# PLANEJAMENTO DE ROTA DE VEÍCULO ARTICULADO EM AMBIENTE DE ESTACIONAMENTO *

**Dayana Cardoso**
Affiliation
UFRGS
Porto Alegre
dscardoso@inf.ufrgs.br

**Walter Frank**
Affiliation
UFRGS
Porto Alegre
walter.frank@inf.ufrgs.br

**ABSTRACT**

O planejamento de trajetórias para veículos não holonômicos articulados enfrenta a explosão combinatória do espaço de estados ($10^{10}$ configurações) e a inadequação de métricas Euclidianas para capturar restrições cinemáticas. Este trabalho investiga a eficácia do Heuristic Guided Reinforcement Learning (HuRL) na aceleração do treinamento de agentes de Aprendizado por Reforço para tarefas de estacionamento. Utilizando o algoritmo Soft Actor-Critic (SAC), avaliamos comparativamente três estratégias de reward shaping: ausência de heurística (baseline), distância Euclidiana e distância topológica via Breadth-First Search (BFS). Os resultados demonstram que a heurística baseada em BFS, ao incorporar a geometria dos obstáculos, supera significativamente as demais abordagens, permitindo a convergência robusta e altas taxas de sucesso, enquanto a heurística Euclidiana estagna em mínimos locais e o baseline falha em aprender a tarefa. A utilização de heurísticas permite que o agente de RL concentre seu aprendizado no desvio de obstáculos e otimização temporal. Resultados experimentais demonstram convergência significativamente mais rápida que métodos puramente baseados em RL, mantendo robustez superior a métodos puramente geométricos.

***K*eywords** Heurística · Politica Ótima · Aprendizado por Reforço

## 1 Introdução

O planejamento de trajetórias para veículos não holonômicos articulados,como trator com reboques, apresenta desafios tanto computacionais quanto teóricos. Do ponto de vista computacional, o espaço de estados cresce exponencialmente: considerando um mapa de 100 m $\times$ 100 m com discretização das coordenadas cartesianas em passos de 10 cm ($10^6$ posições), do ângulo de orientação $\theta$ em passos de 1 (360 valores) e do ângulo de articulação $\beta$ em passos de 1 limitado a aproximadamente 180 valores devido às restrições físicas de canivete, obtém-se uma ordem de grandeza de $10^{10}$ estados possíveis. Essa explosão combinatória torna abordagens de busca exaustiva impraticáveis, mesmo com técnicas modernas de otimização.

Paralelamente, técnicas de aprendizado por reforço (RL) têm demonstrado capacidade de aprender políticas complexas através da interação com o ambiente. No entanto, algoritmos de RL puros frequentemente requerem milhões de amostras para convergir, especialmente em espaços de estados de alta dimensionalidade como o presente caso. Uma abordagem promissora para acelerar o aprendizado é a incorporação de conhecimento prévio através de heurísticas que guiem a exploração do agente.

Neste contexto, Cheng *et al.* [1] propuseram o *Heuristic-Guided Reinforcement Learning* (HuRL), um *framework* que integra informações heurísticas ao processo de aprendizado através da técnica de *reward shaping* baseada em potencial. Esta abordagem visa "encurtar" o horizonte efetivo do problema, fornecendo sinais de recompensa densos que aceleram a convergência do agente. Entretanto, a eficácia do HuRL é criticamente dependente da qualidade e adequação da função heurística empregada.

* *Citation*: **Authors. Title. Pages.... DOI:000000/11111.**

O presente trabalho investiga o impacto de diferentes funções heurísticas na eficiência do treinamento de agentes de RL para a tarefa de estacionamento autônomo de veículos articulados. Especificamente, avalia-se comparativamente três estratégias de *reward shaping*: (i) um *baseline* sem informação heurística auxiliar, submetido a recompensas esparsas; (ii) uma heurística baseada na distância Euclidiana entre o veículo e o objetivo; e (iii) uma heurística topológica fundamentada no algoritmo de busca em largura (*Breadth-First Search* - BFS), que computa a distância geodésica considerando a geometria dos obstáculos.

