https://doi.org/10.31256/Us7Hc6M

SUB-SPARC: Investigation of Imperfect Teachers

Garry Clawson[1]
School of Computer Science,
University of Lincoln, UK
18685030@students.lincoln.ac.uk

***Abstract*—Early literacy and language skills are a significant precursor to children's later educational success. Using social robots as a learning companion is one method to augment efforts to support this aim. SPARC has been shown as a useful way to combine reinforcement learning with human experts to reduce the state search space and for human experts to intervene before any negative reward actions take place, however human expertise is not consistent. In this paper we investigate if SPARC can be counter-useful as expertise levels decline. Results from the study suggest that once a skill level below 90% of an expert is breached, alternative LfD methods maybe more suitable.**

I. Introduction

Early literacy and language skills are a significant precursor to children's later educational success. High quality pre-school programs that provide social interaction, alphabetic and vocabulary knowledge can prevent academic failure in later years [1]. Social robots as learning companions may provide an opportunity to augment these efforts. Learning from Demonstration (LfD) techniques help robots to achieve this in complex social environments. However, the lack of real time control of the robot during LfD training where social robots are asked to perform tasks such as teaching or care giving, iterating through potential states and actions to identify a suitable policy, may be sub-optimal or even harmful [2]. Wizard-of-Oz (WoZ) [3] and Supervised Progressively Autonomous Robot Competencies (SPARC) [4] are both well known strategies to mitigate this by providing expert input at each stage of the state-action training iteration. However, these approaches do not take into account sub-optimal expert teacher performance. Here, we investigate a sub-optimal strategy using a modified SPARC algorithm, SUB-SPARC, to identify the impact on policy convergence when expert teachers have a variation in skill level.

II. Background

Previous research in planning approaches where control may switch back and forth between a human teleoperator and autonomous control are well surveyed [5]. However, few works utilising sub-optimal models of performance are available. Rigter et al. [6] employing a Multi-Arm Bandit (MAB) problem, considers episodic problems that lead to binary outcomes. They assume there is a cost for asking the human, and a cost for failing the episode. Iterative algorithms, such as DAgger [7], retrieves expert actions from the supervisor in all encountered states and aggregates the revised state-action pairs into the training data for later usage. Self-Imitation Learning by Planning (SILP) [8], an extension of DAgger, proposes a method to assist in reinforcement learning in motion planning tasks, while minimising extra computational burden on the trainer. Deschuyteneer [9] extends this work, and builds on SPARC to introduce several updates; Monte Carlo Update (QMC-SPARC), SPARC with Hindsight Experience Replay (HER-SPARC) and DQN (Deep Q-Learning) with SPARC. Sub-optimal teaching is also introduced using QMC-SPARC by replacing expert teacher guidance by a random action rate ϕ where $0 \leq \phi \leq 1$. However, these studies do not address the challenge of a bounded skill level nor compare to a baseline model. This investigation aims to address this gap using a modified SPARC algorithm where a teacher is of bounded quality ϕ where, $\alpha \leq \phi \leq \beta$. This study aims to further the understanding that sub-optimal teaching may have on LfD policy generation.

III. Method

a) Experiment Design: A desktop grid world simulation utilising SPARC as a base LfD model was used. Q-Learning was completed on the grid-world environment to derive a best practice state-action pair, $Q(s, a)$, for every index position. The derived policy π was reviewed and substituted as an expert human teacher to allow the expert human intervention process to be automated. The derived Q-learning policy for the environment provided a baseline for comparison with, WoZ, SPARC original, In Real Life Person (IRLP), and the new sub optimal SPARC algorithm, known as SUB-SPARC going forwards. Learning was completed over 50 episodes.

b) Environment: A three row by four column grid world was instantiated as the environment. Index position [2,3] was the goal state with a reward of +1. Index position [1,3] was a failure state with a reward of -1. Index position [1,1] was also a negative reward position of -1. This position was created to force a maze like feature when traversing the grid world.

c) Algorithm: The SUB-SPARC algorithm (Algorithm 1), was a modified extension of SPARC, except rather than query a policy from a perfect Q-learning policy π that mimics the expert teacher, it utilised a random number ϕ where ϕ is $\alpha \leq \phi \leq \beta$. Both lower and upper bounds were defined to enable a skill level range to be assigned.

[1]This work was supported by the Engineering and Physical Sciences Research Council [EP/S023917/1]

Algorithm 1: Algorithm used in SUB-SPARC

while *learning* **do**
  a = action with the highest $Q(s, a)$ value look at location used with a;
  while *query expert teacher policy* π **do**
    ϕ = *random value;*
    if $\alpha \leq \phi \leq \beta$ **then**
      a = random assigned action, r = -0.05;
    else
      a = expert policy assigned action;
      reward, r = 0.05;
    end
    execute a, and transition to s' $Q(s, a) \leftarrow Q(s_t, a_t) + \alpha\ (r_{t+1} + \gamma\ (max_a\ Q(s_t, a)) - Q(s_t, a_t))$
  end
end

IV. Results and Discussion

Figure 1 shows three comparative learning algorithms and Q-Learning used as a baseline for learning performance. The fewer actions an algorithm requires to achieve a high stable reward, the better the learning from demonstration performance.

