Aprendizado por Reforço Baseado em Modelo para Controle de Baixo-Nível de Robôs com Pernas

Francisco Affonso Pinto

Marcelo Becker

Escola de Engenharia de São Carlos

francisco.affonso02@usp.br

Objetivos

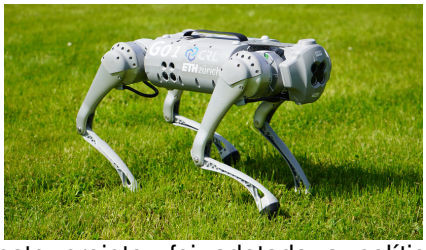
O controle de baixo-nível em sistemas robóticos, como quadrúpedes, torna-se complexo devido a necessidade de controlar estados que envolvem relações altamente não-lineares. Para lidar com essa complexidade, o estado da arte tem recorrido a controles preditivos, com base em modelos dinâmicos simplificados (4).

Nesse contexto, visando o desenvolvimento de um controlador sem essas restrições, o objetivo desse projeto é empregar um *Model-Based Reinforcement Learning* (MB-RL) como controlador de baixo-nível de quadrúpedes para melhorar a performance dos controladores atuais, utilizando sua capacidade de exploração.

Métodos e Procedimentos

Ao utilizar o aprendizado por reforço (RL), é fundamental formular o problema como um Processo de Decisão de Markov (MDP). Nesse contexto, o problema é modelado por um espaço de observações que descreve o estado do sistema, as ações que ele pode executar, e uma função de recompensa que avalia o desempenho em relação à tarefa desejada (controle de baixo nível). Assim, o processo é abordado treinando uma política que maximiza a soma esperada das recompensas.

Figura 1: Unitree Go1 (Imagem por Kang et al. 2023)



Neste projeto, foi adotada a política *Soft Actor-Critic* (SAC) (1), devido à sua eficiência superior em relação a outros modelos. A SAC utiliza duas redes neurais: uma para gerar as ações de controle e outra para avaliar seu desempenho.

Agora será descrito como o processo MDP é interpretado neste projeto. Para o espaço de observação, foram utilizadas as seguintes informações sobre o robô:

$$o_k = [g, q, \dot{q}, v_{xyz}, \omega_{xyz}, a_k, a_{k-1}, cmd] \quad (1)$$

onde g é o vetor gravidade, q a posição das juntas, $\dot{q}$ as velocidades das juntas, v_{xyz} a velocidade linear da base, ω_{xyz} a velocidade angular da base, a_k a ação atual, a_{k-1} a ação anterior, e cmd o comando de baixo nível.

Continuando, a ação de controle é referente a posição de cada junta, enquanto o comando de baixo-nível representa as velocidades lineares e angulares que a base do robô deve atingir.

As funções que recompensam a exploração do robô foram baseadas em (3), incluindo tanto o rastreamento direto do comando de baixo nível quanto funções auxiliares que penalizam comportamentos indesejados, como saltos.

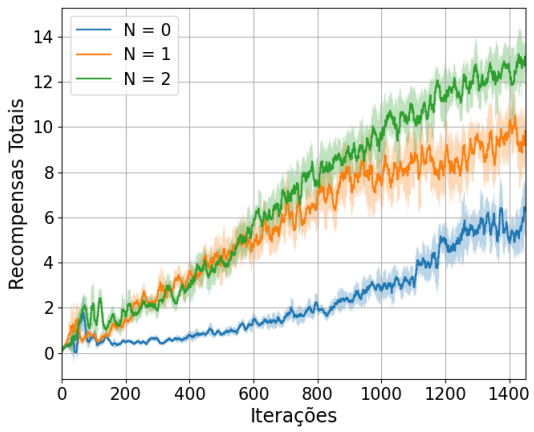
Além disso, foi utilizado uma rede neural *Multi-Layer Perceptron* (MLP) para aprender predizer os passos futuros do robô, reiterando a capacidade das redes neurais profundas de serem aproximadores universais de funções. Essa rede conseguirá predizer informações uteis para o treinamento da política.

Por fim, o treinamento da política foi inspirado em (2), onde o modelo aprendido é usado para ampliar o conjunto de dados utilizado no treinamento da política. A cada iteração, é gerado um dataset inicial a partir de um número fixo de passos. Esse conjunto é utilizado para treinar o MLP, que em seguida é empregado para gerar N passos adicionais, formando o dataset final que será utilizado no treinamento da política SAC.

Resultados

Como objeto de estudo para os resultados foi utilizado o quadrúpede Unitree Go1, ilustrado na Fig. 1, e o treinamento do método proposto foi realizado no simulador Isaac Gym.

Figura 2: Resultado do aprendizado da política conforme N passos adicionais



Foram utilizados 1500 robôs operando simultaneamente, com cada iteração consistindo no treinamento da política com 18 passos reais, acrescidos de N passos artificiais. O desempenho do método está ilustrado na Fig. 2.

Conclusões

Ao analisar os resultados, percebe-se que o método proposto alcança maiores valores de recompensa com menos iterações. Isso é resultado do uso do modelo aprendido para expandir o conjunto de dados de treinamento, o que acelera o processo de aprendizado da política.

Agradecimentos

Os autores agradecem à Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pelo auxílio financeiro concedido através da Bolsa de Iniciação Científica (2023/17678-1).

Referências

[1] Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning* (2018), PMLR.

[2] Janner, M., Fu, J., Zhang, M., and Levine, S. When to trust your model: Model-based policy optimization. *Advances in neural information processing systems 32* (2019).

[3] Margolis, G. B., and Agrawal, P. Walk these ways: Tuning robot control for generalization with multiplicity of behavior. In *Conference on Robot Learning* (2023), PMLR.

[4] Winkler, A. W., Bellicoso, C. D., Hutter, M., and Buchli, J. Gait and trajectory optimization for legged systems through phase-based end-effector parameterization. *IEEE Robotics and Automation Letters* (2018).