Improving planning and MBRL with temporally-extended actions

Palash Chatterjee, Roni Khardon

Indiana University, Bloomington

pecey.github.io/MBRL-with-TEA

Motivation

Real world

Continuous time dynamics. Evolves continuously with time.

Approximate model

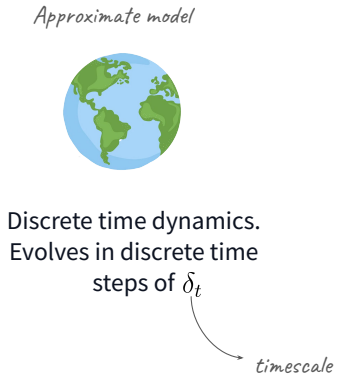
Discrete time dynamics. Evolves in discrete time steps of δ_t

timescale

Motivation

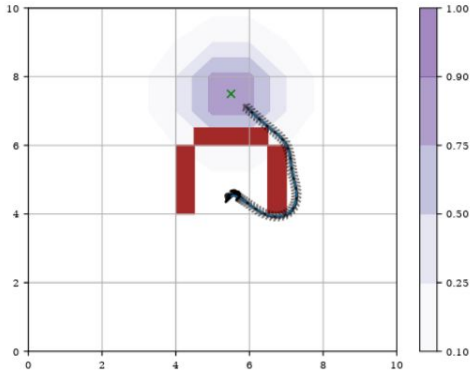
Small timescale → Good local approximation.

But agent **needs larger number of decisions** to solve problems, thus **increasing the complexity**.

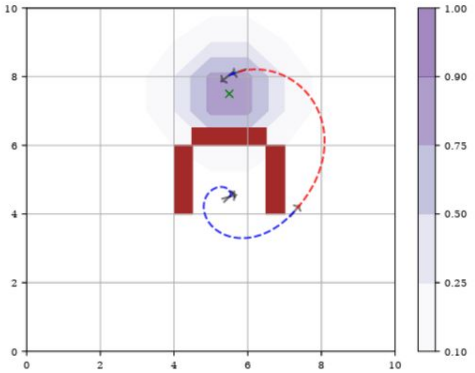


Motivation

Primitive actions



Can we do this?

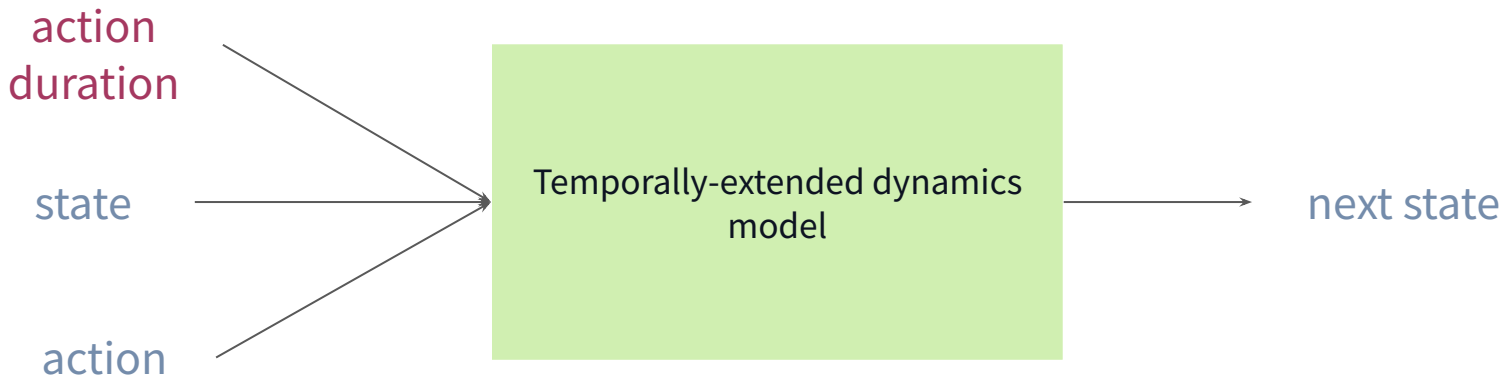


$$|\mathcal{A}| = 2$$

Our Contribution

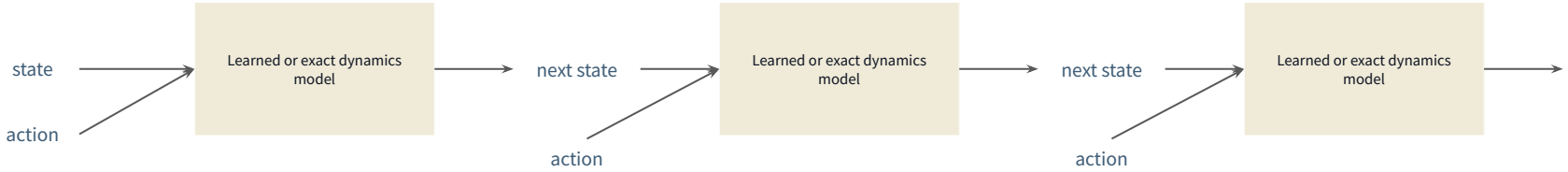
Let the planner **optimize for the action as well as the action duration**, leading to **improved planning and learning** performance.

