Temporal Difference Learning for Model Predictive Control

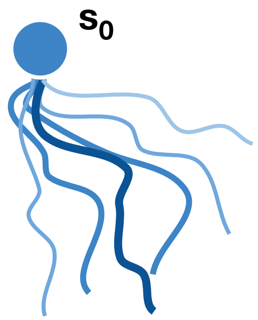
Nicklas Hansen, Xiaolong Wang*, Hao Su*

ICML 2022

UC San Diego

Data-Driven Model Predictive Control (*MPC*)

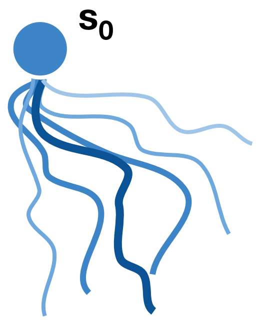
- Plan using a ***learned*** model of the environment
- Objective $\mathbb{E}_{\Gamma \sim \Pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right]$ intractable



(repeat for ∞ steps)

Data-Driven Model Predictive Control (*MPC*)

- Plan using a ***learned*** model of the environment
- Objective $\mathbb{E}_{\Gamma \sim \Pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t r(\mathbf{s}_t, \mathbf{a}_t)\right]$ intractable
- Instead find ***locally optimal*** trajectory $\mathbb{E}_{\Gamma \sim \Pi_\theta}\left[\sum_{t=0}^{H} \gamma^t r\left(\mathbf{s}_t, \mathbf{a}_t\right)\right]$
- **Two major challenges:**
  - Compounding model errors
  - Cost of long-horizon planning



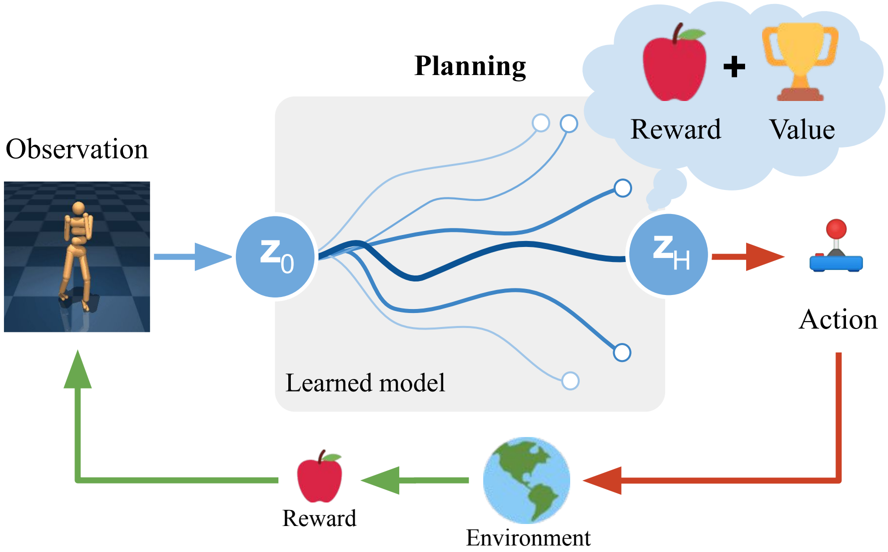
(repeat for H steps)

How can TD-learning help MPC?

