Supplemental Materials for:
Environmental statistics and the trade-off between model-based and TD learning in humans

Dylan A. Simon
Department of Psychology
New York University
New York, NY 10003
dylex@nyu.edu

Nathaniel D. Daw
Center for Neural Science and Department of Psychology
New York University
New York, NY 10003
nathaniel.daw@nyu.edu

1   Model-based learning rate

We have shown that the TD system is most encumbered in cases of low environmental noise. In psychology and neuroscience, it has been suggested that the S–R learning system is evolutionarily older, and so it would make sense that a model-based system developed specifically to address these deficiencies. If this is the case, such a system may well have become specialized for these low noise conditions, and in particular may only be able to learn relatively quickly, for example with a learning rate of 1. This is also broadly consistent with psychological theory, in which a cognitive, declarative system is thought to depend on exact rules. It would also allow a trade-off between systems solely on the basis of performance, without the need for a cost factor. Here we analyze the performance of such a system, which we will refer to as MB1.

1.1   Theory

When adopting a learning rate of 1, the uncertainty in estimating a variable, X, undergoing a Gaussian diffusion process will be:

$$U_X(1) = \left\langle (\hat{X} - \bar{X})^2 \right\rangle = \sigma^2 + \varepsilon^2 \tag{1}$$

When so limited, an MB1 system attempting to learn values in an MDP will have an uncertainty of this form. In the simple case of two equi-probable outcomes, this uncertainty is still smaller than MC whenever the volatility is sufficiently large:

$$U(1) < U_{\mathrm{MC}}(\alpha^*) \quad \Longleftrightarrow \quad \sigma > \frac{2\varepsilon^2}{d}$$

In particular, MB1 still outperforms MC in most of the critical regions, as shown by the dashed lines in Figure 1.

1.2   Simulation

We performed the same simulations described in the text using a version of the model-based learner using a reward learning rate of 1. The results comparing this to SARSA(0) are shown in Figure 2, and indeed MB1 still out-performs the TD system in the critical areas.

1.3   Human behavior

While we do not explicitly test whether the human model-based system is capable of averaging, there are many suggestions that it is not. If this is the case, it may be that the decrease in model-based effect with increased noise in our data is due instead to the increasing disadvantage of MB1.

1
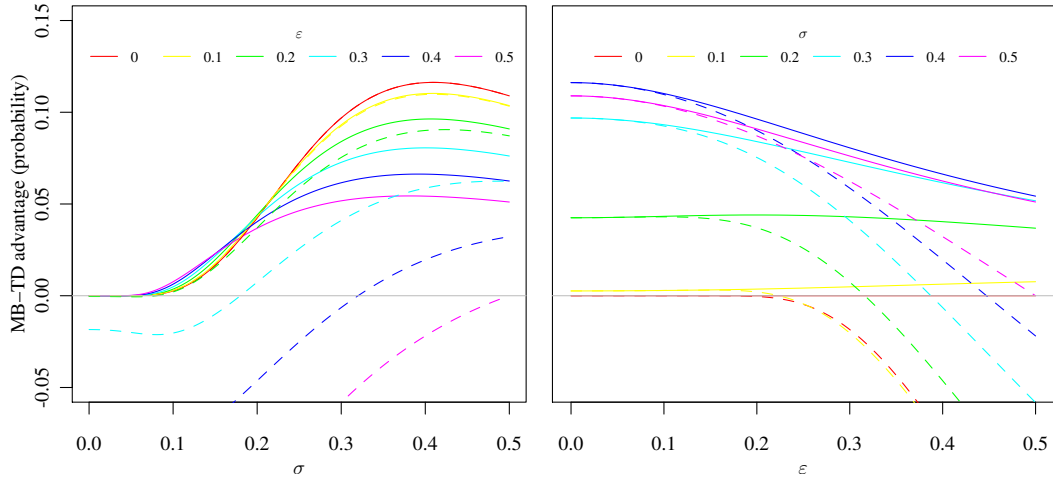
Figure 1: Difference in theoretical success rate between MB (solid) or MB1 (dashed) and MC.
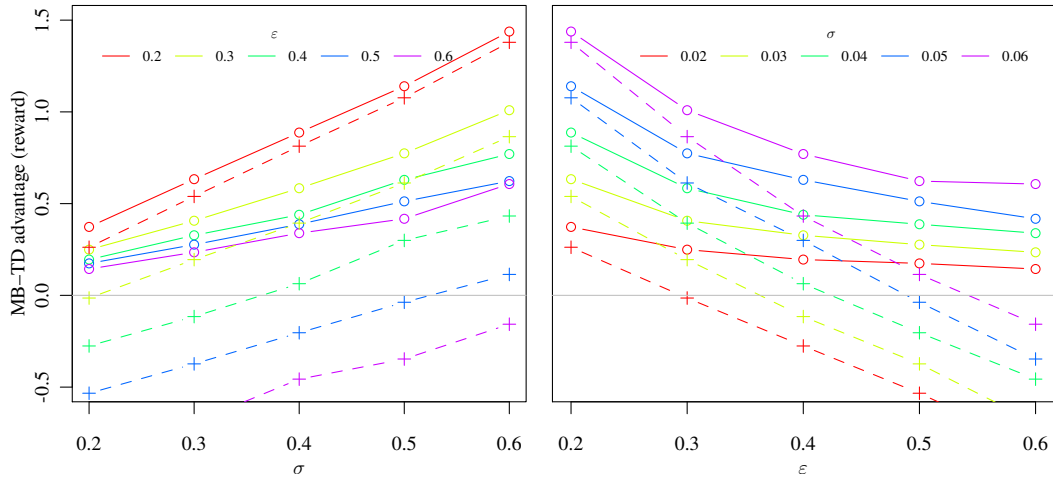
Figure 2: Difference in reward obtained between MB (solid) or MB1 (dashed) and SARSA(0).

