Author response for 'When to use parametric models in reinforcement learning?'

We sincerely appreciate the reviewers' time and effort to provide useful and insightful reviews, and in this case of course especially for their helpful comments and question about our paper.

On code and reproducibility

We took care that our experiments are reproducible, and we will recheck this carefully on acceptance. We are happy to report that our main experiment has already been successfully reproduced by others, without access to our code, which is a good validation that the paper contained sufficient details. We of course appreciate that releasing code can help speed up research (and can sometimes help clarify important details), and intend to release accompanying code, and we also think it is important to make sure the paper itself contains sufficient detail to fully reproduce the results.

On model benefits

One reviewer raises interesting and insightful points about the usefulness of models that generalise better than a policy or value. We agree that this is an important aspect that is worthwhile to discuss in some detail.

An important, and perhaps obvious, conclusion is: it is not just important how accurate the model is (though this clearly matters), but also the model is used. We agree with the reviewer that models that generalise better than a policy or value seem especially interesting, when these are attainable. Interestingly if we then use the model in the same way as we could use replay (as opposed to, for instance, using it to plan forward for behaviour), then only accuracy does not always suffice, even in such benign settings.

It might be interesting to explain an experiment that we conducted, but that did not make it into the paper (yet). We set up an experiment in which the true transition dynamics were quadratic, and the model was also chosen to be quadratic. Non-surprisingly, the model very quickly learnt to match the true dynamics, essentially perfectly. To our initial surprise, the forward Dyna algorithm didn't perform well even in this presumably benign setting—it performed far worse than the replay-based algorithm and the parameters of the value function would often diverge. We first suspected a bug but careful examination revealed this was, instead, a failure of the sort that is now discussed in Section 3. In the appendix we chose a simpler (two-state) example to concretely illustrate the more general theory around this failure, but perhaps it is useful to include this quadratic example in the paper as well, as another demonstration that even a model that generalises perfectly can fail if not used with care.

To be clear, of course we agree that a perfect model, if available, can help attain performance that should surpass that attained from using replay instead. We are just pointing out that using the model only to generate fictional data in the same way (and from the same states) as replay would may not be the easiest way to benefit from an imprecise model.

Of course, it can be hard to know a priori when a model will be accurate enough to rely on, especially for longer trajectories where compounding model errors can be a problem [cf., e.g., Talvitie, 2014, 2017, Asadi et al., 2019]. To quote Vladimir Vapnik: *one should (...) never solve a more general problem as an intermediate step* [Vapnik, 1998, Section 0.9]. Any statement of such generality comes with caveats, but it is interesting to consider in this context: when we use replay we at least know that the data is real, and we do not have to question its accuracy.

We agree that it is important to state our findings as clearly as possible (and appropriately scoped, to avoid hinting toward unwarranted overly general conclusions) and we intend to carefully keep polishing the writing to make the paper as clear as possible.

Relatedly we are also considering including further empirical results (including, though not exclusively, more at-scale results, e.g., with backward planning on Atari, as well as perhaps including the experiment with quadratic dynamics described above) to help further elucidate our main points and augment the experiments currently in the paper.

References

K. Asadi, D. Misra, S. Kim, and M. L. Littman. Combating the compounding-error problem with a multi-step model. *CoRR*, abs/1905.13320, 2019.

E. Talvitie. Model regularization for stable sample rollouts. In *UAI*, pages 780–789, 2014.

E. Talvitie. Self-correcting models for model-based reinforcement learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

V. Vapnik. *Statistical learning theory*. Wiley, 1998.