Generalisation Ability of Proper Value Equivalence Models in Model-Based Reinforcement Learning

Author: Severin Bratus *(s.bratus@student.tudelft.nl)*
Supervised by: Jinke He, Frans Oliehoek

Background

In ***model-based reinforcement learning***, the agent attempts to build a representation of the dynamics of the environment it interacts with in order to apply this model to find a better policy.

In other words, the agent ***plans*** with a model. A model may be used to *evaluate* policies – that is, to predict their true *value* in the environment.

Typically, the model is learned in a supervised manner from sample transitions of environment interactions to predict the next state, after taking some action from some previous state.

This is done with a loss that does not take the future use of the model into account, for instance the **maximum likelihood estimate (MLE)** loss, which results in an *objective mismatch* problem – a well-fit MLE model may not perform best in planning.

In 2021, C. Grimm et al. [2] have introduced the notion of **proper value equivalence** – which states that a model is proper value-equivalent to the environment with respect to some set of policies, if it can be used to compute the value functions of the policies as accurately as in the environment.

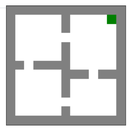
PVE models have been shown to outperform MLE models in planning, however we have supposed that in some settings MLE models would generalise better in policy evaluation.

Research Question

How do predictive models based on MLE and PVE loss objectives compare in of unseen policies?

Methodology

To address the research question, we have compared the **test set value prediction error** of MLE- and PVE-based models in a simple maze-like environment *(shown on the left)*

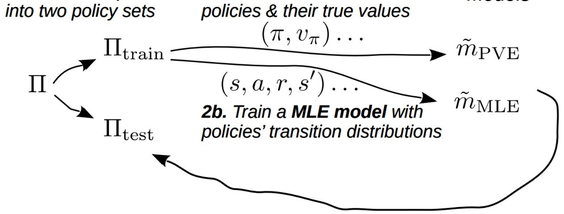


1. *Generate random policies as data, and split into two policy sets*

2a. *Train a* ***PVE model*** *with policies & their true values*

2b. *Train a* ***MLE model*** *with policies' transition distributions*

Models

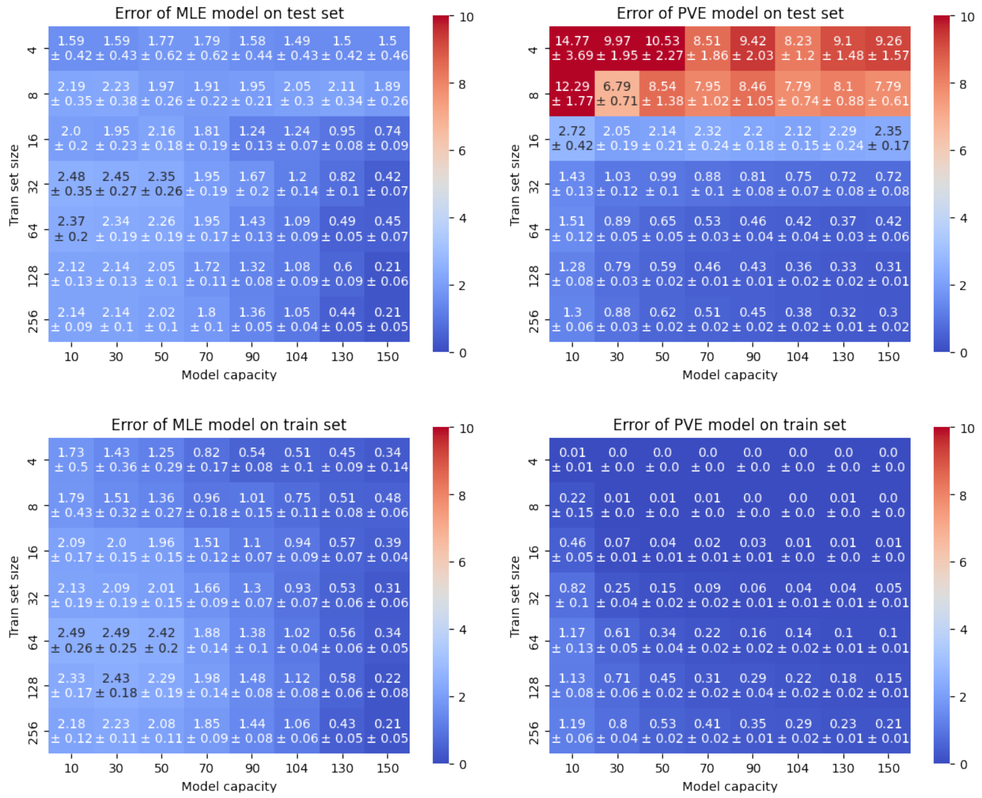


3. *Measure models' value prediction error on the test policy set*

Results

In the Figure below you can see how MLE & PVE compare in varied settings for:
- the training policy set size *(y-axis, top-to-bottom)*
- model representational capacity *(x-axis, left-to-right)*

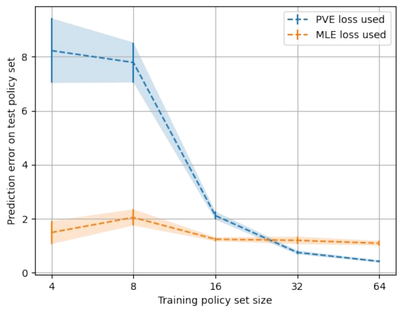
High error is shown in red, low error in blue.



^ PVE models seem to overfit on their training set of policies, as the train set error is much lower than the test set error. This effect is mitigated as the training set grows.

In the Figure to the left is the effect of *training policy set size* on *test set value prediction* error, for one fixed setting of model capacity.

PVE models' test error decreases significantly as the training size is increased, while MLE models' test error stays on the same level.



Conclusions & Future Work

We have demonstrated that, at least in some simple settings, there exists a boundary that divides contexts where MLE is preferable to PVE in policy evaluation generalisation ability, and vice versa.

For potential future work, we recommend exploring the behaviour of this boundary, theoretically, in formal terms, and experimentally, in a more realistic setting.

Limitations

- The experiments were conducted in a simple *tabular* grid-world setting, without function approximation -- results may be different in a more complex setting.
- We have decided not to take statistical error into account, and use expected forms of the loss objectives, instead of the empirical ones.

References

[1] Grimm, Christopher, et al. "The value equivalence principle for model-based reinforcement learning." Advances in Neural Information Processing Systems 33 (2020): 5541-5552.

[2] Grimm, Christopher, et al. "Proper value equivalence." Advances in Neural Information Processing Systems 34 (2021): 7773-7786.

[3] He, Jinke, Thomas M. Moerland, and Frans A. Oliehoek. "What model does MuZero learn?." arXiv preprint arXiv:2306.00840 (2023).