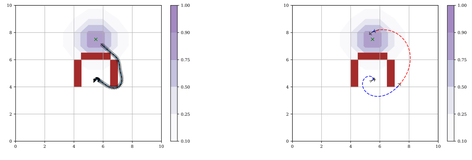
Improving planning and MBRL with temporally-extended actions

Palash Chatterjee, Roni Khardon
Indiana University, Bloomington

I. Summary

From too many decisions to a couple of decisions.



Motivation : Discrete time dynamics with small timescale → long planning horizon → computationally intensive.

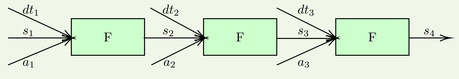
Idea : Treat the duration of the action as an additional optimization variable and optimize alongside other action variables.

Advantages : Shorter planning horizon but deeper search. Faster planning and learning.

II. Key Ideas

Planning with TE dynamics model

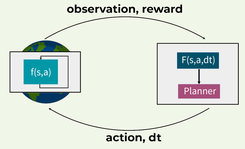
Optimize for action + action duration. Rollout with TE dynamics model.



1. Shorter planning horizons → reduced search space. ($2^{H/m}$ instead of 2^H)
2. Less variables to optimize. $(H/m)(|\mathcal{A}|+1)$ instead of $H|\mathcal{A}|$.
3. In environments with uninformative rewards, can search deeper by simply scaling m. No increase in computation cost.

Obtaining TE dynamics model

MBRL with proposed planner and learned TE dynamics model.



1. Shorter horizon → smaller compounding errors.
2. Less decisions + constant-time evaluation leading to faster evaluation.

Dynamic dt range selection

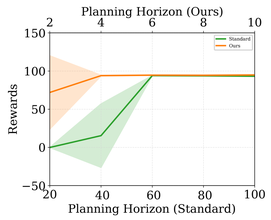
Non-stationary MAB with a UCB-heuristic with EMA of rewards.

$$\arg\max_i \left(\hat{R}_{i,T} + c\sqrt{\frac{2\log T}{N(i,T)}} \right)$$

Here $\hat{R}_{i,T}$ = EMA of rewards of arm i and, $N(i,T)$ = number of times arm i was pulled till time T.

Why non-stationary? Agent's performance depends on model quality which changes after every training iteration.

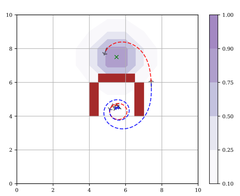
III. Experiments



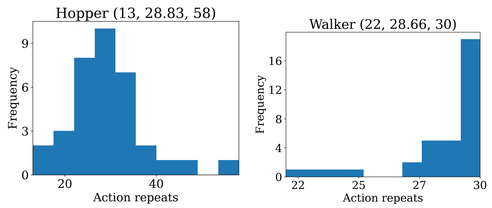
	Rewards ↑		Success Probability ↑	
	Standard	Ours	Standard	Ours
1	90.98 ± 0.26	91.74 ± 1.55	1.00	1.00
2	-0.1 ± 0.01	**88.67 ± 0.65**	0.00	**1.00**
3	-0.1 ± 0.01	**86.15 ± 1.11**	0.00	**1.00**
4	-0.1 ± 0.01	**66.50 ± 43.54**	0.00	**0.80**
5	-0.1 ± 0.01	**83.41 ± 0.54**	0.00	**1.00**

(L) We solve Mountain Car with a shorter planning horizon and (R) we solve more instances of Mountain Car from International Planning Competition.

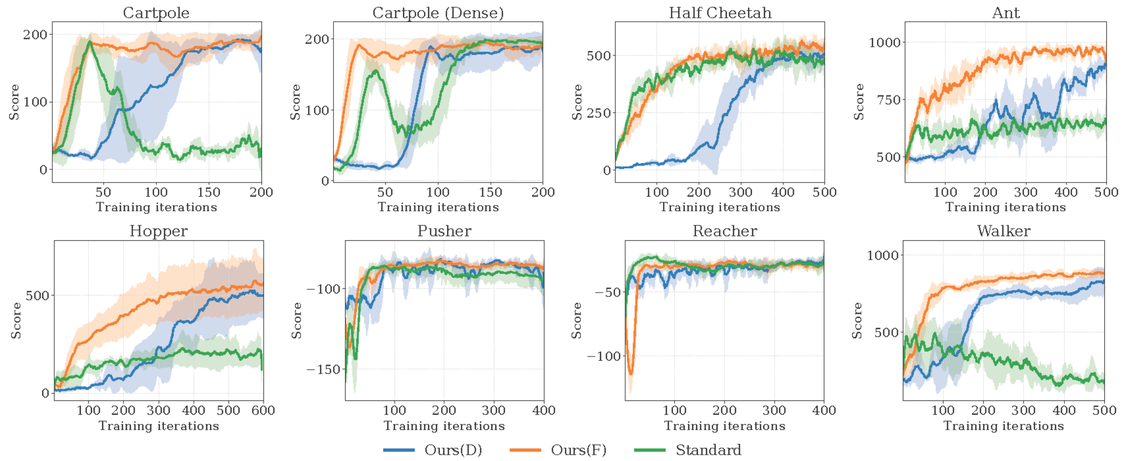
Out of memory error

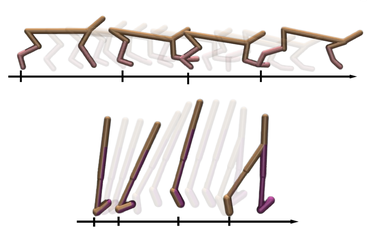


Planning in Dubins Car with $|\mathcal{A}| = 102$. Our method (R) can solve problems where the standard planner (L) requires huge memory.



Distribution of (discretized) action duration during an episode. Agent makes use of the extra flexibility and chooses actions with varying durations.





Training time (in hours)

Env	Standard	Ours (F)	Ours (D)
Cartpole	0.65	**0.4**	**0.5**
Half Cheetah	45	**5**	**5.5**
Ant	41	**4**	**7**
Hopper	40	**4**	**6.5**
Reacher	2.6	**1.6**	**1.5**
Pusher	2.6	**2**	**1.5**
Walker	36.8	**3.1**	**5.8**

Top, Right: Using TE actions leads to better and faster MBRL performance.
Left: Illustration of TE actions in Half Cheetah and Walker.