Wishful Thinking



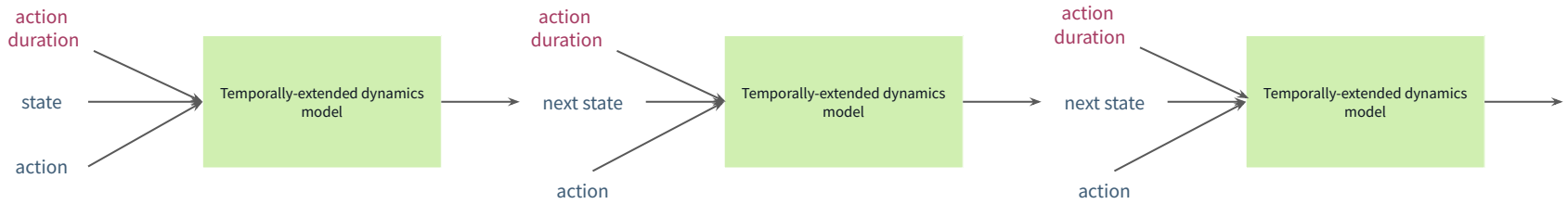
How do we plan if we had access to a temporally-extended dynamics model?

Standard Planning Framework



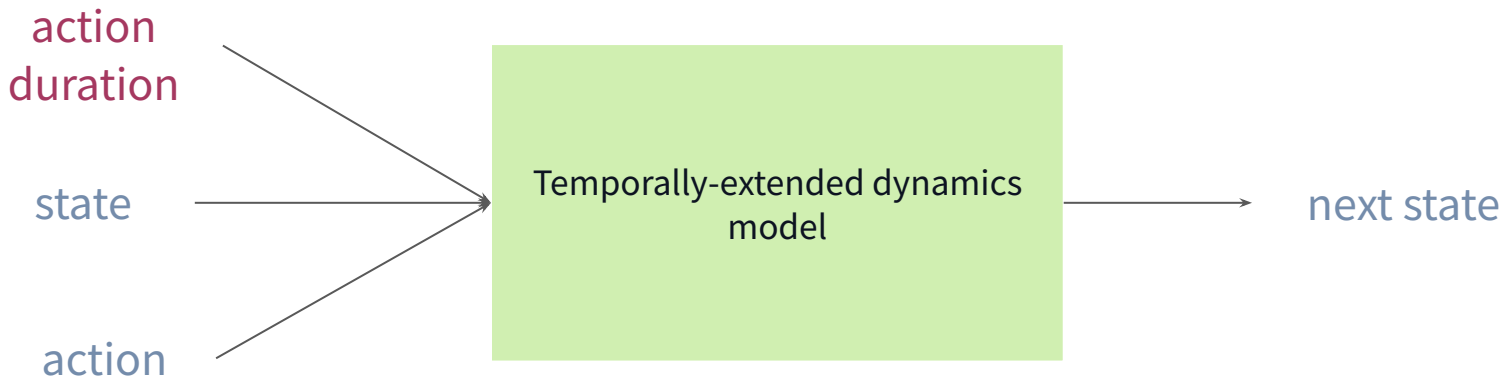
Need to evaluate at fixed time intervals in order to obtain a trajectory.
Can lead to very long planning horizons.

Planning with Temporally-Extended Dynamics Function



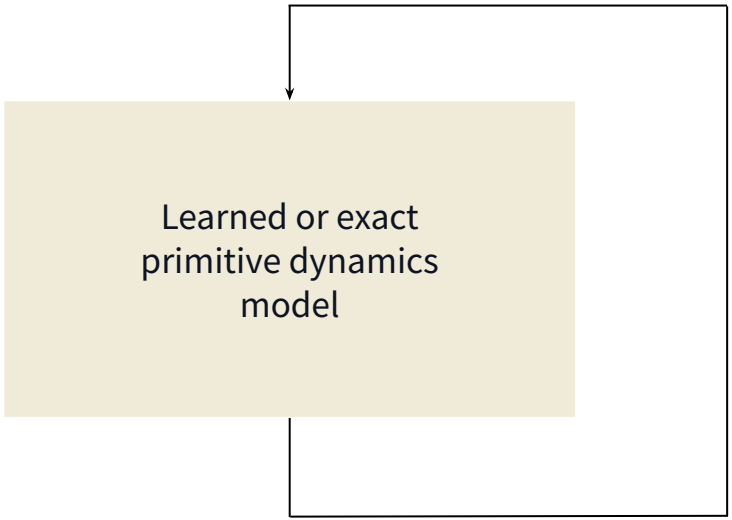
Plan as usual, but with shorter planning horizons.