Utilizando o algoritmo *Soft Actor-Critic* (SAC) como base, os experimentos foram conduzidos em um ambiente de simulação customizado, modelado segundo a API Gymnasium, que reproduz um espaço de estacionamento com $150 \times 150$ metros contendo vagas dispostas em fileiras opostas e obstáculos gerados estocasticamente. O modelo cinemático do veículo articulado segue a formulação de bicicleta, incorporando as dinâmicas de acoplamento entre trator e reboque.

## 2 Metodologia

A metodologia consiste em três pilares: (1) a modelagem de um ambiente de simulação compatível com a API Gymnasium [2], que implementa a dinâmica não-holonômica do sistema trator-reboque baseada no modelo cinemático de bicicleta; (2) a definição de funções de reward shaping baseadas em potencial para mitigar a esparsidade de recompensas; e (3) o treinamento e avaliação comparativa de agentes utilizando o algoritmo Soft Actor-Critic (SAC) [3]

### 2.1 Modelo Cinemático do Veículo Não Holonômico Articulado

O ambiente de treinamento do agente de RL é modelado a partir da cinemática de um veículo articulado descrita por Guan e Jiang [4]. O modelo considera o trator e o reboque como sistemas rigidamente acoplados, onde as ações do agente correspondem aos controles de velocidade longitudinal $v$ e ângulo de esterçamento $\alpha$. O espaço de estados $\mathcal{S}$ é definido por $\mathbf{s} = [x, y, \theta_1, \beta]^\top$, onde $(x, y)$ representa a posição cartesiana do trator, $\theta_1$ sua orientação, e $\beta$ o ângulo de articulação entre trator e reboque.

A dinâmica do sistema é governada pelas seguintes equações diferenciais:

$$\begin{aligned} \dot{x} &= v\cos(\theta_1), \\ \dot{y} &= v\sin(\theta_1), \\ \dot{\theta}_1 &= \frac{v}{D}\tan(\alpha), \\ \dot{\beta} &= -\frac{v}{L}\sin(\beta) - \frac{v}{D}\tan(\alpha). \end{aligned} \tag{1}$$

onde $D$ representa a distância entre o eixo traseiro e dianteiro do trator, $L$ o comprimento do reboque, e $a$ a distância do ponto de articulação ao eixo traseiro do reboque. Esses parâmetros estruturais foram mantidos fixos durante todo o processo de treinamento.

A cada passo de interação, o agente observa o estado atual $\mathbf{s}_t$ e seleciona uma ação $\mathbf{a}_t = [v_t, \alpha_t]^\top$ do espaço de ações $\mathcal{A}$. A transição para o próximo estado $\mathbf{s}_{t+1}$ é obtida através da integração numérica das equações (1) ao longo de um intervalo de tempo $\Delta t$. Esta formulação permite que o agente aprenda políticas $\pi : \mathcal{S} \to \mathcal{A}$ que respeitam as restrições não holonômicas impostas pela cinemática do veículo articulado.

### 2.2 Espaço de Observação

O espaço de observação é constituído por um vetor de estados contínuo $s_t \in \mathbb{R}^{22}$ que integra a cinemática do veículo, percepção exteroceptiva e erros de rastreamento relativos ao objetivo. O vetor de estado é definido por:

$$\mathbf{s}_t = [v, \theta, \beta, \alpha, \mathbf{r}, \rho_g, \psi_{rel}, \Delta\theta_{tr}, \Delta\theta_{tl}]^T \tag{2}$$

onde:

- $v, \alpha$: Velocidade longitudinal e ângulo de esterçamento atuais do trator;

- $\theta, \beta$: Orientação global do trator e ângulo de articulação do reboque;
- $\mathbf{r} \in [0, 1]^{14}$: Vetor de leituras de sensores LIDAR (*raycasts*) para detecção de obstáculos;
- $\rho_g$: Índice de proximidade ao alvo, dado por $(1 + ||\mathbf{p} - \mathbf{p}_{goal}||_2)^{-1}$;
- $\psi_{rel}$: Ângulo de azimute relativo (egocêntrico) em direção ao alvo;
- $\Delta\theta_{tr}, \Delta\theta_{tl}$: Erros de orientação do trator e do reboque, respectivamente, em relação à pose final desejada.

### 2.3 Espaço de Ação

O agente opera em um espaço de ação contínuo bidimensional $\mathbf{a}_t \in \mathbb{R}^2$, controlando diretamente as variáveis de entrada do modelo cinemático. O vetor de controle é definido por:

$$\mathbf{a}_t = [v, \alpha]^T \tag{3}$$

onde $v$ representa o comando de velocidade longitudinal e $\alpha$ o ângulo de esterçamento do eixo dianteiro do trator.

### 2.4 Ambiente físico

O ambiente de simulação consiste em um espaço de estacionamento bidimensional com dimensões de $150 \times 150$ metros, totalizando uma área de $22.500$ m$^2$. A configuração espacial é composta por pares de fileiras de vagas dispostas em orientações opostas, intercaladas por barreiras físicas que delimitam os corredores de circulação. Implementou-se um protocolo de randomização de domínio; a cada episódio de treinamento, uma vaga é selecionada aleatoriamente como posição inicial do veículo, enquanto outra vaga distinta é designada como objetivo de estacionamento. Adicionalmente, as demais vagas recebem uma ocupação aleatória, no qual cada vaga possui probabilidade $p = 0{,}25$ de conter um veículo estacionado, sendo considerado um obstáculo estático.

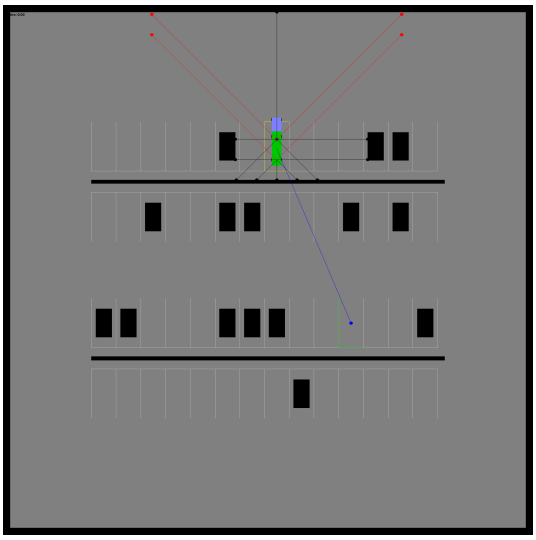

Figure 1: Representação gráfica do ambiente físico de estacionamento.

### 2.5 Heurísticas

A eficácia do reward shaping foi investigada através da comparação de três abordagens distintas para a definição da função de potencial.

- Baseline (Sem heurística): Nesta configuração de controle, o agente é submetido a um regime de recompensa esparsa, recebendo sinais de reforço exclusivamente nos eventos terminais (sucesso ou falha) e penalidades temporais. Esta abordagem serve como linha de base para avaliar o impacto da introdução de sinais de recompensa densos providos pelas heurísticas.

- Heurística Euclidiana: Esta abordagem emprega a distância Euclidiana ($L^2$ norm) entre o centro de massa do veículo e o centróide da vaga de destino como função de potencial. O `reward shaping` é formulado para recompensar o gradiente negativo da distância, incentivando a redução da distância linear a cada passo de tempo. Embora computacionalmente eficiente, esta heurística ignora a topologia do ambiente e restrições cinemáticas, tornando-a suscetível a mínimos locais em ambientes com obstáculos não convexos.
- Heurística Topológica (BFS): Para incorporar a geometria dos obstáculos na função de recompensa, utiliza-se o algoritmo de busca em largura (Breadth-First Search - BFS). O ambiente é discretizado em uma grade de ocupação com resolução de $1m \times 1m$. O mapa de distâncias (considerando obstáculos estáticos) é computado a partir do ponto alvo. A recompensa densa é derivada da redução da distância do menor caminho válido pelo BFS, guiando o agente através de trajetórias livres de colisão e mitigando o problema de mínimos locais.

