PAPER: (og - roxo; alterado - branco)

* Experimental Evaluation
* Simulation Environment
* PPO vs HER
* PPO vs SAC
* SAC vs HER
* Mean Reward and Success Rate Post-Train
* Table III and Figures
* Conclusion and Future Work
* EXPERIMENTAL EVALUATION

To assess the performance of the trained policies, a set of 100 test episodes were conducted for each training configuration, comparing both algorithms and training strategies. Specifically, each algorithm was evaluated with and without curriculum learning.

To ensure robust and reliable performance evaluation, each model variation (PPO, PPO with curriculum, SAC, SAC with curriculum, HER, and HER with curriculum) was independently trained 10 times. This was done to account for variability in training outcomes and reduce the likelihood of overestimating performance due to unusually good or bad runs. Following training, each of the 60 resulting models was evaluated over 100 test episodes, allowing for comprehensive comparisons across algorithms and training strategies, both with and without curriculum learning.

In addition to the final evaluation set, evaluation callbacks were performed every 10,000 timesteps during training.

In addition to the final evaluation set, evaluation callbacks were performed every 10,000 timesteps, per model, during training.

* SIMULATION ENVIRONMENT

\[

R\_{\text{dense}}(t) = 0.5 \cdot r\_{\text{vel}} + r\_{\text{time}} + 20 \cdot \text{passed\\_hoop}

\]

\[

R\_{\text{dense}}(t) = 0.5 \cdot r\_{\text{vel}} + r\_{\text{time}} + 20 \cdot \text{passed\\_hoop}

\]

* PPO vs HER

In figure \ref{fig:mr\_ppo\_her}, it is clear that HER with curriculum shows rapid early progress, reaching a relatively high reward within the first 50,000 timesteps. However, its learning curve becomes noisy and inconsistent, with visible fluctuations in mean reward across training. HER without curriculum performs much worse in early stages but gradually catches up, suggesting that HER benefits from goal relabeling but struggles with unstable gradient updates in complex conditions.

In contrast, both PPO variants show more gradual and stable improvement, with PPO without curriculum slightly outperforming PPO with curriculum by the end of training. This suggests that PPO is less sensitive to curriculum scheduling but can still benefit from a steady learning signal over longer horizons.

Figure \ref{fig:mr\_ppo\_her} underscores the algorithmic divide in training efficacy. HER’s boolean success metric (right axis) struggles markedly across both variants, with neither curriculum nor non-curriculum versions exceeding a 0.4 mean success rate. The absence of meaningful improvement suggests fundamental limitations in HER’s goal-relabeling approach for this task, failing to translate early exploration into consistent successes even at 400k timesteps.

PPO’s continuous reward (left axis) tells a divergent story: the curriculum variant begins outperforming its non-curriculum counterpart at ≈0.3×10⁶ timesteps. This reversal of the typical pattern highlights the task’s unique structure: the handcrafted curriculum likely provides critical intermediate goals that standard PPO exploration misses. The stability of both PPO curves (minimal variance post-convergence) further contrasts with HER’s erratic performance, reinforcing PPO’s suitability for tasks requiring precise, incremental policy optimization.

Surprisingly, curriculum learning fails to enhance HER’s performance while giving PPO a consistent but modest advantage after 300k timesteps. This unexpected outcome suggests that curricula's effectiveness depends more on an algorithm's update mechanism than on reward sparsity alone.

* PPO vs SAC

Figure \ref{fig:mr\_ppo\_sac} shows that SAC with curriculum learning exhibits significant instability in the early stages of training, with noticeable fluctuations in mean reward. In contrast, the other three configurations, SAC without curriculum, PPO with curriculum, and PPO without curriculum, demonstrate more stable and gradual learning curves throughout training.

Despite its stability, PPO requires substantially more timesteps to reach convergence, typically stabilizing after approximately 600,000 steps. On the other hand, SAC converges much earlier, around 90,000 steps, highlighting its superior sample efficiency. This efficiency makes SAC particularly well-suited for environments where faster convergence with fewer interactions is desirable, although its early-stage instability with curriculum indicates that tuning or additional regularization may be beneficial in structured training setups.

Figure \ref{fig:mr\_ppo\_sac} shows that, overall, both algorithms improve their performance over time, but distinct patterns emerge when curriculum learning is applied. SAC demonstrates a stronger initial performance advantage, especially without curriculum, achieving higher mean rewards early on. PPO, on the other hand, requires more timesteps to reach convergence, typically stabilizing after approximately 0.3×10⁶ steps. Despite this slower start, PPO with curriculum steadily improves and even slightly surpasses SAC in the later training stages, showing a strong final performance.

Curriculum learning appears to offer more consistent and significant benefits to PPO, helping reduce variance and accelerate learning during the early and middle phases. In contrast, the impact of curriculum on SAC is less pronounced, suggesting that SAC’s inherent sample efficiency may already mitigate the need for a staged learning progression. These observations indicate that curriculum learning is particularly beneficial for PPO, which relies more heavily on structured task progression to reach optimal performance.

* SAC vs HER

Figure \ref{fig:mr\_sac\_her} provides a direct comparison between SAC and HER. Based on the analysis of the training curves, SAC consistently outperforms HER, particularly in terms of stability and convergence behavior. While HER may exhibit faster early progress under certain conditions, SAC demonstrates more reliable and robust learning, reinforcing its advantage in this task.