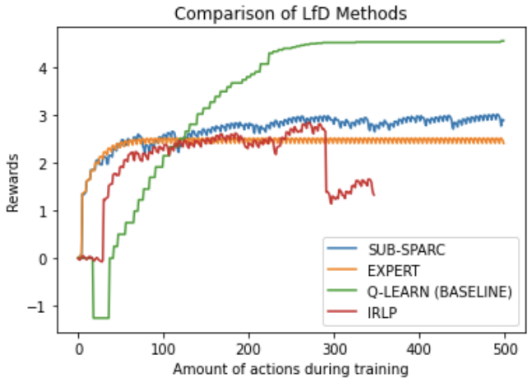

Fig. 1. Comparison of performance of LfD methods plotted against amount of actions taken during training and associated rewards per episode received. The fewer actions an algorithm requires to achieve a high stable reward the better. An In Real Life Person (IRLP) is also shown to provide further comparison.

Figure 2 shows SUB-SPARC set at varying teacher expertise levels with 0% being expert and 100% fully random actions. Utilising a static start point in the grid-world environment meant that once an expert policy was introduced this was repeated for every episode. This yielded the fastest learning rate in the fewest steps possible out performing the Q-Learn process. SUB-SPARC achieved comparative results to SPARC between $0-7\%$ error rate. Between $7\%-10\%$ error rate SUB-SPARC achieved equivalent results to the Q-Learn baseline. Once SUB-SPARC had exceeded a 20% error rate, it required over five times more actions to complete an episode compared to Q-Learn baseline. An expert level below 50% did not return any rewards greater than 0, which demonstrates that minimal learning took place. This suggests that a high quality of consistent human expert interaction during LfD is important to maintain the benefits of SPARC.

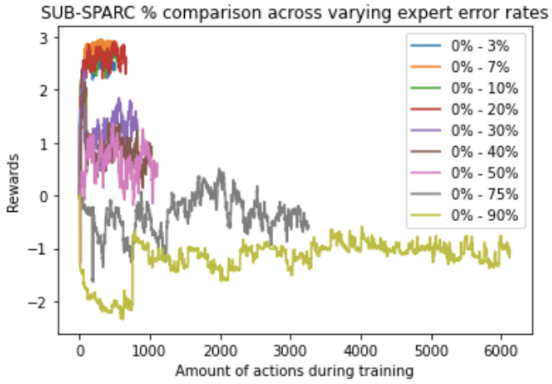

Fig. 2. Comparison of performance of the SUB-SPARC algorithm across varying teacher expertise levels (error rate) measured in %. 0% being expert and 100% being non expert (fully randomised actions within the grid-world). Shorter plots with higher rewards, indicate that the shorter trials achieve faster success, and termination of the episode.

V. Summary and Future Work

The findings of the study suggest that an expertise level below 90% performs worse than the baseline Q-Learn algorithm in a simple grid-world environment. This supports previous work by Deschuyteneer. One limitation of the SUB-SPARC model is the simplified exploration strategy using a static starting point, as well the low level of optimisation for next step selection other than a simple randomised action. Further work investigating the advantages of continually degrading skill levels through each episode of the learning process could more accurately reflect actual human performance over time.

References

[1] Park, H.W., Grover, I., Spaulding, S., Gomez, L. and Breazeal, C., 2019, July. A model-free affective reinforcement learning approach to personalization of an autonomous social robot companion for early literacy education. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, No. 01, pp. 687-694).

[2] Hedlund, E., Johnson, M. and Gombolay, M., 2021, March. The Effects of a Robot's Performance on Human Teachers for Learning from Demonstration Tasks. In Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (pp. 207-215).

[3] Riek, L.D., 2012. Wizard of Oz studies in hri: a systematic review and new reporting guidelines. *Journal of HRI*, 1(1), pp.119-136.

[4] Senft, E., Lemaignan, S., Baxter, P.E. and Belpaeme, T., 2016. SPARC: an efficient way to combine reinforcement learning and supervised autonomy. *Future of Interactive Learning Machines workshop*

[5] Lin, Z., Harrison, B., Keech, A. and Riedl, M.O., 2017. Explore, exploit or listen: Combining human feedback and policy model to speed up deep reinforcement learning in 3d worlds. *arXiv preprint* arXiv:1709.03969.

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

[7] Ross, S., Gordon, G. and Bagnell, D., 2011, June. A reduction of imitation learning and structured prediction to no-regret online learning. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 627-635).

[8] Luo, S., Kasaei, H. and Schomaker, L., 2021, May. Self-imitation learning by planning. In 2021 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4823-4829). IEEE.

[9] Deschuyteneer, J., 2020. Interactive reinforcement learning for real-time robot training. [Masters Thesis, Ghent University]