

Tabela 6 – Índices de desempenho para o degrau 300K $\rightarrow$ 320K.

Método	**IAE**	**SII**	**%OS**	$\mathbf{t_s}$ **(min)**	$\mathbf{t_{ac}}$ **(min)**	$\mathbf{e_\infty}$
(HEDENGREN, 2021)	2,27	2,27	9,12	0,17	0,55	0
(BITARÃES; SILVA; EUZÉBIO, 2022)	1,99	3,19	0	0,13	0,28	0
Proposta	1,80	2,15	0,46	0,15	0,25	0

Fonte: Autoria Própria, 2026.

Os resultados apresentados na Tabela 6 evidenciam que a função de recompensa proposta foi capaz de produzir uma sintonia com desempenho superior aos métodos de referência para o ponto de operação considerado. O controlador obtido por meio da estratégia proposta

apresentou o menor valor de IAE (1,80), indicando melhor desempenho no acompanhamento da referência quando comparado às sintonia propostas por (HEDENGREN, 2021) (2,27) e (BITARÃES; SILVA; EUZÉBIO, 2022) (1,99).

Além da redução do erro de rastreamento, a estratégia proposta também apresentou o menor valor de SII (2,15), representando uma redução de aproximadamente 32,6% em relação ao método de (BITARÃES; SILVA; EUZÉBIO, 2022) e de cerca de 5,3% em comparação à sintonia de (HEDENGREN, 2021). Esse resultado demonstra que estratégia de controle proposta apresentou menores variações, proporcionando uma atuação mais suave do controlador e reduzindo o esforço imposto ao atuador.

Em relação às características transitórias, a sintonia proposta apresentou sobressinal de apenas 0,46%, valor significativamente inferior ao obtido por (HEDENGREN, 2021) (9,12%) e muito próximo daquele obtido por (BITARÃES; SILVA; EUZÉBIO, 2022), que não apresentou sobressinal. O tempo de acomodação também foi o menor entre os métodos avaliados, igual a 0,25 min, evidenciando uma resposta mais rápida ao degrau de referência. Embora o tempo de subida tenha sido ligeiramente superior ao método de (BITARÃES; SILVA; EUZÉBIO, 2022) (0,15 min contra 0,13 min), essa diferença foi pequena e acompanhada por uma resposta mais suave da variável manipulada. Todos os métodos apresentaram erro nulo em regime permanente. Dessa forma, os resultados demonstram que a função de recompensa proposta, baseada simultaneamente nos índices IAE e SII, foi capaz de obter um controlador PI que conciliou elevada precisão no rastreamento da referência, menor esforço de controle e excelente desempenho transitório.

4.2 Treinamento para diferentes pontos de operação

Em seguida, foi avaliado o desempenho do algoritmo ARS utilizando a função de recompensa proposta em uma sequência de mudanças do ponto de operação do processo CSTR, considerando os degraus de referência de 300 K para 320 K, posteriormente para 330 K, 340 K e, por fim, 350 K. Diferentemente do estudo apresentado anteriormente, em que apenas um único degrau foi analisado, este ensaio tem como objetivo verificar a capacidade do controlador em manter um desempenho satisfatório diante de sucessivas alterações na referência, situação frequentemente encontrada em processos industriais.

Durante o treinamento de cada ponto de operação, também foi realizada uma análise dos parâmetros da função de recompensa. Inicialmente, foram avaliadas diferentes combinações dos fatores de ponderação α e β, com o objetivo de investigar a influência de cada termo na qualidade da sintonia obtida. Para cada combinação, o algoritmo foi treinado completamente e, ao final das 3000 iterações, foi selecionada a maior recompensa obtida. Posteriormente, os respectivos índices IAE e SII associados à melhor solução encontrada foram analisados, permitindo selecionar a configuração que apresentasse o melhor compromisso entre erro de rastreamento e suavidade da ação de controle.

4.2.1 Treinamento degrau 320 → 330 K

As Tabelas 7 e 8 apresentam os ganhos do controlador PI obtidos para diferentes combinações dos parâmetros α e β no ponto de operação de 320–330 K. Conforme observado na Tabela 7, a variação de α, mantendo $\beta = 0{,}6$, produziu poucas alterações nos ganhos do controlador para a maior parte das combinações avaliadas. Apenas para $\alpha = 0{,}1$ e $\alpha = 0{,}4$ foram obtidos ganhos significativamente inferiores, indicando soluções de menor desempenho durante o treinamento.

Tabela 7 – Ganhos do controlador PI obtidos para diferentes valores de α, considerando o ponto de operação de 320–330 K e $\beta = 0{,}6$.

α	β	K_p	K_i
0,1	0,6	2,36	5,80
0,2	0,6	27,59	121,59
0,3	0,6	27,59	121,59
0,4	0,6	16,68	80,18
0,5	0,6	27,59	121,59
0,6	0,6	27,59	121,59
0,7	0,6	27,59	121,59
0,8	0,6	27,59	121,59
0,9	0,6	27,86	118,99
1,0	0,6	27,59	121,59

Fonte: Autoria Própria, 2026.

Tabela 8 – Ganhos do controlador PI obtidos para diferentes valores de β, considerando o ponto de operação de 320–330 K e $\alpha = 1$.

α	β	K_p	K_i
1,0	0,1	27,86	227,44
1,0	0,2	27,86	198,44
1,0	0,3	27,91	172,49
1,0	0,4	27,94	154,40
1,0	0,5	27,94	137,78
1,0	0,6	27,59	121,59
1,0	0,7	27,03	108,92
1,0	0,8	22,33	95,87
1,0	0,9	16,59	86,41

Fonte: Autoria Própria, 2026.

