Preference-Based Learning for Dynamic Bipedal Locomotion

Maegan Tucker[1], Noel Csomay-Shanklin[2], and Aaron D. Ames[1,2]

I. Background

While model-based gait generation and control methods yield desirable theoretical properties such as safety certificates, they are sensitive to model uncertainty induced by real-world effects such as slipping, compliance, and flexibilities. In response, the machine learning community has developed several model-free approaches to locomotion including reinforcement learning, and imitation learning. However, these methods often fail when deployed on hardware due to their reliance on reward functions and simulation environments.

To leverage the advantages of model-based methods while accounting for real-world dynamics, we propose a preference-based learning framework that uses sequential human feedback to systematically realize dynamic locomotion directly on hardware. The main advantages of this approach are that it only relies on easy to provide subjective human feedback, mainly pairwise preferences, and it provides actions to sequentially sample, eliminating the guess work of manual tuning.

II. Methods

The objective of the proposed preference-based learning framework is to use subjective feedback to identify the optimal action $a^* = \text{argmin}_{a \in \mathbb{R}^d} U(a)$ of an unknown utility function $U : \mathbb{R}^d \to \mathbb{R}$, in as few iterations as possible. The summarized procedure of each iteration is to 1) select a new action to execute on the system using Thompson sampling, 2) obtain subjective feedback based on the experimental performance, and 3) infer a Bayesian posterior over the utilities $\boldsymbol{U} : \mathbf{A} \to \mathbb{R}$ corresponding to a discrete set of actions $a \in \mathbf{A} \subset \mathbb{R}^d$. Inspired by [1], we model the posterior probability as:

$$\mathcal{P}(\boldsymbol{U} \mid \mathbf{D}) \propto \mathcal{P}(\mathbf{D} \mid \boldsymbol{U})\mathcal{P}(\boldsymbol{U}),$$

where $\mathbf{D}$ is the dataset of user feedback, $\mathcal{P}(\mathbf{D} \mid \boldsymbol{U})$ is the likelihood function, and $\mathcal{P}(\boldsymbol{U})$ is the Gaussian process prior. For more details on the learning framework refer to [2], [3].

III. Experimental Results

We demonstrate the preference-based learning framework towards two model-based methods. First it was applied towards tuning the essential constraints of an HZD optimization problem to realize stable and robust locomotion on AMBER-3M with spring feet (shown in Fig. 1a), without accounting for the added compliance in the model or the controller [2]. Second, the method was applied towards identifying ID-CLF-QP$^+$ controller gains (shown in Fig. 1b) that stabilized the outputs of the 3D biped Cassie without torque chatter [3]. Future work includes unifying the framework to optimize both the generated gaits and the controller simultaneously.

This research was supported by NSF Graduate Research Fellowship No. DGE-1745301, and the Caltech Big Ideas and ZEITLIN Funds.

[1]Authors are with the Department of Mechanical and Civil Engineering, California Institute of Technology, Pasadena, CA 91125.

[2]Authors are with the Department of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, CA 91125.

References

[1] W. Chu and Z. Ghahramani, "Preference learning with gaussian processes," in *Proceedings of the 22nd international conference on Machine learning*, 2005, pp. 137–144.

[2] M. Tucker, N. Csomay-Shanklin, W.-L. Ma, and A. D. Ames, "Preference-based learning for user-guided hzd gait generation on bipedal walking robots," *arXiv preprint arXiv:2011.05424*, 2020.

[3] N. Csomay-Shanklin, M. Tucker, M. Dai, J. Reher, and A. D. Ames, "Learning controller gains on bipedal walking robots via user preferences," *arXiv preprint arXiv:2102.13201*, 2021.

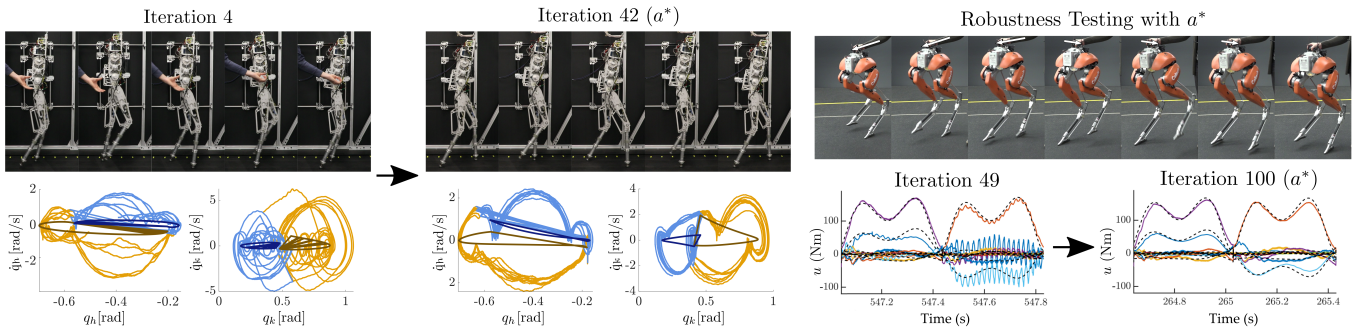

Fig. 1. Experimental results of the preference-based learning framework towards: a) Tuning constraints of the HZD gait generation framework on AMBER-3M with unmodeled spring feet; and b) Tuning ID-CLF-QP$^+$ controller gains on Cassie, with robustness testing of the optimal action a^*.