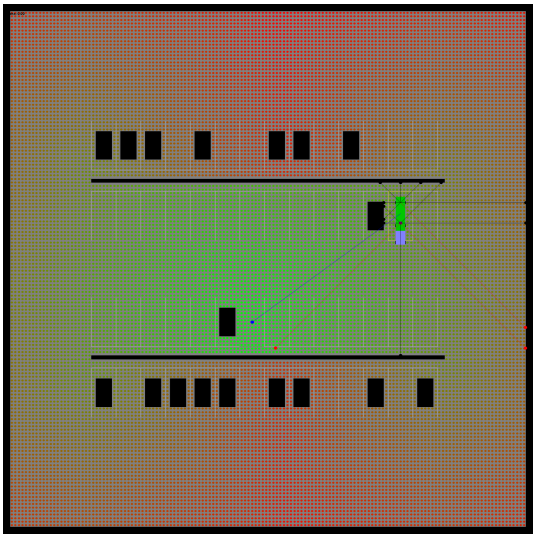

Figure 2: Ambiente de estacionamento com o mapa de distâncias computado pelo algoritmo BFS. A área pontilhada em vermelho representa os pontos mais distantes do objetivo, enquanto a área verde representa os pontos próximos ao objetivo.

### 2.6 Aceleração via HuRL

Seguindo o framework de Cheng et al. [1], formulamos o problema como um Processo de Decisão de Markov (MDP) $\mathcal{M}$. Para acelerar o aprendizado, construímos um MDP remodelado $\tilde{\mathcal{M}}$ utilizando *Potential-Based Reward Shaping* com nossa heurística física.

A nova função de recompensa imediata $\tilde{R}$ fornecida ao agente é:

$$\tilde{R}(x, u, x') = R_{env}(x, u) + \gamma\Phi(x') - \Phi(x), \tag{4}$$

onde:

- $R_{env}$ é a recompensa esparsa do ambiente (ex: +100 na meta, -10 em colisão).
- $\gamma$ é o fator de desconto.
- $\Phi(x) = -h_{phy}(x)$ é a função de potencial derivada da Eq. (**??**).

### 2.7 Algoritmo de Treinamento

O agente de RL (implementado via SAC) aprende uma política $\pi_\theta(u|s_{aug})$. O estado aumentado fornecido à rede neural é $s_{aug} = [z, o_{local}]$, composto pelas coordenadas privilegiadas e pelas leituras locais de sensores (Lidar/Grid).

**Vantagem Teórica:** De acordo com o Teorema 4.1 de Cheng et al., o uso de $h_{phy}$ reduz o "horizonte efetivo" do problema. Como $h_{phy}$ é uma aproximação admissível da dinâmica do veículo, o termo $\gamma\Phi(x') - \Phi(x)$ atua como uma densificação da recompensa que guia o agente continuamente em direção à meta respeitando a cinemática, enquanto o componente $R_{env}$ ensina o agente a desviar de obstáculos não previstos pela heurística analítica.

Desta forma, a abordagem é robusta: se a heurística falhar (ex: um obstáculo em U), o RL aprende a compensar o resíduo (conceito de *Improvable Heuristic*); onde a heurística é precisa (espaço livre), o RL converge quase instantaneamente.