Em contrapartida, conforme apresentado na Tabela 8, a variação de β, mantendo $\alpha = 1$, provocou alterações mais expressivas nos ganhos do controlador, principalmente no ganho integral K_i, que apresentou redução gradual à medida que β aumentou. Esse comportamento indica que o termo associado à suavidade da ação de controle exerce maior influência na sintonia obtida pelo algoritmo ARS. Dentre os valores avaliados, a melhor recompensa foi obtida para $\beta = 0{,}8$.

4.2.2 Treinamento degrau 330 → 340 K

As Tabelas 9 e 10 apresentam os ganhos do controlador PI obtidos para diferentes combinações dos parâmetros α e β no ponto de operação de 330–340 K. Conforme mostrado na Tabela 9, observa-se novamente que a variação de α produziu alterações reduzidas nos ganhos do controlador para valores de α superiores a 0,4, indicando baixa sensibilidade da função de recompensa em relação a esse parâmetro. Os ganhos K_p e K_i convergiram para valores muito próximos, sugerindo estabilidade do processo de treinamento e evidenciando que a ponderação do termo associado ao IAE exerceu pouca influência sobre a sintonia obtida nessa faixa de operação.

Tabela 9 – Ganhos do controlador PI obtidos para diferentes valores de α, considerando o ponto de operação de 330–340 K e $\beta = 0{,}6$.

α	β	K_p	K_i
0,1	0,6	3,07	4,50
0,2	0,6	27,58	74,55
0,3	0,6	27,61	83,05
0,4	0,6	28,41	90,83
0,5	0,6	28,47	96,12
0,6	0,6	28,48	98,74
0,7	0,6	28,48	100,64
0,8	0,6	28,50	102,23
0,9	0,6	28,50	103,53
1,0	0,6	28,50	105,64

Fonte: Autoria Própria, 2026.

Tabela 10 – Ganhos do controlador PI obtidos para diferentes valores de β, considerando o ponto de operação de 330–340 K e $\alpha = 1$.

α	β	K_p	K_i
1,0	0,1	28,40	195,39
1,0	0,2	28,48	153,54
1,0	0,3	28,50	132,18
1,0	0,4	28,50	119,36
1,0	0,5	28,50	110,83
1,0	0,6	28,47	105,64
1,0	0,7	28,50	101,52
1,0	0,8	28,49	95,76
1,0	0,9	28,48	92,38

Fonte: Autoria Própria, 2026.

A análise da Tabela 10 mostra que a influência do parâmetro β apresentou comportamento semelhante ao observado no ponto de operação anterior. À medida que β aumentou, verificou-se uma redução gradual do ganho integral K_i, enquanto o ganho proporcional K_p permaneceu praticamente constante. Esse comportamento confirma que o termo associado à suavidade da ação de controle possui maior influência na definição dos ganhos do controlador.

Dentre os valores avaliados, a maior recompensa foi obtida para $\beta = 0{,}6$, sendo essa combinação selecionada para representar o ponto de operação de 330–340 K.

4.2.3 Treinamento degrau 340 → 350 K

As Tabelas 11 e 12 apresentam os ganhos do controlador PI obtidos para diferentes combinações dos parâmetros α e β no ponto de operação de 340–350 K. Conforme mostrado na Tabela 11, diferentemente dos pontos de operação anteriores, a variação de α passou a influenciar significativamente os ganhos do controlador. À medida que α aumentou, observaram-se alterações graduais tanto no ganho proporcional K_p quanto no ganho integral K_i, indicando maior sensibilidade da função de recompensa ao termo associado ao índice IAE. Esse comportamento evidencia que, nessa região de operação, a ponderação do erro de rastreamento passou a exercer maior influência sobre a sintonia obtida pelo algoritmo ARS.

Tabela 11 – Ganhos do controlador PI obtidos para diferentes valores de α, considerando o ponto de operação de 340–350 K e $\beta = 0{,}6$.

α	β	K_p	K_i
0,1	0,6	23,49	35,55
0,2	0,6	27,27	55,66
0,3	0,6	27,11	68,84
0,4	0,6	29,51	77,31
0,5	0,6	29,53	81,12
0,6	0,6	29,54	84,34
0,7	0,6	29,55	87,06
0,8	0,6	29,57	89,49
0,9	0,6	28,48	92,38
1,0	0,6	29,59	93,66

Fonte: Autoria Própria, 2026.

Tabela 12 – Ganhos do controlador PI obtidos para diferentes valores de β, considerando o ponto de operação de 340–350 K e $\alpha = 1$.

α	β	K_p	K_i
1,0	0,1	31,52	216,60
1,0	0,2	29,56	174,20
1,0	0,3	29,56	140,25
1,0	0,4	29,58	118,03
1,0	0,5	29,59	103,66
1,0	0,6	29,59	93,66
1,0	0,7	29,59	86,02
1,0	0,8	29,57	75,42
1,0	0,9	29,57	75,42

Fonte: Autoria Própria, 2026.