The Big Question



How do we get a temporally-extended dynamics model?

Key Idea 2



Gets the work done for exact model.

Can lead to compounding error in case of learned models.

Too slow.

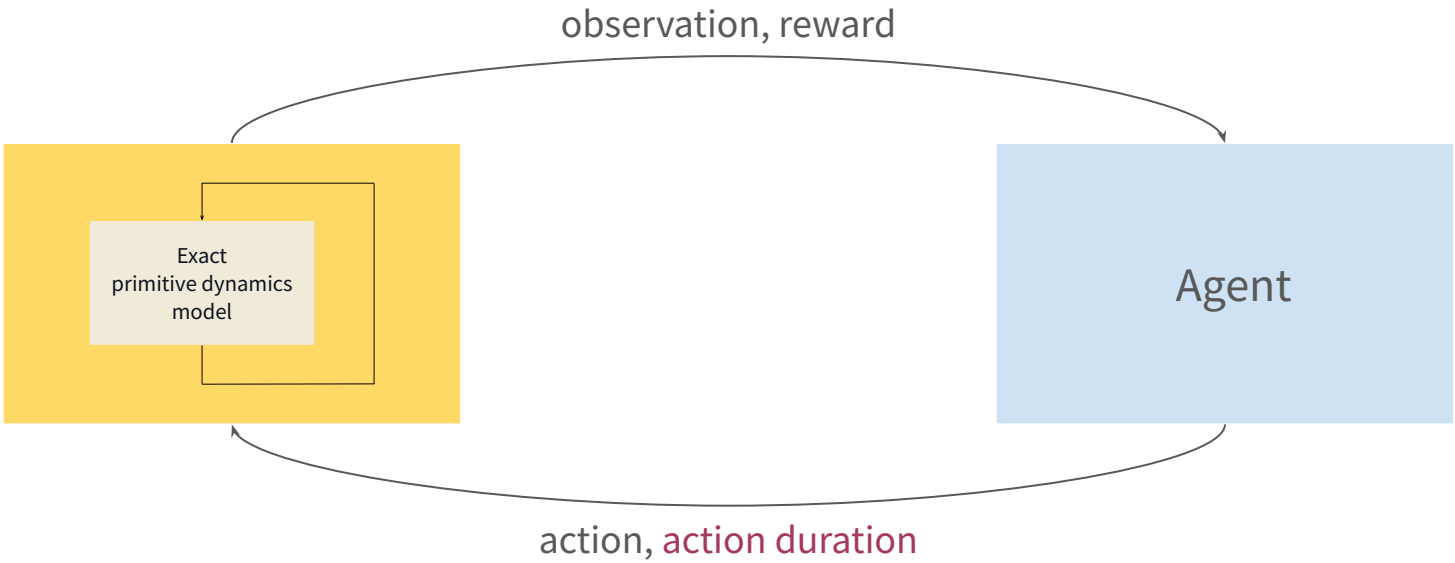
Key Idea 2

Learned temporally-extended dynamics model

Gets the work done.

Constant time evaluation for a temporally-extended action.

Learning in practice



observation and reward are due to an temporally-extended action.

Range for action duration

What range should the planner search over?

