Adaptive Online Planning for Continual Lifelong Learning

Kevin Lu[1], Igor Mordatch[2], Pieter Abbeel[1]

[1] UC Berkeley, [2] Google Brain

Abstract

We study an online continual lifelong learning scenario where the underlying dynamics of the world can change and the agent has no ability to reset, meaning that mistakes compound catastrophically into the future.

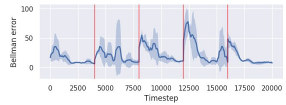
Model-based planning is effective immediately given the model, but is prohibitively expensive and can be biased by the planning horizon. Model-free learning is slow, leading to crucial mistakes that can land the agent in an area where learning is difficult, and can struggle to adapt to world changes.

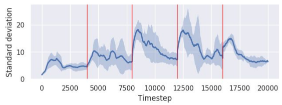
What do we need to learn effectively in this setting with limited computation?

Method

AOP augments MPC with model-free learning of value and policy networks. Model-based planning allows for effective control early in life, and model-free learning allows for reduced computation and stable behavior.

We quantify uncertainty as the disagreement of an ensemble of value functions. At the beginning of training, standard deviation is high. At world changes, the value function weakly predicts the value of the current state (Bellman error). When below a threshold, AOP uses a lower planning horizon.





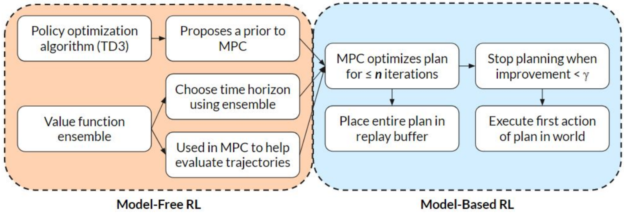
AOP executes more planning iterations when planning iterations improve the planned trajectory, using the value function to guide planning and exploration. The policy is preferred when it is strong compared to the planned trajectory.

Planning

AOP uses significantly less planning than comparable algorithms over a lifetime.

AOP-TD3	AOP-BC	POLO	MPC-8	MPC-3
11.39%	11.40%	37.50%	100%	37.50%

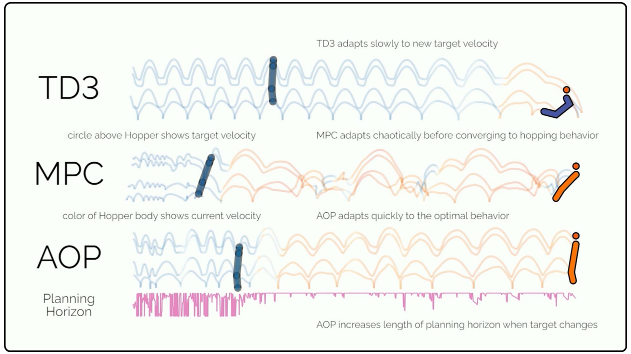
Adaptive Online Planning



Fast Adaptation

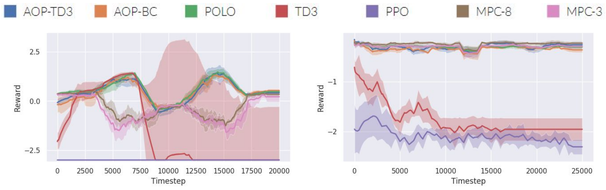
The hopper tries to achieve a target forward velocity. The blue outline of the hopper shows when the agent is tasked with switching from a slow (blue) to a fast (orange) target velocity.

TD3 adapts slowly and falls over, while MPC acts chaotically, leading to suboptimal behavior. AOP quickly adapts to the new world setting, increasing its level of planning accordingly to do so.



Changing Worlds

AOP achieves the performance of expensive model-based planners with ~11% of their level of planning, avoiding the performance degradation over time of model-free methods.

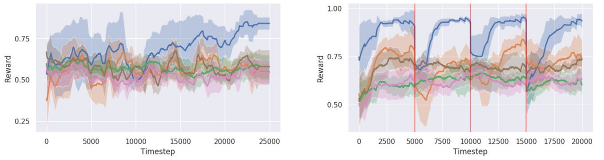


Hopper (C)

Maze (C) [D]

Novel States

Furthermore, AOP demonstrates learning; model-based planning only stays static in performance. Left: AOP learns to generalize to unseen worlds. Right: AOP performs better in mazes it sees many times (red lines show new worlds).



Maze (C) [S]

Maze (N) [S]

More Information

Website: bit.ly/aop_neurips

arXiv: arxiv.org/abs/1912.01188

Code: github.com/kzl/aop