A análise da Tabela 12 mostra que o comportamento em relação ao parâmetro β permaneceu semelhante ao observado nos pontos de operação anteriores, com redução progressiva do ganho integral K_i à medida que β aumentou, enquanto o ganho proporcional

K_p apresentou pequenas variações. Entretanto, diferentemente dos demais pontos de operação, a maior recompensa foi obtida para a combinação $\alpha = 0{,}7$ e $\beta = 0{,}6$, evidenciando que, nessa região de operação, uma menor ponderação do termo associado ao erro de rastreamento proporcionou um melhor equilíbrio entre desempenho e suavidade da ação de controle. A análise da Tabela 12 mostra que o comportamento em relação ao parâmetro β permaneceu semelhante ao observado nos pontos de operação anteriores, com redução progressiva do ganho integral K_i à medida que β aumentou, enquanto o ganho proporcional K_p apresentou pequenas variações. Entretanto, diferentemente dos demais pontos de operação, a maior recompensa foi obtida para a combinação $\alpha = 0{,}7$ e $\beta = 0{,}6$, evidenciando que, nessa região de operação, uma menor ponderação do termo associado ao erro de rastreamento proporcionou um melhor equilíbrio entre desempenho e suavidade da ação de controle.

Nos treinamentos correspondentes aos pontos de operação de 300–320 K, 320–330 K e 330–340 K, observou-se que a variação do parâmetro α não proporcionou melhorias significativas na recompensa obtida nem nos índices de desempenho do controlador. Em razão disso, optou-se por manter $\alpha = 1$, concentrando a análise na influência do parâmetro β. As melhores soluções foram obtidas para $\beta = 0{,}6$, $\beta = 0{,}8$ e $\beta = 0{,}6$, respectivamente.

Tabela 13 – Melhores parâmetros obtidos pelo algoritmo ARS para cada ponto de operação.

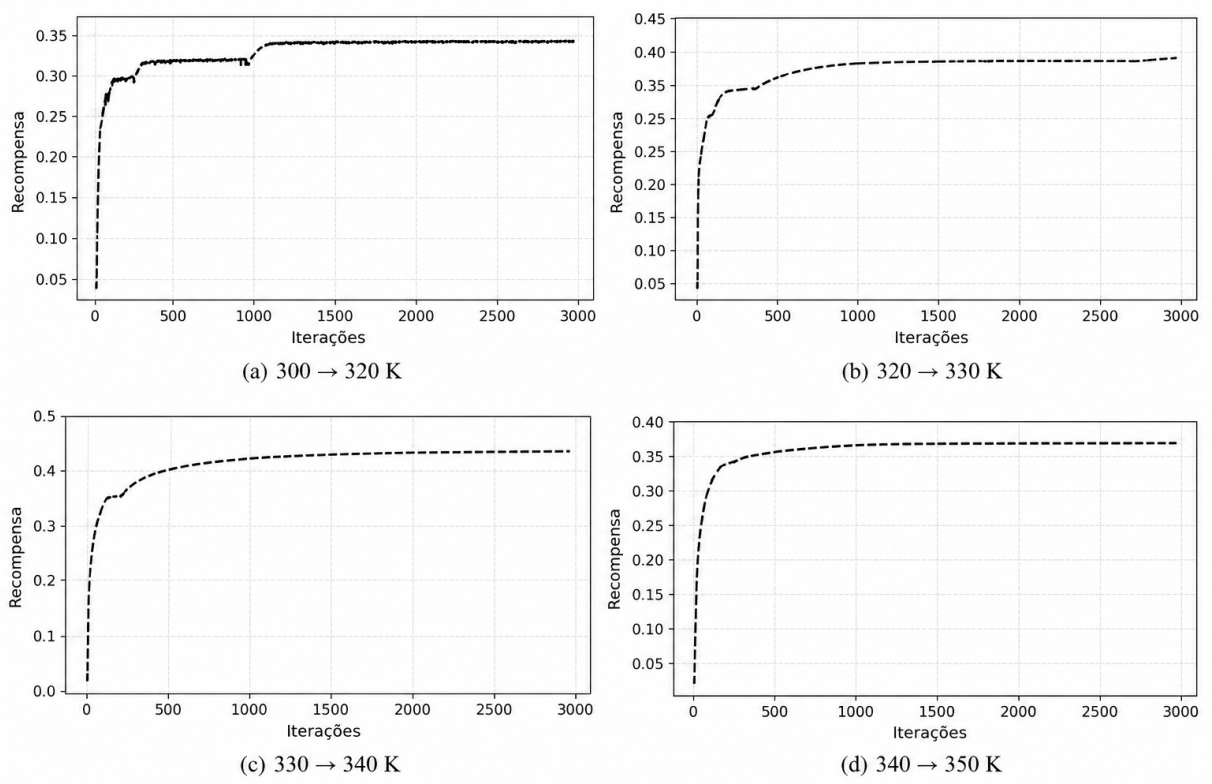
Referência	$\boldsymbol{\alpha}$	$\boldsymbol{\beta}$	**Recompensa**	$\mathbf{K_p}$	$\mathbf{K_i}$
300–320	1,0	0,6	0,3232	8,29	45,34
320–330	1,0	0,8	0,3945	22,33	95,87
330–340	1,0	0,6	0,4386	28.47	105,64
340–350	0,7	0,6	0,3707	29,55	87,06

Fonte: Autoria própria, 2026.

Entretanto, para o ponto de operação de 340–350 K, verificou-se um comportamento distinto. Nesse caso, a redução do parâmetro α resultou em uma melhora da função de recompensa e dos índices de desempenho, sendo a melhor sintonia obtida para $\alpha = 0{,}7$ e $\beta = 0{,}6$. Esse resultado indica que, para regiões de operação mais elevadas, a dinâmica do processo passa a exigir uma ponderação diferente entre os termos da função de recompensa, favorecendo um melhor equilíbrio entre a minimização do erro de rastreamento e a suavização da variável manipulada.