2

2 Extension of theory to continuous MDPs

The theoretical results can be extended directly to simple types of continuous MDPs. We will consider the extension of the case given in the text, where each action leads to one of two other states with equal probability. For this we must assume that each reward is sampled on average equally often (a strong form of ergodicity) or not at all. This allows us to identify a single σ parameter that applies identically to all rewards between each sample. Along this line we also ignore the model-based benefit of sharing values between states with common outcomes (which is admittedly often considered a key advantage), as the specific disadvantage of MC methods will still play a role.

Consider the problem of estimating the value function, Q, under a policy, π:

$$\bar{Q}_t^\pi(s) = \left\langle \sum_{i=0} \gamma^i R_{t+i}(s_{t+i}, \pi(s_{t+i})) \right\rangle$$

The Gaussian diffusion process on the reward values induces a similar process on Q itself, which may either be sampled directly by observing trajectories (as in the case of TD(1)) or indirectly by sampling R and computing $\hat{Q}$ (as in the case of model-based). Because the stochastic transitions smooth over multiple rewards, the effective volatility will decrease:

$$\left\langle (\bar{Q}_{t+1} - \bar{Q}_t)^2 \right\rangle = \sum_{i=0} \frac{\gamma^{2i}}{2^i} \sigma^2 = \frac{\sigma^2}{1 - \frac{\gamma^2}{2}}$$

As for noise, we get a result analogous to the episodic case, where variance between the values of different states increases the effective noise:

$$\left\langle (Q - \bar{Q})^2 \right\rangle = \frac{\varepsilon^2 + \frac{\gamma^2}{4} \operatorname{var} \bar{Q}}{1 - \gamma^2}$$

This increased noise only affects those strategies that learn by taking samples of Q, like TD. As a result, the optimal learning rate for TD will decrease, resulting in increased uncertainty. (TD(0) can also be analyzed in this framework by solving the noise process induced by value updates and the Kalman uncertainty recursively.)

As we extend this to more complex MDPs, the exact trade-off will depend on many factors, but in general the uncertainty of a TD system is dominated by volatility, while the model-based learner is hindered more by noise. The pattern should also hold for other noise models, including discontinuous change points, as long as they are i.i.d. between different rewards in the environment. Finally, different learning algorithms may differ in their rate of convergence to stationarity, which we expect will be affected by similar factors.

3