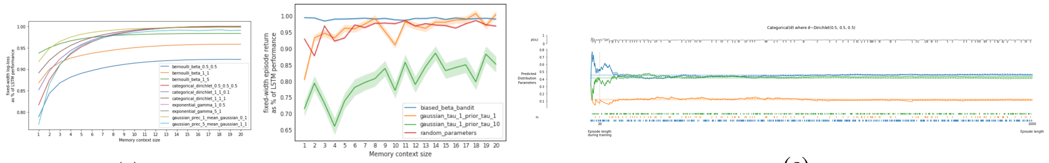
Overall comments We thank the reviewers for their detailed and helpful comments. We followed the reviewers' suggestions and have significantly expanded our discussion of related work into a separate section, incorporating the references mentioned (R1, R3). As suggested by R1 and R4 we have changed our writing to a more neutral tone, and mention the claims flagged by R1 as directions to study more rigorously based on our findings, with appropriate caveats regarding convergence and inclusion of the optimal policy in the search space. We also performed two experiments investigating memory-limited agents (including no memory) as suggested by R3, and generalization (R2, R3).



(a)

(b)

(c)

Figure 1: **a, b**: Effect of reducing memory on prediction and bandits respectively **c**: Extended generalisation.

Reduced-memory baselines (R3) As suggested by R3, we added an additional baseline: an agent with reduced memory capacity. This agent was implemented using a feedforward network given the n most recent timesteps. Results are shown in Figure 1A, B. These results also provide some information on outcomes when the Bayes-optimal policy is not in the set of policies - an interesting question suggested by the comments of R1 and R2.

Further relevant work (R1, R3) R1 mentions a number of papers on connections between RNNs and Bayes filtering, and highlights two papers which define new neural architectures for Bayes filtering. Unlike these papers, our manuscript focuses on the way in which Bayes filtering naturally arises from metalearning. This point—although clear once spelled out formally—is not widely appreciated in the metalearning community, as R1 notes (apart from the paper by Grant et al.). We also appreciate the pointers to Rosenschein and Kaelbling's work on situated/embedded automata. R3 highlights two related approaches to memory-based metalearning using fast and slow weights and a differentiable memory which we now include in an expanded related work section, as well as a section on interpretability.

Clarification of simulation results (R3) R3 refers to the expressivity of the map between the latent states of the Bayes-optimal and RNN agents as a possible weakness in our analysis. This map is regularised by a PCA of the RNN agent's latent state before the MLP. Importantly, a failure of simulation is a failure of injectivity: if a single state in one agent must be mapped to two distinct states in another then simulation fails, and no amount of expressivity in the map between states will save it. This occurs when two trajectories lead to the same state in one agent but not another (for instance if exchangeability has not been fully learned). We have clarified this in our discussion.

Is Bayes-optimality a surprising result? (R1, R2) R1 and R2 point out that our main findings are not completely surprising. While the behavioral correspondence is expected, it is theoretically unclear how (fully trained) RNN meta-learners internally organize sufficient stats. The fact that our particular structural analysis method is successful suggests a smooth and fairly structured organization of internal states in a (quasi-?) Euclidean space (interestingly similar organizational patterns have also been found in biological neural networks: doi 10.1101/461129). This, in turn, might also play an interesting role in a theoretical understanding of generalization (see next point).

Limitations of claims and Bayes-optimality in the generalization regime (R2, R1, R3) Our claims regard trained RNN meta-learners "on-distribution". We are currently not aware of theoretical predictions for generalization in meta-learners. We report positive results for evaluation on extended episodes in Fig. 1 in the paper, and we have also run experiments testing generalisation substantially beyond the 20 timestep horizon during training, by evaluating up to 1,000 timesteps (Figure 1C above, not cherry-picked). Meta-learned solutions generalise surprisingly well, suggesting that a generalisable update rule has been learned, but the mechanisms underlying this are not understood empirically or theoretically. In the absence of an adequate explanation, we would prefer to leave these mechanisms to further study.

Other comments: R1: Sample-inefficiency of meta-learning as a regression problem. Thanks for highlighting the connection to MC RL and the discussion in AIMA. We agree with the core issue and note that sample-efficiency was not a primary concern in our study - though it is a very active area of research: arXiv 2006.16507 and 2006.05094. Although Figures 11 and 12 show convergence 1-3 orders of magnitude before training stops, this is still sample-inefficient.

R2: comparison at the "computational level" —we did intend this in the Marrian sense, and now cite appropriately.

R2: Alternative hypotheses. We agree that these are interesting points to follow up on. However, if meta-training converges properly (regardless of problem complexity) there is no behavioral deviation from Bayes-optimality and the theory says little about the computational structure. Before convergence or off meta-distribution, a theoretical understanding is currently lacking.

R4: using the Bayes-optimal agent to generate inputs leads to small numerical changes, but no change to the conclusions.

R4: SGD baseline. We assume this refers to learning by SGD on a single sample from the environment prior. We were unable to implement this in time for the rebuttals, but do not expect that such a solution would generalise.