Os ganhos finais obtidos para cada ponto de operação são apresentados na Tabela 13. Observa-se que, à medida que a temperatura de operação aumenta, os ganhos do controlador também sofrem alterações, evidenciando que a dinâmica não linear do reator CSTR modifica-se ao longo da faixa de operação. Esse comportamento reforça a necessidade de realizar uma sintonia específica para cada região de funcionamento, justificando a utilização do algoritmo ARS como ferramenta de ajuste automático do controlador.

Figura 9 – Curvas das recompensas do algoritmo ARS para diferentes pontos de operação.



Fonte: Autoria própria, 2026

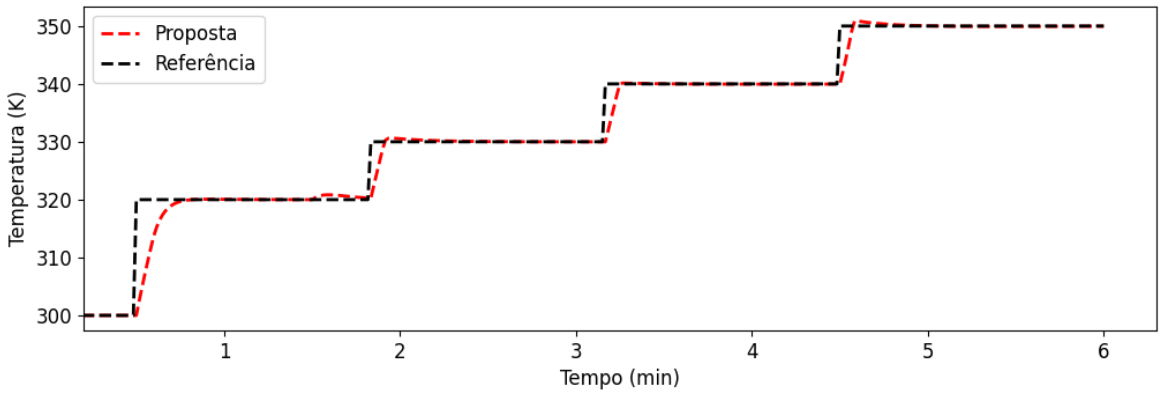
As curvas de aprendizagem apresentadas na Figura 9 ilustram a evolução da função de recompensa durante o treinamento em cada um dos pontos de operação considerados. Em todos os casos observa-se um comportamento semelhante ao obtido no treinamento inicial, caracterizado por um rápido crescimento da recompensa nas primeiras iterações, seguido por uma redução gradual das oscilações até atingir uma região de estabilização. Esse comportamento indica a convergência da política aprendida pelo algoritmo ARS.

Além disso, verifica-se que as maiores recompensas foram obtidas nas etapas finais do treinamento. Para os pontos de operação de 320–330 K, 330–340 K e 340–350 K, a melhor solução foi encontrada na última iteração (3000), enquanto para o ponto de operação de 300–320 K a melhor recompensa ocorreu na iteração 2375. Esses resultados evidenciam que o número de iterações adotado foi suficiente para permitir a convergência do algoritmo e a obtenção de ganhos do controlador adequados para cada região de operação do processo.

As respostas da variável de processo e da variável manipulada são apresentadas nas Figuras 10 e 11, respectivamente. Observa-se que o método ajustou ganhos distintos para cada região de operação, evidenciando a adaptação do controlador às diferentes características dinâmicas do processo. Verifica-se que o controlador foi capaz de acompanhar todas as mudanças de referência sem apresentar instabilidades ou oscilações persistentes, mantendo erro em regime permanente praticamente nulo ao longo de toda a simulação. Além disso, nota-se que a temperatura da camisa apresenta transições suaves entre os diferentes pontos de operação,

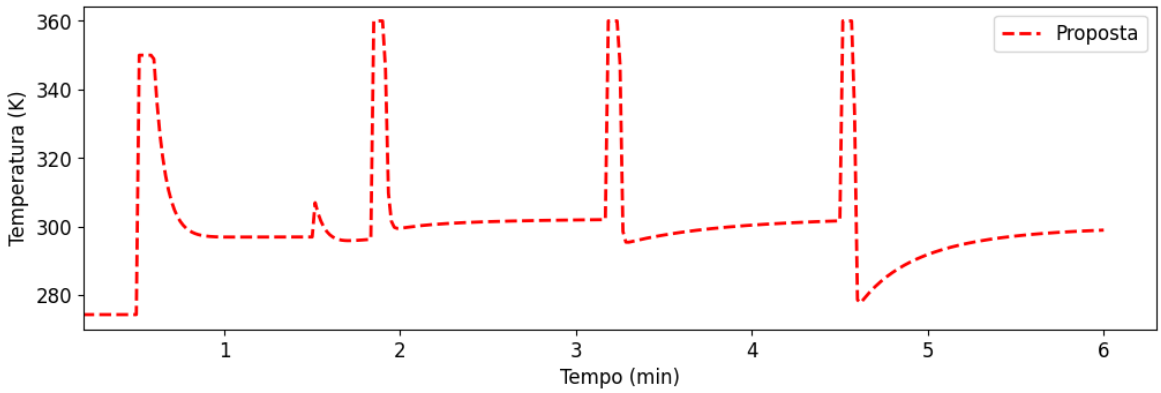
indicando que a função de recompensa proposta conseguiu limitar variações excessivas na ação de controle.

Figura 10 – Curva da saída processo (T) com a sintonia proposta para diferentes pontos de operação.



Fonte: Autoria Própria, 2026.

