Active Robot Learning from Demonstrations for Sub-Optimal Human Teaching

Muhan Hou[a,*]

[a]Department of Computer Science, Vrije Universiteit Amsterdam
ORCID (Muhan Hou): https://orcid.org/0009-0008-2195-5224

Abstract. Learning from Demonstrations (LfD) allows robots to acquire skills for various tasks by imitating how humans perform them. However, untrained human users are not necessarily good teachers for the robot learners. Even though they may be domain experts in the target task itself (i.e., optimal in performing the task), the overall distribution of the demonstrations they provide may not be most beneficial for robot learning (i.e., sub-optimal in teaching the task). To achieve robust robot learning under sub-optimal human teaching while taking human factors into account, my research focuses on developing active LfD algorithms that empower robots to take more initiative by actively querying human demonstrations that may better support robot learning. Additionally, my research examines human factors beyond user experience and further investigates how active LfD may influence human teaching strategy after experiencing active guidance, attempting to extend the active LfD paradigm to foster a reciprocal learning loop between human teachers and the robot learner.

1 Introduction

Learning from Demonstrations (LfD) has achieved great success in robotics, enabling robots to acquire all kinds of skills by imitating how humans perform the tasks [2, 16, 17, 26]. However, humans are not necessarily good teachers for robots. Even when human users are domain experts in the target task and able to demonstrate the optimal action to take for every state encountered in their demonstrations (i.e., optimal in performing the task), the overall distribution of demonstrations they choose to provide may not be optimal for robot learning (i.e., sub-optimal in teaching the task) [12].

One intuitive strategy is to adopt the uniform distribution as the target for human demonstrations and cover as many diverse areas of the task space as possible. However, without proper guidance, the natural distribution of human demonstrations often tends to be imbalanced and biased [11, 27], which can be detrimental to robot learning [5, 29]. To actively guide human teachers to overcome sub-optimal teaching, my research begins with the conventional LfD setting where demonstrations are provided before policy learning (i.e., offline demonstrations) and has developed an algorithm that queries offline demonstrations from humans and shapes them toward target distributions to better facilitate robot learning [11].

However, the uniform coverage strategy may not necessarily be optimal for robot learning [12]. Certain critical areas of the state space that are encountered less frequently can be much harder for the control policy to generalize to (e.g., encountering an oncoming vehicle in autonomous driving). These areas may require more demonstrations than more common but simpler situations (e.g., driving straight with no vehicles nearby). Furthermore, the probability distribution of the robot running into different areas of the state space is dependent on the evolving control policy, which is non-stationary and constantly updates its action distributions over states throughout learning. Therefore, it is impractical to determine the optimal distribution of human demonstrations a priori before robot learning begins [12, 13].

Instead of learning with *offline* human demonstrations in a *demo-then-training* manner, previous efforts [6, 7, 20] have been made to learn with *online* human teaching input in a *demo-while-training* manner. However, similar to the issues in the settings with offline human demonstrations, what human teachers choose to demonstrate may not be most beneficial to robot learning. Situations that human teachers perceive as easy to learn may prove difficult for learning agents to generalize, and vice versa. And this mismatch may consequently lead to redundant or missing coverage of the state space. Furthermore, since human teachers are expected to provide teaching input while the robot is learning the task, deciding when to intervene becomes an additional challenge for them. This may lead human teaching to be even more sub-optimal without any external guidance. To address sub-optimal human teaching with online demonstrations, the second phase of my research enables the robot to take an active role in the learning process by designing active LfD algorithms that optimize the timing and content of the robot queries for episodic human demonstrations [12, 14].

In addition to the robot learning performance, human factors are also critical to take into account when a human teacher is involved in the learning loop. Previous work [12, 20] has made considerable efforts to improve user experience, but has underexamined other equally important aspects beyond it, such as the influence on human teaching strategies after experiencing robot guidance. To this end, the third phase of my research further extends its focus to reciprocal human-robot interactive learning, designing active LfD algorithms that aim to benefit both robot learning and human teaching.

To summarize, aiming to overcome sub-optimal human teaching in the context of robot learning from human demonstrations, my research develops active learning algorithms that may: 1) actively shape the imbalanced distribution of offline human demonstrations for the conventional demo-then-training LfD setting; 2) optimize the timing and content of queries for online human demonstrations to improve robot learning performance; 3) take into account the influence

[*] Email: m.hou@vu.nl.

of LfD algorithm design on human teaching and build a reciprocal learning loop between human teachers and the robot learner.

2 Active Learning with Offline Human Demonstrations

Aiming to overcome the sub-optimal human teaching in the conventional LfD setting with offline human demonstrations, one of my research has developed an active learning algorithm that enables the robot to shape the distribution of human demonstrations by actively guiding its interaction with humans [11]. Previous work attempted to solve the imbalanced demonstration distribution after the data are already collected, either via data curating approaches [1, 19, 21] or revising the cost function to unbias policy learning [5, 9, 29]. Few of them paid careful attention to data abundance during the data acquisition process and put efforts to maintain a balanced data distribution from the early phase.