## 3 Resultados

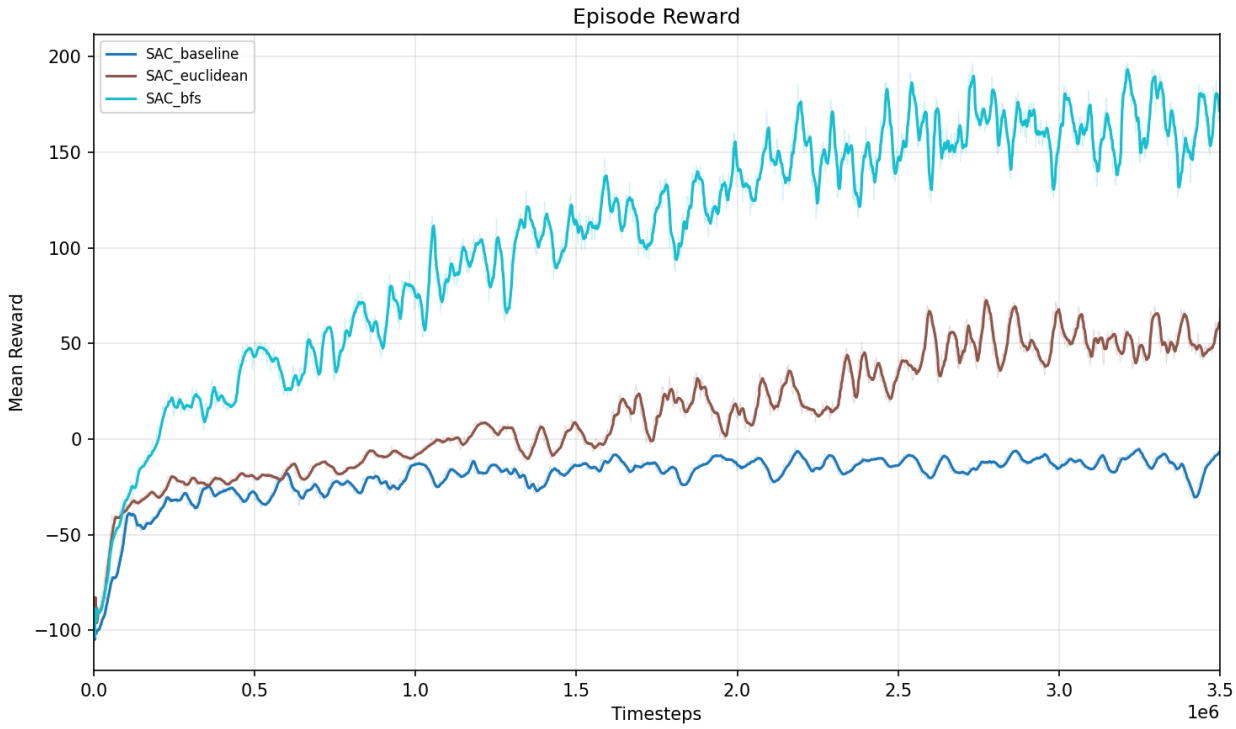


Figure 3: Recompensa por episódio dos agentes treinados com heurísticas euclidiana e menor caminho (BFS) com baseline

A análise das curvas de aprendizado evidencia a superioridade da heurística BFS no reward shaping. O modelo SAC + BFS apresentou convergência robusta para valores elevados de recompensa. Em contrapartida, o modelo SAC + Euclidiano estagnou em patamares inferiores, limitado por mínimos locais. O Baseline estabilizou em valores negativos, indicando o aprendizado de uma política de "sobrevivência" (evitar colisões) sem, contudo, solucionar a tarefa de navegação.

Definida como a conclusão efetiva da manobra de estacionamento, a taxa de sucesso corrobora a eficácia da informação topológica. A heurística BFS obteve desempenho significativamente superior, superando a complexidade dos obstáculos. A heurística Euclidiana demonstrou eficácia limitada, restringindo-se majoritariamente a cenários onde o vetor de direção ao objetivo não apresentava obstruções físicas.

O gráfico de duração de episódio revela outro aspecto; a heurística mais 'inteligente' não apenas elevou a taxa de sucesso, mas também a velocidade de resolução do objetivo, visto que a linha referente ao agente com heurística BFS apresenta um nível menor que as demais. Isso significa que o agente com heurística BFS tende a chegar ao objetivo em menos passos temporais.

## 4 Conclusão

O presente estudo demonstrou a eficácia do framework Heuristic Guided Reinforcement Learning (HuRL) aplicado ao planejamento de movimento de veículos articulados, um domínio caracterizado por dinâmica não-holonômica e horizontes longos. Os resultados indicam que a aplicação de reward shaping baseado em potencial é determinante para superar a esparsidade de recompensas, viabilizando a convergência do algoritmo SAC onde a abordagem baseline convergiu apenas para comportamentos de sobrevivência (não-colisão).