Figura 11 – Curva da entrada do processo (T_c) com a sintonia proposta para diferentes pontos de operação.



Fonte: Autoria Própria, 2026.

Ao final da simulação, a sintonia proposta apresentou um IAE acumulado de 8,82 e um SII acumulado de 9,45, evidenciando um bom desempenho tanto no rastreamento da referência quanto na redução do esforço de controle durante toda a operação.

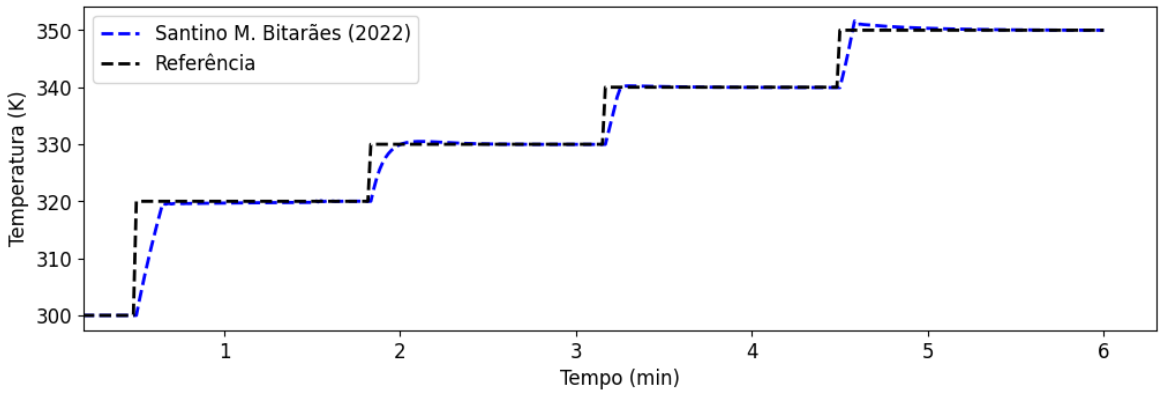
Para fins de comparação, foi realizada a mesma sequência de simulações utilizando os parâmetros de sintonia propostos por (BITARÃES; SILVA; EUZÉBIO, 2022), apresentados na Tabela 14. As respostas obtidas são mostradas nas Figuras 12 e 13. Embora o controlador também acompanhe adequadamente as mudanças de referência, observa-se uma maior intensidade nas variações da variável manipulada ao longo das transições, refletindo diretamente no índice SII.

Tabela 14 – Sintonia proposta por (BITARÃES; SILVA; EUZÉBIO, 2022).

Referência	K_p	K_i
300–320	50,39	52,40
320–330	7,37	30,21
330–340	21,47	27,06
340–350	36,79	56,52

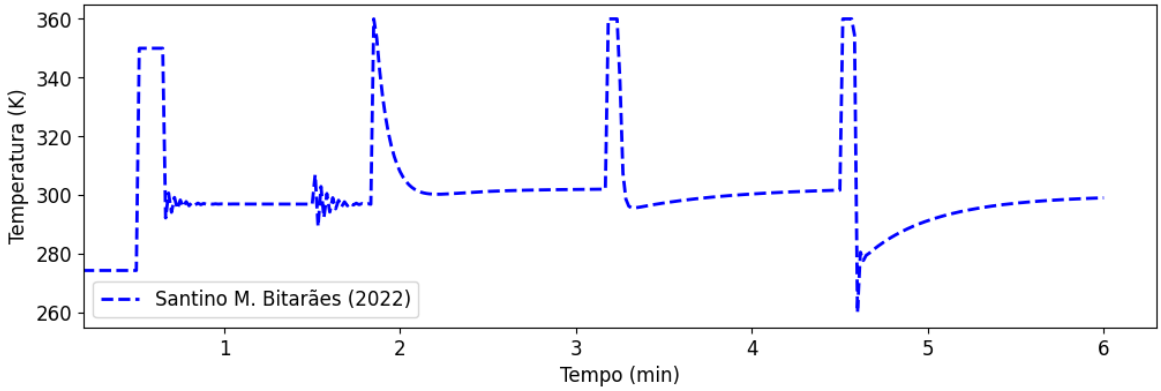
Fonte: Adapatado de (BITARÃES; SILVA; EUZÉBIO, 2022).

Figura 12 – Curva da saída do processo (T) com a sintonia de (BITARÃES; SILVA; EUZÉBIO, 2022) para diferentes pontos de operação.



Fonte: Autoria Própria, 2026.

Figura 13 – Curva da entrada do processo (T_c) com a sintonia de (BITARÃES; SILVA; EUZÉBIO, 2022) para diferentes pontos de operação.



Fonte: Autoria Própria, 2026.

Os resultados acumulados obtidos com o método de (BITARÃES; SILVA; EUZÉBIO, 2022) foram IAE igual a 9,08 e SII igual a 11,90. Comparando-se esses resultados com a proposta deste trabalho, observa-se uma redução aproximada de 2,9% no IAE e de 20,5% no SII. Esses resultados demonstram que a nova função de recompensa permitiu reduzir simultaneamente o

erro acumulado e a variação da ação de controle, produzindo uma sintonia mais equilibrada para diferentes condições de operação.