To solve the demonstration imbalance in the early phase of data acquisition, we explicitly took into account the influence of robots on human teacher behaviors and enabled the robot to actively guide its interaction with humans to shape the distribution of collected data [11]. More specifically, we formalized such an active data collection process into a discrete finite-horizon Markov Decision Process (MDP) to maintain data balance against uncertainties during the data collection process. Results for the experiments of simulated data collection verified our method's generalization capability to actively shape the resulting distribution into various target distributions, along with its robustness to different levels of uncertainties during the data collection process. Furthermore, we verified our method's efficacy in real-world robot tasks and demonstrated improved robot learning performance in unseen situations when robot policies were trained with demonstrations of more balanced distributions shaped by our active data collection method.

3 Active Learning with Online Human Demonstrations

Due to the intractability of determining the optimal demonstration distribution a priori, my research has also developed algorithms that optimize the sequence of robot queries for episodic human demonstrations by actively deciding both when and what to query throughout the learning process.

Previous work [4, 6, 7, 8, 18, 23, 25] attempted to solve the problem by querying isolated state-action pairs from human teachers whenever the robot perceives high uncertainty. However, these approaches require frequent context switching, which imposes high cognitive demands on humans and increases the risk of errors or noise in providing immediate interventions. To overcome these challenges, my research has developed an active learning algorithm that enables the robot to actively request episodic demonstrations (i.e., from an initial state to a terminal state) for better learning performance and improved user experience [12]. More specifically, we construct a trajectory-based uncertainty measurement of the robot policy based on temporal difference errors of episodic policy roll-out and utilize it to optimize the decision of when to query and what to query in a trajectory-based feature space. We test our method on three simulated navigation tasks with sparse rewards, continuous state-action spaces, and increasing levels of difficulty. Results indicate that our method converges to expert-level performance significantly faster in both experiments with oracle-simulated demonstrators and real human expert demonstrators, while also achieving improved perceived task load and consuming significantly less human time.

To further explore the potential of active LfD in robotics, my research has also extended beyond the learning-from-scratch setting to the policy transfer scenario, where the common issue of covariance shift in LfD can be further exacerbated by discrepancies in task space between the source and target domains [14]. We have extended our EARLY framework to the transfer learning scenario and simultaneously optimized the problems of when to query and what episodic demonstrations of the target task to query during the policy transfer process. We validate the effectiveness of our approach in 8 robot policy transfer scenarios, involving policy transfer across diverse environment characteristics, task objectives, and robotic embodiments. Results from simulations and preliminary sim-to-real experiments demonstrate that our method achieves significantly higher success rates and greater sample efficiency in target tasks that other baseline methods struggle to address.

4 Reciprocal Human-Robot Interactive Learning

Going beyond common human factors like user experience, my research also examines the influence of active guidance on human teaching, aiming to design an active LfD algorithm that may benefit both robot learning and human teaching. Specifically, this work is inspired by Curriculum Learning (CL) [3], the general idea of which is to break the original hard task into a sequence of sub-tasks with gradually increasing difficulty. Such scaffolding not only benefits robot learning [10, 28], but also is intuitive to common human users as it well aligns with how humans naturally learn new tasks [15, 24].

In our approach [13], we construct a sequence of curricula with gradually increasing difficulties by resetting the environment to states that are progressively farther away from the task goal area and closer to the initial state space of the original task. Queries for demonstrations will be iteratively generated based on the latest maintained curricula, which will not only gather demonstrations for the robot to imitate but also automate the curriculum expansion and iteration process, guiding policy exploration. We evaluate our method on four simulated robotic tasks with sparse rewards, achieving much better converging policy performance and sample efficiency. A further user study shows that our method takes less human time and fewer failed demonstration attempts while improving per-guidance teaching performance, post-guidance teaching adaptability, and teaching transferability to unseen tasks.

5 Future Work

Our work so far has assumed that human teachers are domain experts. In reality, they are often neither experts in task execution nor in teaching. Future research will aim to design active LfD algorithms that address this dual sub-optimality by optimizing queries based not only on the demonstrations most beneficial for policy learning but also on the human teacher's capacity to provide them.

Additionally, our current work assumes access to a reset function that allows the environment to be initialized from any desired state for querying demonstrations, which is impractical in real-world settings. Our future work will reframe the problem within the context of Autonomous Reinforcement Learning [22] without assuming a reset function. We aim to develop active querying strategies that balance the reachability of queried states under the current policy with the expected learning gains, enabling the robot to learn both environment resetting and the primary task itself.

References

[1] A. Amini, W. Schwarting, G. Rosman, B. Araki, S. Karaman, and D. Rus. Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 568–575. IEEE, 2018.

[2] P. J. Ball, L. Smith, I. Kostrikov, and S. Levine. Efficient online reinforcement learning with offline data. In *International Conference on Machine Learning*, pages 1577–1594. PMLR, 2023.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009.

