## Review-3

### When to use parametric models in reinforcement learning?

Growing advances in model-based reinforcement learning have yielded data-efficient methods. These methods learn a model of the dynamics of the world and align behaviors of the agent with the model's beliefs. However, it is often unclear as to when to use a model for acting and planning since models may inherently be imperfect in nature and can steer the agent towards a sub-optimal policy in the case of inaccurate beliefs. The work investigates the usage of parameterized models and their relationship to replay memory in various reinforcement learning settings. Replay memory is thought of as analogous to a model being used for behavioral updates. Such behavioral models are used for training agents in a sample-efficient manner in comparison to pure model-based approaches.

Parameterized models serve as an internal component for the agent in order to plan for future steps. However, planning is hindered by inaccurate beliefs in the long-horizon and computational expense since the model is updated to learn complete dynamics of the environment. Pure model-based approaches collect data from the environment and train the agent based on the model being updated by the agnet's policy. Another method for using models is by executing inverse updates which predicts the dynamics backwards. Learning an inverse model is beneficial since it only provides fictional states and does not harm agent's behaviors. On the other hand, a replay-based approach collects data from the environment and simply updates the agent's policy based on these samples. A stark contrast in learning with these methods can be observed on the basis of (1) scalability and (2) performance. Firsly, the conventional planning approach is scalable in the number of planning steps and reduces data dependency as the planning horizon is improved. Secondly, inverse model learning is comparable replay in terms of performance under high stochasticity. This indicates that a pure model-based approach is not robust to changes in environment dynamics. The claim is further strengthened by carrying out a large-scale study of Atari 2600 games with the novel data-efficient Rainbow which uses fewer samples to demonstrate better performance in comparison to state-of-the-art Simulated Policy Learning (SimPLe) method.

Insights related to the usage of an inverse model and the failure to learn a forward model in the presence of the deadly triad are significant contributions introduced by the work. Additionally, the data efficient Rainbow accurately depicts the sufficiency and benefits of a replay-based approach in the case of large-scale learning. However, the work falls short of addressing two major points in model-based learning. Firstly, while the experiments highlight replay-based learning as an alternative to learning models, they fall short of explaining when and how the two schemes may be combined for optimal learning. For instance, one may conjecture that collecting environment samples in a replay and using them to update the model would improve data-efficiency while the work does not throw any light on this insight. Secondly, the imrpoved data-efficient Rainbow is compared to SimPLe which is the only model-based method for Atari 2600 games. A better basis of comparison could be wherein the replay memory of Rainbow is replaced by a model. This way, performance of the same Rainbow agent could be compared between the two schemes.

Introduction of the modified Rainbow is one of the novel contributions of the work. The data-efficient Rainbow significanlty outperforms the canonical implementation and yields significant performance improvements in comparison to SimPLe. The improved additionally utilizes fewer samples to converge towards better reward metrics when compared to SimPLe. Apart from its empirical analysis, the work answers the broader question on the usage of parameteric models by theoretically demonstrating a failure to learn in the case of the deadly triad setting. These claims pave way for new questions related to the combination of models with replay-based approaches. Moreover, the work can be further generalized to the off-policy case and for different choices of function approximators in different settings such as continuous control.