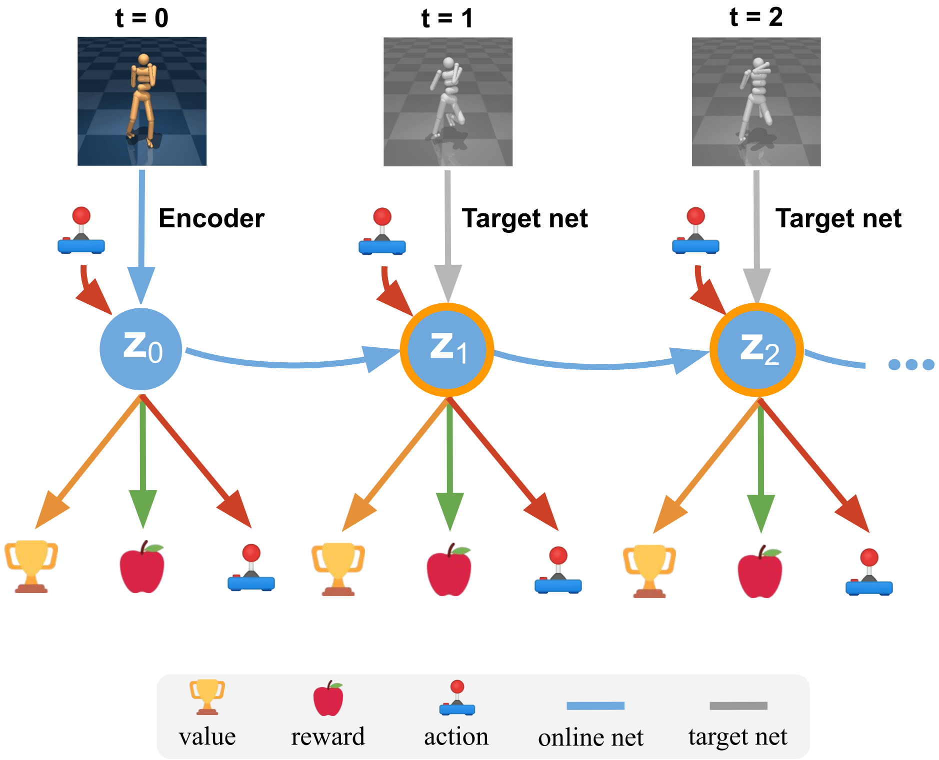
Inference *(planning)*

- Planning in latent space
- Return estimate:

$$\mathbb{E}_{\Gamma}\left[\underbrace{\gamma^{H} Q_{\theta}(\mathbf{z}_{H}, \mathbf{a}_{H})}_{\textbf{Value}} + \underbrace{\sum_{t=0}^{H-1} \gamma^{t} R_{\theta}(\mathbf{z}_{t}, \mathbf{a}_{t})}_{\textbf{Rewards}}\right]$$

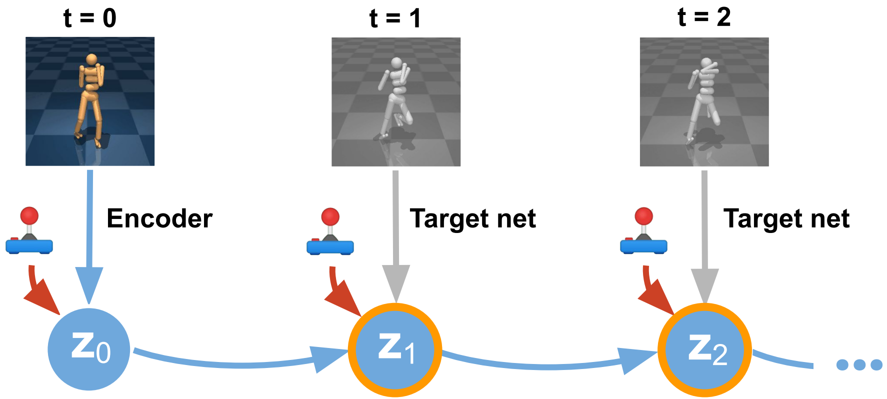


TD-MPC

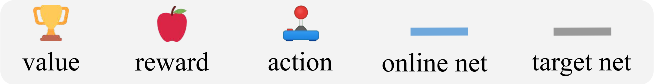


t = 0
t = 1
t = 2
Encoder
Target net
Target net
z_0
z_1
z_2
value
reward
action
online net
target net

TD-MPC



Minimize diff. between **recurrent prediction** and **target encoding**

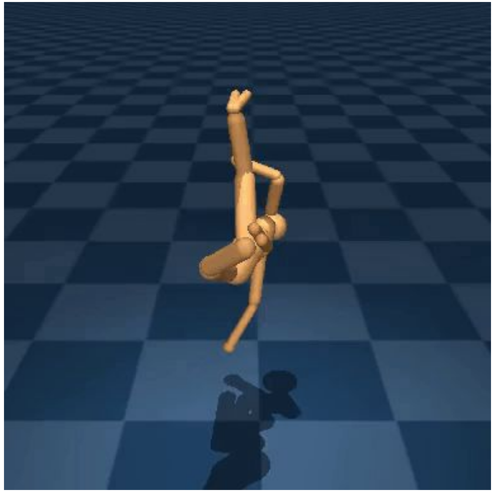


Results

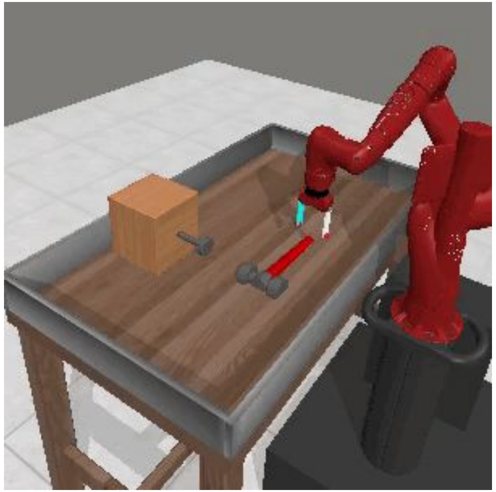
TD-MPC solves ***challenging*** continuous control problems

Dog Run

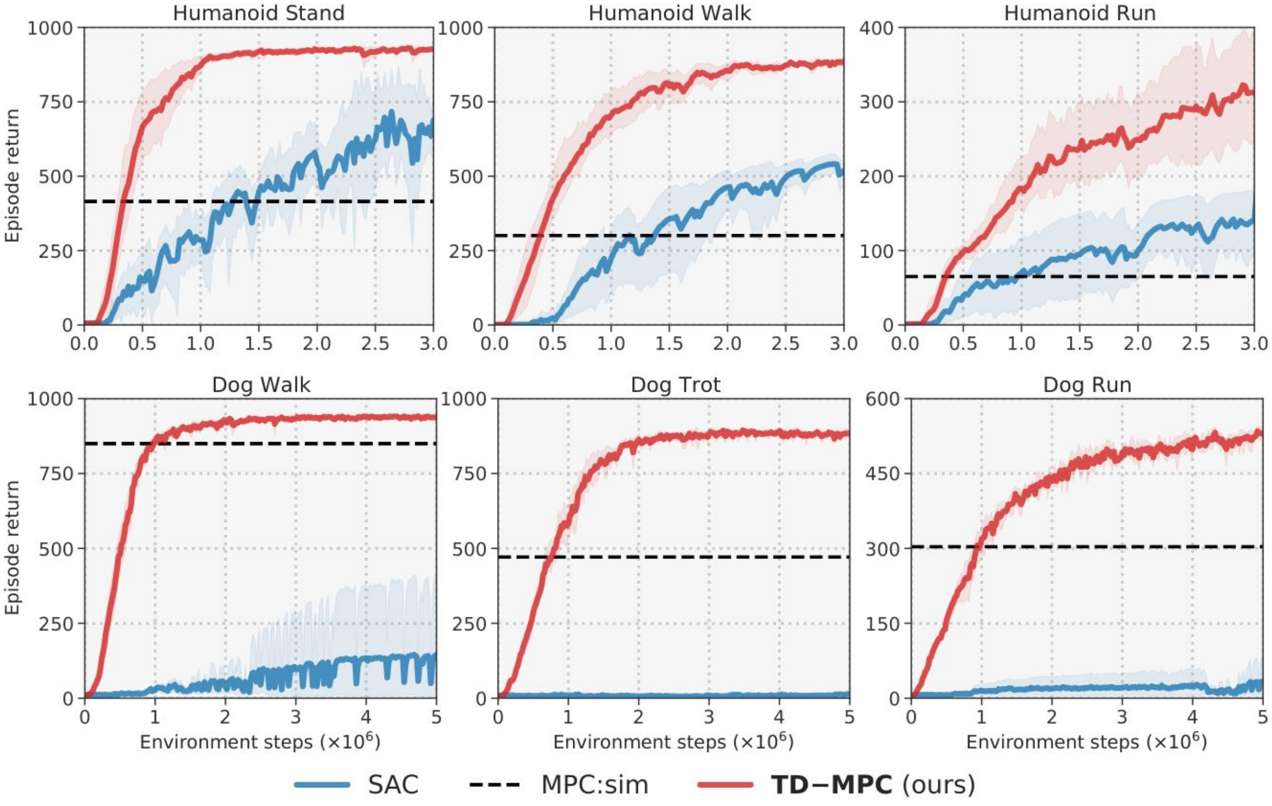
Humanoid Run



Hammer



Results



Humanoid Stand
Humanoid Walk
Humanoid Run
Dog Walk
Dog Trot
Dog Run
Episode return
Environment steps (×10⁶)
SAC
MPC:sim
TD−MPC (ours)

Poster: 6-8pm today

nicklashansen.github.io/td-mpc