We thank reviewers R1, R2, R3, and R4 for their constructive and helpful feedback.

Scope and significance (R4). Our work derives the normative solution to the problem of how to dynamically allocate noisy and limited memory resources for reinforcement learning. This work could have implications for machine learning in the long run but our intended audience is currently neuroscientists, since memory access is inherently noisy in neural circuits. We aim to use this work to generate testable predictions as to what should be observed in neural circuits when learning complex tasks that require memory access. One possibility is that more neurons are devoted to specific state-action pairs (in parietal cortex or in the basal ganglia, where q-values are putatively encoded) or memory of the q-values might be sampled for a longer duration when higher precision is warranted, thus modulating the speed-accuracy trade-off characteristic of decision making with sequential sampling. We aim to explore these ideas in future work.

Disambiguating DRA from approaches in RL (R4). R4 is correct in pointing out that other groups have proposed alternative approaches to deal with memory limitations in RL, such as using regularization (SAC [A1]), or using neural networks for representing policy and value functions, and even compressing state representations with graph Laplacian [A2]. Our work is meant to complement these previous studies. SAC, for instance, directly penalizes the policy entropy while maximizing reward to encourage exploration. In DRA, we penalize precise representations of (q-)values instead. The use of Laplacian in RL [A2], on the other hand, hints at yet another problem involving efficient use of memory – compact representation of states (*e.g.*, *chunking*) – which is something we very much look forward to addressing in future work. We modified the *Related work* section to discuss and point out how these proposals complement our work.

Justifying our assumptions (R4). In our work, we assume that agents have limited capacity to store (q-)values in their memory, but that they can observe and store states and actions perfectly. This choice is deliberate. It allowed us to develop a normative solution to the problem of allocating resources to items in memory. As far as we know, we are the first ones to provide such a solution. We agree with R4 that it would be important in the future to combine our approach with alternatives that focus instead on regularizing the policy or compressing states, including heuristic approaches such as truncation of planning trees, a strategy suggested Huys et al. [A3]. However, we hope that R4 will agree with us and other reviewers that the existence of other approaches does not take anything away from our contribution: We propose a normative solution to an important problem that "*is of high interest to the community (R2)*", and which will eventually make concrete experimental predictions. "*This is not a paper searching for state of the art results, and it should not be treated as such; rather, it is an exposition of a particular idea, and it did well to explore it (R1)*".

Convergence and baseline (R1). Following a suggestion from R1, we found with new analyses that the asymptotic values of memory precision, σ_*, are largely independent of the choices of initial value σ_0. We also found that convergence speed does not depend on the difference between the initial and optimal asymptotic values. Also, we can confirm that letting $\lambda \to 0$ reduces DRA to SARSA. As suggested by R1, we pick $\lambda = 0$ as the baseline to compare convergence speed and found that DRA is $1.5\times$ and $1.4\times$ slower in the grid-world and mountain car tasks respectively.

Comparison with equal resource allocation (R1, R3). We tested DRA against a model that allocates equal resources to all memories for the grid-world and mountain car tasks (ref. Fig. 3d), and report a $2\times$ and $1.3\times$ improvement in the objective (Eq.1) respectively with flexible resource allocation (DRA). We added these results to the revised manuscript.

Sampling trajectories (R3). Conceptually, the sampling of memories is similar to Dyna, with the difference being that instead of randomly sampling individual memories, DRA samples entire trajectories *on-policy* (though they could also be drawn from stored episodic memories). In DRA, replays are entirely forward. In our current implementations, we sample trajectories at the end of each trial from the starting state for that trial to termination. In principle, we could also do so multiple times during the trial, *e.g.* at each state, and not necessarily until termination.

Generalizing DRA (R2, R4). We aim to address non-independent memories and continuous state spaces in future by considering a Gaussian process prior over the q-values and using GPSARSA [A4] instead of SARSA to update the mean q-values, and extend to non-tabular settings by incorporating compact state representations, *e.g.*, [A2].

Response to remaining comments. **R1:** We have now clarified the meaning of "more" or "less" resources in the text, but we insist that our arguments apply to all systems that are restricted to sample from value distributions but cannot access its mean and precision directly. We depicted the normalized entropy in Fig. 1 simply to ease visualization, but performed appropriate checks as mentioned earlier. **R2:** The mountain car task was included to demonstrate general applicability of DRA to arbitrary problems. We now dedicate a section in the main text to *Related work* and discuss [A1, A2, A5]. **R3:** Figs. 1b & 1c come from different simulations. We have rewritten the caption for Fig. 3 (also suggested by **R4**) clarifying the y-axis confusion in 3f. We also mention that a_{dec} & $t_{\text{non-dec}}$ come from simulations. **R4:** In all tasks, we systematically varied the values of λ & σ_{base} and report that the qualitative results hold for arbitrary values of these parameters with σ_{base} having a slightly stronger effect than λ.

Code release We intend to release the code on GitHub as soon as the submissions are no longer anonymous.

References. [A1] Haarnoja et al. *ArXiV* 2018. [A2] Wu et al., *ArXiV* 2018. [A3] Huys et al. *PNAS* 2015. [A4] Engel et al. *ICML* 2005. [A5] Mattar & Daw. *Nat. Neuro.* 2019.