[4] M. Cakmak and A. L. Thomaz. Active learning with mixed query types in learning from demonstration. In *Proc. of the ICML workshop on new developments in imitation learning*. Citeseer, 2011.

[5] B. César-Tondreau, G. Warnell, K. Kochersberger, and N. R. Waytowich. Towards fully autonomous negative obstacle traversal via imitation learning based control. *Robotics*, 11(4):67, 2022.

[6] M.-H. Chen, S.-A. Chen, and H.-T. Lin. Active reinforcement learning from demonstration in continuous action spaces. In *AI and HCI Workshop at the 40th International Conference on Machine Learning (ICML), Honolulu, Hawaii, USA*, 2023.

[7] S.-A. Chen, V. Tangkaratt, H.-T. Lin, and M. Sugiyama. Active deep q-learning with demonstration. *Machine Learning*, 109(9):1699–1725, 2020.

[8] S. Chernova and M. Veloso. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research*, 34:1–25, 2009.

[9] D. Dresvyanskiy, W. Minker, and A. Karpov. Deep learning based engagement recognition in highly imbalanced data. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings 23*, pages 166–178. Springer, 2021.

[10] C. Florensa, D. Held, M. Wulfmeier, M. Zhang, and P. Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*, pages 482–495. PMLR, 2017.

[11] M. Hou, K. Hindriks, A. Eiben, and K. Baraka. Shaping imbalance into balance: Active robot guidance of human teachers for better learning from demonstrations. In *2023 32nd IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1737–1744. IEEE, 2023.

[12] M. Hou, K. Hindriks, G. Eiben, and K. Baraka. "give me an example like this": Episodic active reinforcement learning from demonstrations. In *Proceedings of the 12th International Conference on Human-Agent Interaction*, pages 287–295, 2024.

[13] M. Hou, K. Hindriks, A. Eiben, and K. Baraka. Active robot curriculum learning from online human demonstrations. In *2025 20th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 810–818. IEEE, 2025.

[14] M. Hou, K. Hindriks, A. Eiben, and K. Baraka. Robot policy transfer with online demonstrations: An active reinforcement learning approach. In *2025 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2025.

[15] K. A. Krueger and P. Dayan. Flexible shaping: How learning in small steps helps. *Cognition*, 110(3):380–394, 2009.

[16] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 6292–6299. IEEE, 2018.

[17] A. Nair, A. Gupta, M. Dalal, and S. Levine. Awac: Accelerating online reinforcement learning with offline datasets. *arXiv preprint arXiv:2006.09359*, 2020.

[18] B. Packard and S. Ontanón. Policies for active learning from demonstration. In *AAAI Spring Symposia*, 2017.

[19] M. J. Procopio, J. Mulligan, and G. Grudic. Coping with imbalanced training data for improved terrain prediction in autonomous outdoor robot navigation. In *2010 IEEE International Conference on Robotics and Automation*, pages 518–525. IEEE, 2010.

[20] M. Rigter, B. Lacerda, and N. Hawes. A framework for learning from demonstration with minimal human effort. *IEEE Robotics and Automation Letters*, 5(2):2023–2030, 2020.

[21] G. Sawadwuthikul, T. Tothong, T. Lodkaew, P. Soisudarat, S. Nutanong, P. Manoonpong, and N. Dilokthanakul. Visual goal human-robot communication framework with few-shot learning: a case study in robot waiter system. *IEEE Transactions on Industrial Informatics*, 18(3): 1883–1891, 2021.

[22] A. Sharma, K. Xu, N. Sardana, A. Gupta, K. Hausman, S. Levine, and C. Finn. Autonomous reinforcement learning: Formalism and benchmarking. In *International Conference on Learning Representations*, 2021.

[23] D. Silver, J. A. Bagnell, and A. Stentz. Active learning from demonstration for robust autonomous navigation. In *2012 IEEE International Conference on Robotics and Automation*, pages 200–207. IEEE, 2012.

[24] B. F. Skinner. Teaching machines. *Scientific American*, 205(5):90–106, 1961.

[25] K. Subramanian, C. L. Isbell Jr, and A. L. Thomaz. Exploration from demonstration for interactive reinforcement learning. In *Proceedings of the 2016 international conference on autonomous agents & multiagent systems*, pages 447–456, 2016.

[26] M. Vecerik, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothörl, T. Lampe, and M. Riedmiller. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*, 2017.

[27] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi. Delving into deep imbalanced regression. In *International conference on machine learning*, pages 11842–11851. PMLR, 2021.

[28] Y. Zhang, P. Abbeel, and L. Pinto. Automatic curriculum learning through value disagreement. *Advances in Neural Information Processing Systems*, 33:7648–7659, 2020.

[29] W. Zhou, S. Bohez, J. Humplik, N. Heess, A. Abdolmaleki, D. Rao, M. Wulfmeier, and T. Haarnoja. Forgetting and imbalance in robot lifelong learning with off-policy data. In *Conference on Lifelong Learning Agents*, pages 294–309. PMLR, 2022.