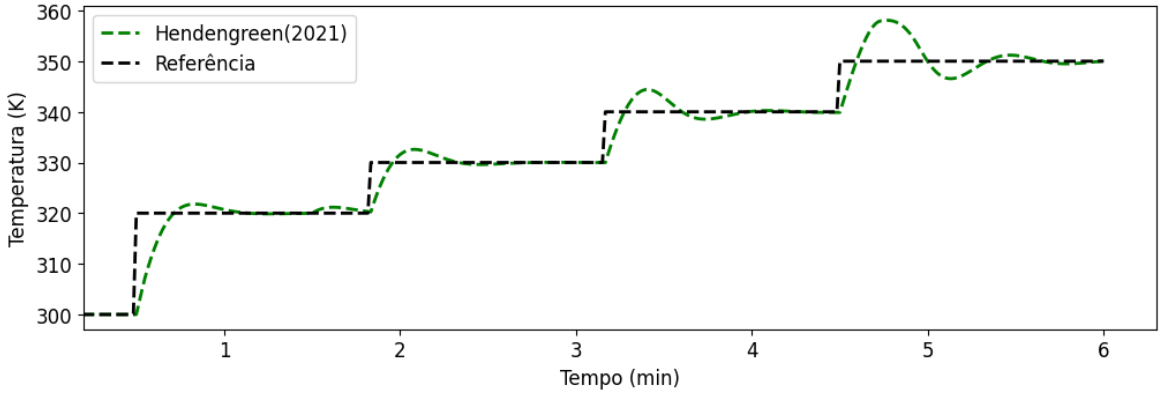
Por fim, realizou-se a comparação com o controlador proposto por (HEDENGREN, 2021), empregando os mesmos ganhos PI para todos os pontos de operação como consta na Tabela 15. As respostas do sistema são apresentadas nas Figuras 14 e 15. Observa-se que, embora o controlador mantenha a estabilidade do processo, sua resposta apresenta maior erro acumulado durante as sucessivas mudanças de referência, consequência da utilização de um único conjunto de parâmetros para todas as regiões de operação.

Tabela 15 – Sintonia proposta por (HEDENGREN, 2021).

Referência	K_p	K_i
300–320	4,62	40,44
320–330	4,62	40,44
330–340	4,62	40,44
340–350	4,62	40,44

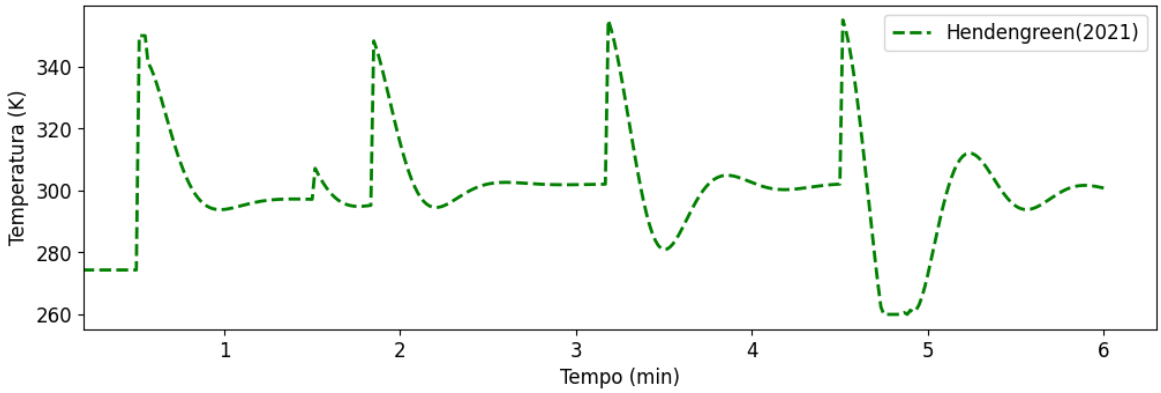
Fonte: Autoria própria, 2026.

Figura 14 – Curva da saída do processo (T) com a sintonia de (HEDENGREN, 2021) para diferentes pontos de operação.



Fonte: Autoria Própria, 2026.

Figura 15 – Curva da entrada do processo (T_c) com a sintonia de (HEDENGREN, 2021) para diferentes pontos de operação.



Fonte: Autoria Própria, 2026.

Para o controlador proposto por (HEDENGREN, 2021), os índices acumulados obtidos foram IAE igual a 14,47 e SII igual a 11,01. Em comparação com esse método, a sintonia proposta reduziu o IAE em aproximadamente 39%, além de apresentar um SII cerca de 14% inferior. Esses resultados evidenciam que o algoritmo ARS, aliado à nova função de recompensa baseada nos índices IAE e SII, foi capaz de obter uma sintonia mais adequada para diferentes pontos de operação do processo CSTR.

Tabela 16 – Índices de desempenho acumulado dos controladores para diferentes pontos de operação.

Método	**IAE**	**SII**
(HEDENGREN, 2021)	14,47	11,01
(BITARÃES; SILVA; EUZÉBIO, 2022)	9,08	11,90
Proposta	8,82	9,46

Fonte: Autoria Própria, 2026.

De forma geral, os resultados obtidos demonstram que a estratégia proposta apresentou o melhor desempenho dentre os métodos avaliados para o cenário de múltiplos degraus, alcançando simultaneamente o menor erro acumulado e o menor esforço de controle. Esse comportamento indica que a função de recompensa desenvolvida neste trabalho favorece um melhor compromisso entre precisão de rastreamento e suavidade da ação de controle, características desejáveis em aplicações industriais sujeitas a frequentes mudanças nas condições de operação.