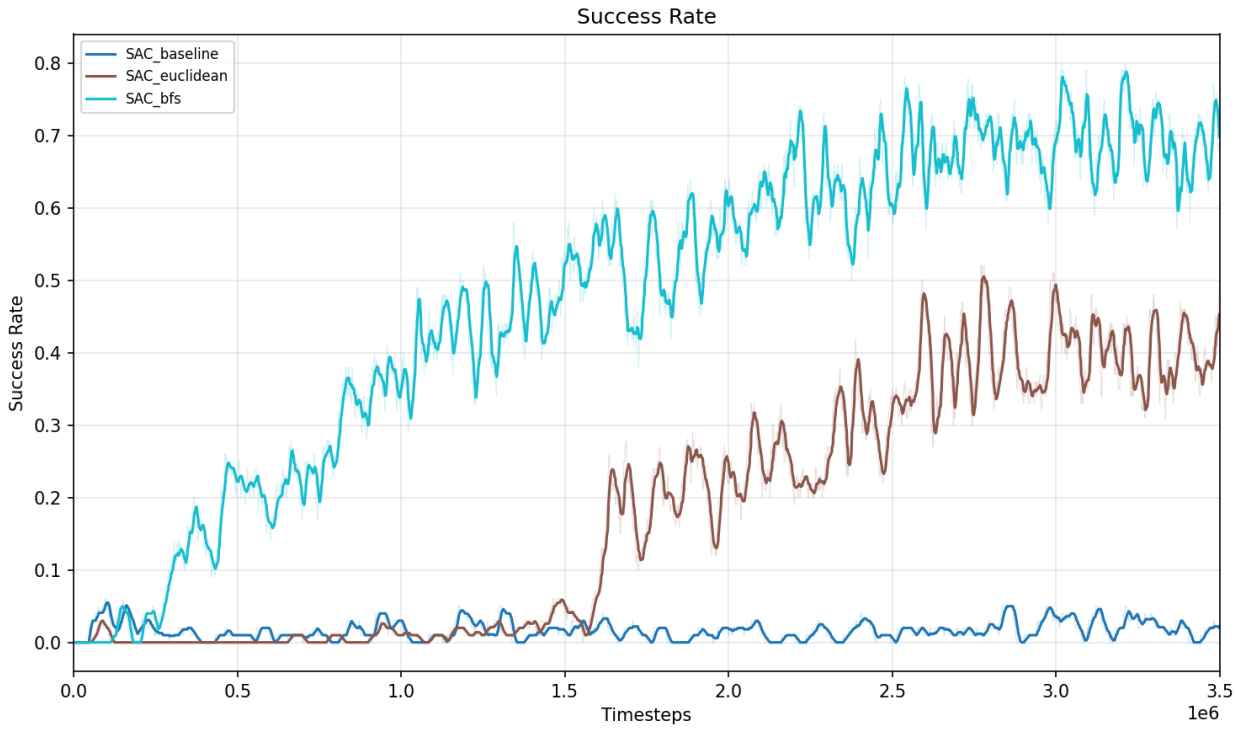


Figure 4: Taxa de sucesso dos agentes treinados com heurísticas euclidiana e menor caminho (BFS) com baseline