✗ **Set up as a hyper-parameter.**

✓ **Dynamic selection.**

Dynamic selection of range for action duration

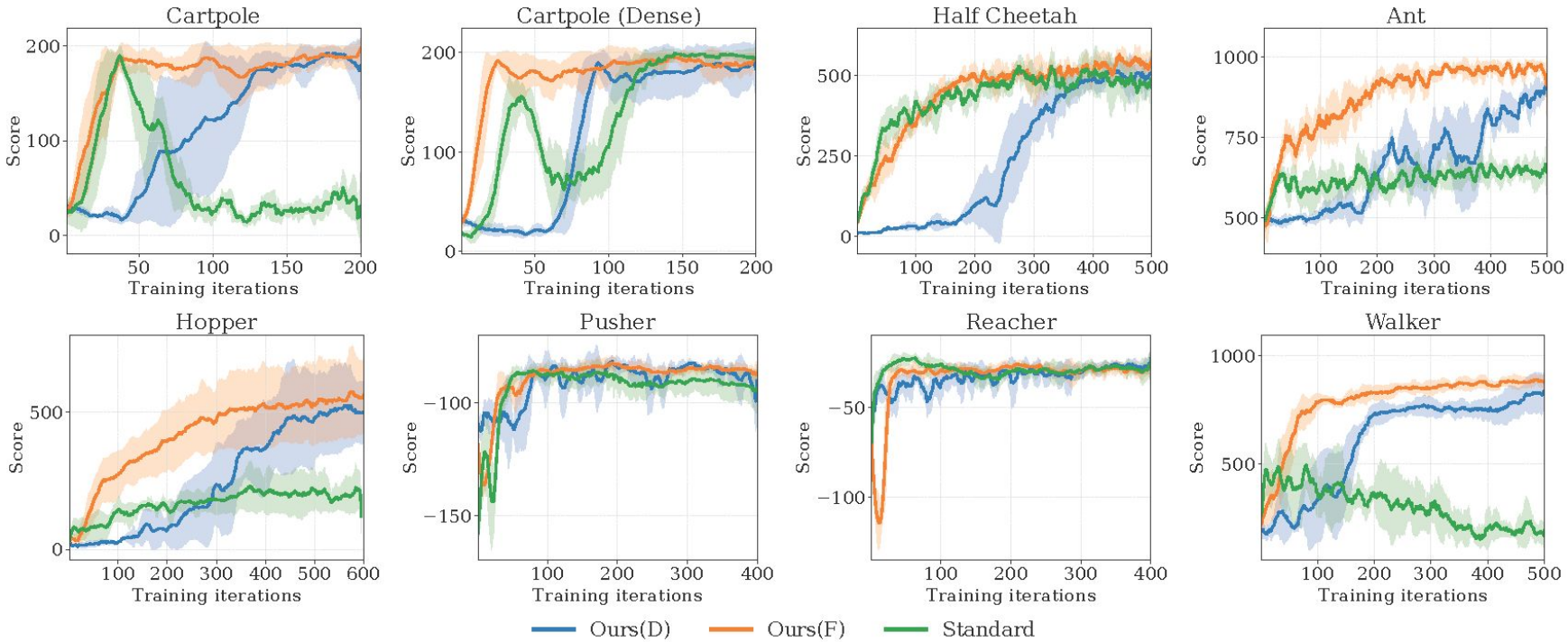
Treat as a **non-stationary multi-armed bandit** problem with **exponentially spaced candidates** and a **UCB-heuristic with exponential-moving average** of rewards for arm selection.

Results - Planning

Solves problems with shorter planning horizon and which are not solvable with standard planning.

Please see the paper for detailed results.

Results - MBRL with temporally-extended actions



Competitive in all environments, significantly better in some.

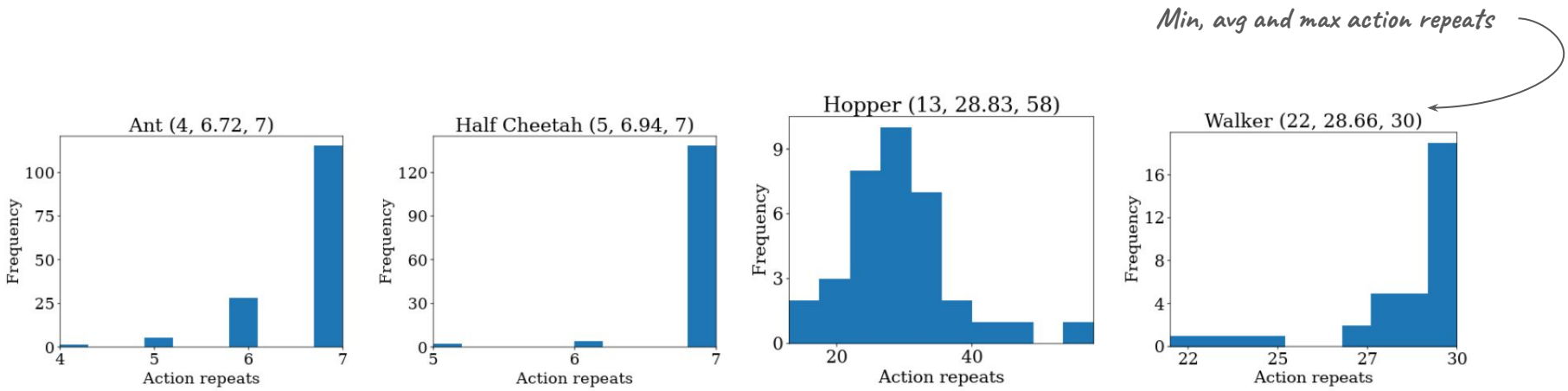
Results - MBRL with temporally-extended actions

Approximate 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**

Significant speedup in most environments.

Results - Comparison to action repeats



Distribution of discretized action duration is concentrated in some environments, while being dispersed in others.

Conclusion

TE actions → action duration as pseudo action variable.

Planning → TE actions + iterate over exact dynamics model.

MBRL → TE actions + learned TE dynamics + non-stationary MAB.

Please see the paper for more details and results.

Thank you.