5 Conclusões e Trabalhos Futuros

Este trabalho teve como objetivo desenvolver uma estratégia de sintonia de controladores PI para o processo simulado CSTR utilizando o algoritmo de Aprendizado por Reforço ARS, empregando uma nova função de recompensa baseada simultaneamente na minimização do erro de rastreamento IAE e da variação da ação de controle SII. A proposta buscou obter controladores capazes de proporcionar um compromisso entre precisão, suavidade da ação de controle e menor esforço imposto ao atuador.

A avaliação da estratégia foi realizada por meio de comparações com duas abordagens de referência encontradas na literatura. No primeiro cenário, considerando um único ponto de operação, verificou-se que a função de recompensa proposta foi capaz de reduzir significativamente a variação da ação de controle e o tempo de saturação do atuador em relação ao método baseado exclusivamente na minimização do erro, mantendo desempenho de rastreamento semelhante. Esses resultados evidenciaram que a consideração simultânea dos índices IAE e SII durante o treinamento favorece a obtenção de controladores mais equilibrados para aplicações industriais.

Em seguida, foram realizados ensaios envolvendo sucessivas mudanças de referência em diferentes regiões de operação. Nesse cenário, a estratégia proposta apresentou o menor erro acumulado e o menor esforço de controle entre os métodos avaliados, demonstrando que o algoritmo ARS foi capaz de adaptar os parâmetros do controlador às diferentes características dinâmicas do processo. Os resultados obtidos evidenciaram que a utilização de ganhos específicos para cada faixa de operação proporciona desempenho superior à utilização de um único conjunto de parâmetros para toda a região de funcionamento do reator.

Assim, conclui-se que a utilização do algoritmo ARS associada à função de recompensa proposta constitui uma alternativa eficiente para a sintonia automática de controladores PI aplicados ao processo CSTR. Os resultados obtidos confirmam que a abordagem foi capaz de produzir controladores com desempenho equilibrado entre qualidade de rastreamento, suavidade da ação de controle e redução do esforço imposto ao atuador, atendendo aos objetivos estabelecidos para este trabalho.

Como perspectivas para trabalhos futuros, sugere-se a investigação da metodologia proposta em outros processos não lineares e de maior complexidade, bem como sua aplicação em controladores PID e em estratégias de controle mais avançadas, como controle preditivo baseado em modelo (MPC). Além disso, recomenda-se a avaliação de outros algoritmos de aprendizado por reforço, a proposição de novas funções de recompensa que incorporem restrições operacionais e critérios econômicos e, por fim, a validação experimental da metodologia em uma planta piloto ou em um sistema físico real, permitindo verificar sua aplicabilidade em condições práticas de operação.

REFERÊNCIAS

ANTONELLI, R.; ASTOLFI, A. Continuous stirred tank reactors: easy to stabilise? **Automatica**, Elsevier, v. 39, n. 10, p. 1817–1827, 2003. Citado 4 vezes nas páginas 10, 11, 25 e 26.

ÅSTRÖM, K. J.; HÄGGLUND, T. Pid control. **IEEE Control Systems Magazine**, v. 1066, p. 30–31, 2006. Citado 2 vezes nas páginas 9 e 17.

BITARÃES, S. M.; SILVA, M. T.; EUZÉBIO, T. A. Sintonia de controladores pi baseada em augmented random search: Estudo de caso do processo cstr. In: **Congresso Brasileiro de Automática-CBA**. [S.l.: s.n.], 2022. v. 3, n. 1. Citado 11 vezes nas páginas 5, 6, 9, 11, 27, 32, 35, 36, 42, 43 e 45.

BITARãES, S. M. **Controle por aprendizagem por reforço aplicado aos processos: CSTR e espessador**. Dissertação (Dissertação (Mestrado em Instrumentação, Controle e Automação de Processos de Mineração)) — Universidade Federal de Ouro Preto, Ouro Preto, MG, Brasil, 2022. Citado 3 vezes nas páginas 12, 27 e 29.

BORASE, R. P.; MAGHADE, D.; SONDKAR, S.; PAWAR, S. A review of pid control, tuning methods and applications: Rp borase et al. **International Journal of Dynamics and Control**, Springer, v. 9, n. 2, p. 818–827, 2021. Citado 2 vezes nas páginas 9 e 12.

CAMPOS, M. C. M. M. de; TEIXEIRA, H. C. G. **Controles Típicos de Equipamentos e Processos Industriais**. 1. ed. São Paulo: Editora Blucher, 2006. ISBN 9788521203988. Citado 5 vezes nas páginas 9, 17, 18, 19 e 20.

CASSOL, G.; CAMPOS, G.; THOMAZ, D.; CAPRON, B.; SECCHI, A. Reinforcement learning applied to process control: a van der vusse reactor case study. In: **Computer Aided Chemical Engineering**. [S.l.]: Elsevier, 2018. v. 44, p. 553–558. Citado na página 12.

CHEN, D.; PENG, Y.; ZHENG, T.; WANG, H.; QU, C.; CHENG, L. Adviser-actor-critic: Eliminating steady-state error in reinforcement learning control. **arXiv preprint arXiv:2502.02265**, 2025. Citado na página 12.

DING, Y.; REN, X.; ZHANG, X.; LIU, X.; WANG, X. Multi-phase focused pid adaptive tuning with reinforcement learning. **Electronics**, MDPI, v. 12, n. 18, p. 3925, 2023. Citado na página 12.

ESMAEILI, M.; SHAYEGHI, H.; NEJAD, H. M.; YOUNESI, A. Reinforcement learning based pid controller design for lfc in a microgrid. **COMPEL-The international journal for computation and mathematics in electrical and electronic engineering**, Emerald Publishing Limited, v. 36, n. 4, p. 1287–1297, 2017. Citado na página 12.

