Scope

This project may be instantiated as a Master's project or thesis. Prior knowledge in Reinforcement Learning and Deep Learning is recommended.

Problem setting

One of the most severe and fundamental issues arising in Reinforcement Learning (RL) from offline datasets with non-linear function approximation is the problem of value-overestimation [1]. In contrast to online learning methods, which can alleviate this issue by broader sampling around the assumed maximum, offline methods have no access to the underlying mechanism performing exploration. Current Offline Reinforcement Learning methods are therefore commonly based on the assumption that out-of-distribution actions (and states, for that matter) yield a worse return than the coverage of the dataset. [2] Optimization is hence only performed within the scope of the data. This "conservatism" on the flip side may introduce another form of bias, namely that the collected data is somewhat optimal (or at least covers enough information to infer an optimal policy). Depending on the quality of the dataset, this assumption can lead to unsatisfying performance.

Hypothesis

Most model-based RL methods approximate the dynamics of a given MDP only over one or few consecutive timesteps. In many cases, this simplifies dynamics estimation in that the considered horizon is commonly shorter than the full temporal horizon of a task [2]. In addition, the aspects covered by a learned dynamics model commonly differ (at least to some extent) from the long-term value function. We therefore assume that, in some or hopefully many cases, the generalization capabilities of dynamics model and value-function are different. And we further assume that this can be exploited in the offline setting.

Proposed Method

If we can derive uncertainties in value- and dynamics estimation, similar to [3,4,5], we can trigger exploratory rollouts within the learned model and limit the horizon proportionally to the balanced uncertainties. This then extends the dataset provided in an offline setting in order to reduce the degree of required conservatism.

Roadmap

1. Proof-of-Concept with the true model. Train a value-function on an offline dataset. Implement model-based rollouts on the true model and show that performance increases as a sanity check. Then infer an uncertainty-measure from the value-function and perform truncated rollouts. Again, verify that we can get an increase in performance.
2. Train value-function and dynamics model independently on a given task and evaluate uncertainties on the state and action spaces. Verify that there are areas apart the dataset where the uncertainties differ.
3. Balance uncertainties in model and value-function and use that to perform truncated rollouts. Evaluate on a range of continuous control tasks.

References

[1] Stabilizing off-policy q-learning via bootstrapping error reduction. Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. (NeurIPS 2019)

[2] Offline reinforcement learning: Tutorial, review, and perspectives on open problems. Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. (preprint, 2020)

[3] Uncertainty weighted actor-critic for offline reinforcement learning. Yue Wu, Shuangfei Zhai, Nitish Srivastava, Joshua M. Susskind, Jian Zhang, Ruslan Salakhutdinov, and Hanlin Goh. (ICML 2021)

[4] Why so pessimistic? estimating uncertainties for offline RL through ensembles, and why their independence matters. Seyed Kamyar Seyed Ghasemipour, Shixiang Shane Gu, and Ofir Nachum. (preprint 2022)

[5] Pessimistic bootstrapping for uncertainty-driven offline reinforcement learning. Chenjia Bai, Lingxiao Wang, Zhuoran Yang,

Zhihong Deng, Animesh Garg, Peng Liu, and Zhaoran Wang. (ICLR 2022)

Organization

Write an email to nrprojects@informatik.uni-freiburg.de with a clear reference to this project proposal.