Figure \ref{fig:mr\_sac\_her} illustrates that a clear contrast emerges between the two methods: while SAC shows steady and substantial improvements in mean reward across training timesteps, HER maintains a relatively low and fluctuating performance throughout.

Curriculum learning appears to benefit SAC modestly, providing a slight increase in performance and reduced variance early on. In contrast, HER's performance remains consistently low and noisy, regardless of curriculum use. While curriculum introduces minor improvements in HER's early learning phase, it fails to sustain long-term gains, and its variance remains high.

* Mean Reward and Success Rate Post-Train

In the post-training evaluation, HER with curriculum learning achieved a significantly lower mean reward (µ ≈ 0.3) compared to HER without curriculum (µ ≈ 0.6). (5) The success rate followed the same trend, with approximately 30% for HER with curriculum and 60% without (6). This consistent performance gap suggests that curriculum learning, while generally intended to guide learning through structured progression, may have negatively interfered with HER’s goal relabeling mechanism, particularly in this sparse-reward setting. As the goal distribution shifted across difficulties, it likely introduced inconsistencies in the replay buffer, resulting in less stable learning and reduced overall performance.

In the post-training evaluation, HER with curriculum learning achieved a substantially lower mean reward and success rate (µ ≈ 0.16) compared to HER without curriculum (µ ≈ 0.30), performing nearly half as well in terms of reward, contradicting the typical expectation that curriculum learning improves performance. While both versions of HER performed poorly overall, the consistent performance gap suggests that curriculum learning may have negatively interacted with HER’s goal relabeling strategy. Specifically, in a sparse-reward setting, the shifting goal distribution imposed by curriculum stages may have introduced inconsistencies into the replay buffer. These inconsistencies could have impaired the relevance and effectiveness of relabelled goals, leading to less stable learning and reduced overall performance

In contrast, PPO and SAC exhibited minimal differences between their curriculum and non-curriculum variants. For PPO, the success rate with curriculum was slightly lower (by approximately 0.1), while SAC showed virtually identical success rates and mean rewards across both conditions. This indicates that both algorithms are robust to curriculum variations, likely due to their reward shaping and exploration strategies being less sensitive to changes in task distribution. These findings underscore that while curriculum learning can be beneficial, its impact is highly dependent on the underlying algorithm, and in particular, goal-conditioned methods like HER may require more careful curriculum design to avoid destabilizing learning.

In contrast to HER, both PPO and SAC benefited from curriculum learning, though to different extents. PPO's success rate increased from 0.79 to 0.90 and its mean reward improved from 82.14 to 87.40 with curriculum, indicating a meaningful performance gain. Similarly, SAC saw an increase in success rate from 0.73 to 0.88 and a boost in mean reward from 79.33 to 85.72 when trained with curriculum. These results suggest that while PPO and SAC do not rely on curriculum to function effectively, they can leverage it to enhance stability and final performance, possibly due to their ability to exploit dense reward signals and structured exploration.

These findings reinforce that the effectiveness of curriculum learning is highly algorithm-dependent. While PPO and SAC gain measurable improvements, HER’s performance degrades under curriculum, likely due to the incompatibility between staged task progression and HER’s goal relabelling mechanism. This highlights the need for more careful curriculum design when integrating it with **goal-conditioned methods**, especially in sparse-reward settings.

* Conclusion and Future Work

The results demonstrate that SAC achieves superior sample efficiency compared to PPO, converging approximately six times faster while maintaining stable learning curves throughout training. This efficiency makes SAC particularly suitable for robotic applications where training time and computational resources are constrained. The off-policy nature of SAC also provides advantages in terms of data utilization and learning stability in continuous control domains. Surprisingly, the integration of curriculum learning with HER showed detrimental effects on performance, with success rates dropping from 60% to 30% when curriculum was applied. This finding suggests that goal relabeling mechanisms in HER may be sensitive to changes in task distribution during structured difficulty pro gression. The inconsistencies introduced in the replay buffer as goal distributions shift across curriculum levels appear to destabilize the learning process, highlighting the need for careful consideration when combining these techniques.

The results demonstrate that SAC exhibits greater sample efficiency compared to PPO, converging approximately three times faster while maintaining smoother and more stable learning curves. This efficiency reinforces SAC’s suitability for robotic applications where data collection is costly, and training time is constrained. Its off-policy nature allows for more effective reuse of past experiences, contributing to its learning stability in continuous control tasks. Additionally, both SAC and PPO benefited from curriculum learning, with SAC's success rate rising from 73% to 88%, and PPO's from 79% to 90%, underscoring curriculum's role in enhancing performance when paired with compatible learning strategies.

In contrast, the integration of curriculum learning with HER had a clearly detrimental effect. Both HER’s success rate and mean reward dropped to nearly halved. This suggests that HER's goal relabelling mechanism may be particularly sensitive to shifts in task distribution introduced by curriculum schedules. As the goal space evolves during training, the replay buffer may accumulate off-distribution or misaligned transitions, thereby undermining HER’s sample efficiency and stability. These results highlight the need for more sophisticated curriculum strategies when applying structured progression in goal-conditioned, off-policy methods like HER.

* Table III and Figures

Figures 2, 3, 4, 5, 6

Table III

CODE: