Meeting
09/09/2020

Shuo Zhang

VI. RELATED WORK

[1]: MPC+MFRL; While MFRL's strength in exploration allows us to train a better forward dynamics model for MPC, MPC improves the performance of the MFRL policy by sampling-based planning; MBRL's data efficiency + MFRL's level of performance. But MPC planning only occurs in evaluation phase, thus actions generated by MPC planning has nothing to do with MFRL's policy update.

[2]: Model-based rollouts are used to compute targets of value function training, thus accelerating value function learning. In a MFRL(e.x. DDPG) algorithm, target-Q of S_0 is computed using both H steps dynamics model rollouts and target-Q of S_H. In TD learning, H only equals 1.

[3]: STEVE is similar to MVE and is shown better than MVE. STEVE is shown to be able to balance between the errors in the model and the Q-function estimates. Instead of a fixed H step, Value expansion of STEVE is computed as a weighted sum of M*N*L ensemble trajectories' mean of each of H steps, where M,N,L are numbers of different dynamic model parameters, reward function parameters and Q-value parameters. The weight is proportional to the inverse of variance as uncertainty measure.

[4]: iLQG is used as a model-based controller to do rollouts in dynamics model, from which the imaginary data is also incorporated into the training of model-free Q-learning.

REFERENCES

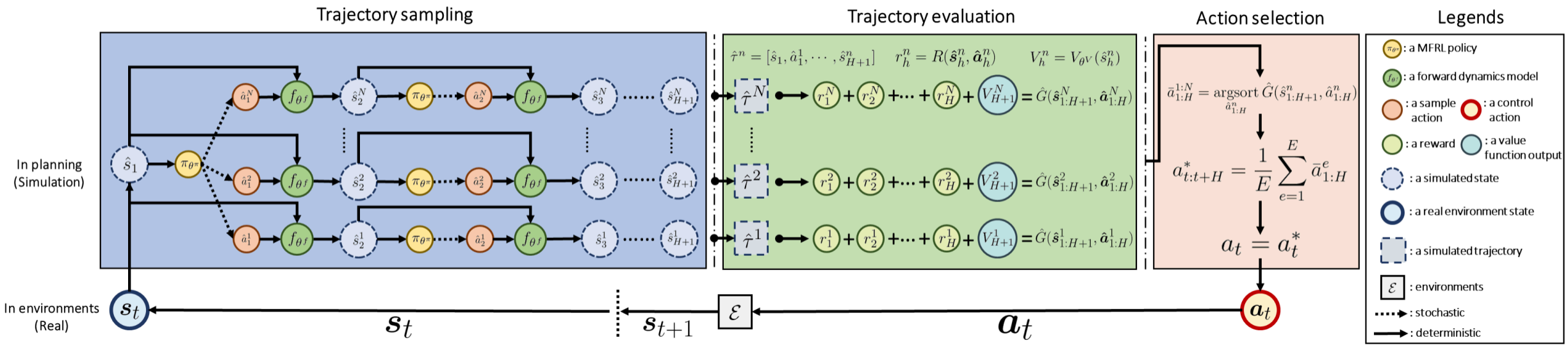
[1] Zhang-Wei Hong, Joni Pajarinen, and Jan Peters. Model-based lookahead reinforcement learning. *arXiv preprint arXiv:1908.06012*, 2019.

[2] Vladimir Feinberg, Alvin Wan, Ion Stoica, Michael I Jordan, Joseph E Gonzalez, and Sergey Levine. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*, 2018.

[3] Jacob Buckman, Danijar Hafner, George Tucker, Eugene Brevdo, and Honglak Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Advances in Neural Information Processing Systems*, pages 8224–8234, 2018.

[4] Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. In *International Conference on Machine Learning*, pages 2829–2838, 2016.

MPC-MFRL


Trajectory sampling
Trajectory evaluation
Action selection
Legends
In planning
(Simulation)
In environments
(Real)
$\hat{\tau}^n = [\hat{s}_1, \hat{a}_1^1, \cdots, \hat{s}_{H+1}^n]$
$r_h^n = R(\hat{\boldsymbol{s}}_h^n, \hat{\boldsymbol{a}}_h^n)$
$V_h^n = V_{\theta^V}(\hat{s}_h^n)$
$\bar{a}_{1:H}^{1:N} = \underset{\hat{a}_{1:H}^n}{\text{argsort}}\, \hat{G}(\hat{s}_{1:H+1}^n, \hat{a}_{1:H}^n)$
$a_{t:t+H}^* = \frac{1}{E}\sum_{e=1}^{E} \bar{a}_{1:H}^e$
$a_t = a_t^*$
$\boldsymbol{s}_t$
$\boldsymbol{s}_{t+1}$
$\mathcal{E}$
$\boldsymbol{a}_t$
: a MFRL policy
: a forward dynamics model
: a sample action
: a control action
: a reward
: a value function output
: a simulated state
: a real environment state
: a simulated trajectory
: environments
: stochastic
: deterministic

MPC-MFRL (new ideas?)

- So far, MPC does not affect trpo update directly. We can try to incorporate MPC into TRPO policy update as well? (MPC not in evaluation phase, but the generated action could guide policy network's update? (like supervised learning))
- i.e. Rollouts in dynamics models could not only be considered for the target of value function, but also for the 'target' of actions?)
- Also, so far contemporary methods try to incorporate rollouts in dynamics models for the target of value function, not for the guidance of actions.