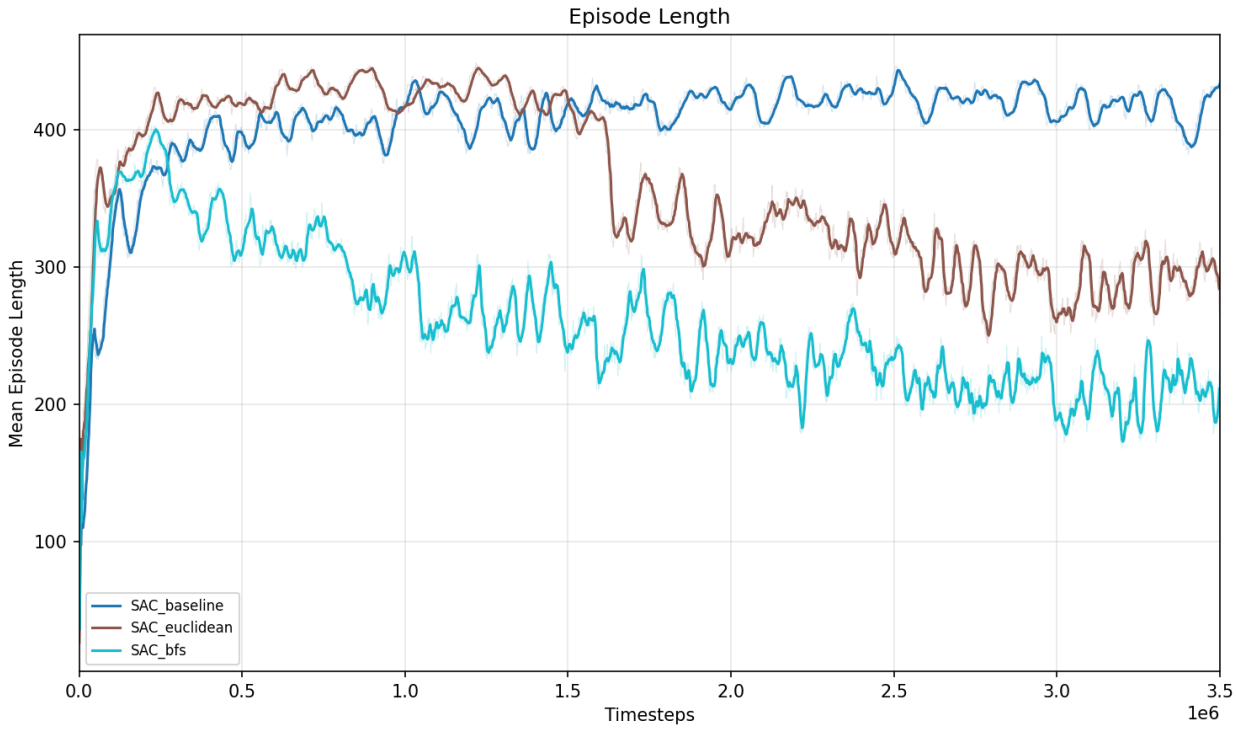


Figure 5: Duração de episódio dos agentes treinados com heurísticas euclidiana e menor caminho (BFS) com baseline

A análise comparativa evidenciou que o ganho de desempenho é fortemente correlacionado à fidelidade topológica da informação heurística. Enquanto a heurística Euclidiana sofreu com mínimos locais, a abordagem baseada em

BFS mostrou-se superior ao incorporar a geometria dos obstáculos no cálculo do potencial, fornecendo gradientes de recompensa densos e consistentes. Conclui-se, portanto, que a hibridização de métodos de busca clássicos com Aprendizado por Reforço Profundo constitui uma estratégia robusta e necessária para a resolução eficaz de tarefas de navegação complexa.

### 4.1 Trabalhos futuros

Com base nos resultados obtidos e nas limitações identificadas, propõem-se as seguintes direções para a continuidade da pesquisa:

- Coordenadas Privilegiadas: No decorrer do desenvolvimento, tomamos conhecimento das coordenadas privilegiadas [5], que podem trazer informações diretivas para o modelo que tem potência de levar a um aprendizado mais rápido, ficando como trabalho futuro.
- Outras arquiteturas de RL: Convém investigar o desempenho de outros algoritmos off-policy (ex: TD3) ou on-policy (ex: PPO), bem como abordagens baseadas em modelo (Model-Based RL), sob o paradigma HuRL.
- Incorporação de Restrições Não-Holonômicas na Heurística: A substituição do BFS (que assume movimento omnidirecional) por algoritmos como *Hybrid A** ou *RRT**, que consideram o raio mínimo de curvatura do veículo, poderia gerar potenciais de recompensa ainda mais informativos, especialmente em manobras de precisão.

**Nota:** O desenvolvimento dos algoritmos e parte da redação foram apoiados por modelos de linguagem de grande escala (LLMs), utilizados como ferramenta para acelerar a prototipagem, o tuning de hiperparâmetros e a documentação dos experimentos.

## References

[1] Ching-An Cheng, Andrey Kolobov, and Adith Swaminathan. Heuristic-guided reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13550–13563, 2021.

[2] Mark Towers, Ariel Kwiatkowski, Jordan Terry, John U Balis, Gianluca De Cola, Tristan Deleu, Manuel Goulão, Andreas Kallinteris, Markus Krimmel, Arjun KG, et al. Gymnasium: A standard interface for reinforcement learning environments. *arXiv preprint arXiv:2407.17032*, 2024.

[3] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[4] Chenyong Guan and Yu Jiang. A tractor-trailer parking control scheme using adaptive dynamic programming. *Complex & Intelligent Systems*, 8(3):1835–1845, 2022.

[5] Mario Rosenfelder, Henrik Ebel, Jasmin Krauspenhaar, and Peter Eberhard. Model predictive control of non-holonomic systems: Beyond differential-drive vehicles. *Automatica*, 152:110972, 2023.