HEDENGREN, J. **Temperature control of a stirred reactor**. 2021. Citado 9 vezes nas páginas 5, 6, 11, 26, 32, 35, 36, 44 e 45.

JESAWADA, H.; YERUDKAR, A.; VECCHIO, C. D.; SINGH, N. A model-based reinforcement learning approach for pid design. **arXiv preprint arXiv:2206.03567**, 2022. Citado na página 12.

JÚNIOR, L.; EULER, G. et al. Sistema de controle data driven para colunas de destilação utilizando deep reinforcement learning. Universidade Federal de Campina Grande, 2023. Citado na página 12.

LAKHANI, A. I.; CHOWDHURY, M. A.; LU, Q. Stability-preserving automatic tuning of pid control with reinforcement learning. **arXiv preprint arXiv:2112.15187**, 2021. Citado na página 12.

LIMA, M. d. O.; BRAGA, A. P. d. S.; PRAÇA, P. P.; LIMA, W. d. S.; SOUSA, R. M.; MOTA, L. F. E. P. Controle com estrutura pid fuzzy aplicado a plantas industriais. In: SIMPÓSIO BRASILEIRO DE AUTOMAÇÃO INTELIGENTE. [S.l.], 2013. Citado na página 12.

LIMA, R. M. Aplicação de algoritmos de reinforcement learning para controle de nível de um tanque. Universidade Federal do Rio de Janeiro, 2018. Citado na página 12.

LUZ, E. M. L.; CAARLS, W. Comparison of reinforcement learning techniques for controlling a cstr process. **Brazilian Journal of Chemical Engineering**, Springer, v. 42, n. 1, p. 345–356, 2025. Citado na página 12.

MANIA, H.; GUY, A.; RECHT, B. Simple random search provides a competitive approach to reinforcement learning. **arXiv preprint arXiv:1803.07055**, 2018. Citado 4 vezes nas páginas 10, 12, 22 e 30.

MARIZ, R. L. Controle preditivo e aprendizado por reforço aplicados a tanques conectados em ambiente de nuvem sob demanda. Edifes, 2024. Citado na página 12.

MATE, S.; PAL, P.; JAISWAL, A.; BHARTIYA, S. Simultaneous tuning of multiple pid controllers for multivariable systems using deep reinforcement learning. **Digital Chemical Engineering**, Elsevier, v. 9, p. 100131, 2023. Citado na página 12.

OGATA, K. **Engenharia de Controle Moderno**. 5. ed. São Paulo: Pearson Prentice Hall, 2010. ISBN 978-85-4301-375-6. Citado 2 vezes nas páginas 20 e 21.

OLIVEIRA, Y. M. S. de; SILVA, I.; SANTOS, D. A. S. dos; CRUZ, A. F. dos S.; SILVA, T. de O.; CARVALHO, G. A. Uso da técnica de reinforcement learning para controle de nível de tanque. **Apoena**, v. 6, p. 153–164, 2023. Citado na página 12.

RAJESWARAN, A.; LOWREY, K.; TODOROV, E. V.; KAKADE, S. M. Towards generalization and simplicity in continuous control. **Advances in neural information processing systems**, v. 30, 2017. Citado na página 22.

REGO, R. C. B.; ARAÚJO, F. M. U. de. Nonlinear control system with reinforcement learning and neural networks based lyapunov functions. **IEEE Latin America Transactions**, IEEE, v. 19, n. 8, p. 1253–1260, 2021. Citado na página 12.

RUSSELL, S.; NORVIG, P. Inteligência artificial: uma abordagem moderna. **GEN LTC**, 2022. Citado 4 vezes nas páginas 9, 13, 14 e 17.

SALIMANS, T.; HO, J.; CHEN, X.; SIDOR, S.; SUTSKEVER, I. Evolution strategies as a scalable alternative to reinforcement learning. **arXiv preprint arXiv:1703.03864**, 2017. Citado na página 22.

SHUPRAJHAA, T.; SUJIT, S. K.; SRINIVASAN, K. Reinforcement learning based adaptive pid controller design for control of linear/nonlinear unstable processes. **Applied Soft Computing**, Elsevier, v. 128, p. 109450, 2022. Citado na página 28.

SOMEFUN, O. A.; AKINGBADE, K.; DAHUNSI, F. The dilemma of pid tuning. **Annual Reviews in Control**, Elsevier, v. 52, p. 65–74, 2021. Citado na página 9.

SPIELBERG, S.; TULSYAN, A.; LAWRENCE, N. P.; LOEWEN, P. D.; GOPALUNI, R. B. Toward self-driving processes: A deep reinforcement learning approach to control. **AIChE journal**, Wiley Online Library, v. 65, n. 10, p. e16689, 2019. Citado 2 vezes nas páginas 9 e 12.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. 2. ed. Cambridge, MA: MIT Press, 2018. Citado na página 9.

SUTTON, R. S.; BARTO, A. G. et al. **Reinforcement learning: An introduction**. [S.l.]: MIT press Cambridge, 1998. v. 1. Citado 3 vezes nas páginas 15, 17 e 30.

WAYNE, B. **Process control: modeling, design, and simulation**. [S.l.]: Prentice Hall, 2020. Citado na página 27.

YU, X.; XU, S.; FAN, Y.; OU, L. Self-adaptive lsac-pid approach based on lyapunov reward shaping for mobile robots. **Journal of Shanghai Jiaotong University (Science)**, Springer, v. 30, n. 6, p. 1085–1102, 2025. Citado na página 12.