Dopamine Ramps Are a Consequence of Reward Prediction Errors

Samuel J. Gershman
sjgershm@mit.edu
Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, U.S.A.

Temporal difference learning models of dopamine assert that phasic levels of dopamine encode a reward prediction error. However, this hypothesis has been challenged by recent observations of gradually ramping stratal dopamine levels as a goal is approached. This note describes conditions under which temporal difference learning models predict dopamine ramping. The key idea is representational: a quadratic transformation of proximity to the goal implies approximately linear ramping, as observed experimentally.

1 Introduction

Temporal difference (TD) learning is arguably the most successful account of dopamine function in the basal ganglia (Glimcher, 2011; Niv & Schoenbaum, 2008; Schultz, Dayan, & Montague, 1997). According to this account, phasic dopamine signals a reward prediction error—the discrepancy between observed and predicted reward—and this signal is used to improve future predictions. Recently, Howe, Tierney, Sandberg, Phillips, & Graybiel (2013) reported a form of dopaminergic activity that appears (at first glance) to fly in the face of the prediction error hypothesis: as a rat approaches the goal in a maze, dopamine levels in the striatum gradually ramp up, peaking when the rat arrives at the goal. As Niv (2013) pointed out, this observation is unanticipated by existing TD models.

In this note, we describe conditions under which the TD model predicts ramping. The essential assumption pertains to the representation of space: provided that proximity to the goal is encoded by a convex transformation, ramping will be observed. In particular, the near-linear ramping that Howe et al. (2013) observed occurs when the proximity transformation is quadratic.

2 Temporal Difference Learning

TD learning is an algorithm for estimating the value function (expected discounted future return):

$$V_t = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right], \tag{2.1}$$

Neural Computation **26, 467–471** (2014)
doi:10.1162/NECO_a_00559

where t indexes time, r_t is the reward delivered at time t, and $\gamma \in [0, 1)$ is a discount factor. Here the expectation is taken over possibly random sequences of rewards; in the remainder of this note, we assume for simplicity that rewards and transitions are deterministic.

TD models of the dopamine system typically assume a linear approximation of the value function

$$V_t = \sum_k x_t(k) w_t(k), \tag{2.2}$$

where $x_t(k)$ is the kth feature of the state at time t and $w_t(k)$ is the feature weight. The weight vector is updated according to

$$w_{t+1}(k) = w_t(k) + \alpha \delta_t x_t(k), \tag{2.3}$$

where $\alpha \in [0, 1)$ is a learning rate and δ_t is the reward prediction error:

$$\delta_t = r_t + \gamma V_{t+1} - V_t. \tag{2.4}$$

TD models of the dopamine system assert that δ_t is represented by the firing of midbrain dopamine neurons and, by extension, dopamine levels in the striatum.

3 Modeling Spatial Navigation

To model the experiment of Howe et al. (2013), we assume a very simple feature representation consisting of a single feature encoding proximity to the goal. Thus, $x_t = 1$ when the rat has reached the end of the maze, declining with distance from the end. Letting p_t denote the proximity to the goal, we define $x_t = f(p_t)$, where f is a transformation that encodes the neural representation of proximity. We assume that p_t monotonically increases with t (i.e., the rat is continuously progressing toward the goal). Under the assumption of a deterministic environment, we can analytically solve for the asymptotic value of the feature weight (Bradtke & Barto, 1996),

$$w_\infty = \frac{\sum_{t=1}^{T} x_t r_t}{\sum_{t=1}^{T} x_t (x_t - \gamma x_{t+1})}, \tag{3.1}$$

where T denotes the goal state. Here $r_t = 1$ if $t = T$, and 0 otherwise. Note that $w_\infty \geq 0$ provided that $x_t \geq 0$ and $r_t \geq 0$ for all t.

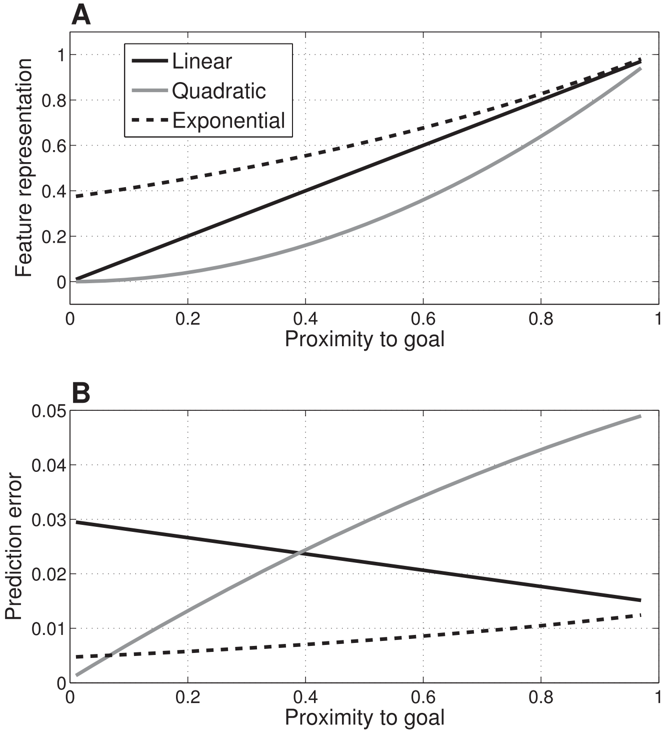

Figure 1: Illustration of dopamine ramping. (A) Proximity representation. (B) Prediction error as a function of proximity.

4 Why Does Ramping Occur?

We can express the asymptotic prediction error as

$$\delta_t = r_t + w_\infty[\gamma f(p_{t+1}) - f(p_t)]. \tag{4.1}$$

We will focus on the time points prior to the goal, $t < T$, and thus the reward term is always 0. To get a ramping prediction error as the rat approaches the goal, we require that $\gamma f(p_{t+1}) - f(p_t)$ is a monotonically increasing function of t (recall that p_t is a monotonically increasing function of t). This is satisfied by any strictly convex proximity transformation when $\gamma \approx 1$. We obtain an approximately linear ramp when the proximity transformation is quadratic, $f(p) = \kappa p^2$, where κ is a constant. To see this, note that $f'(p) = 2\kappa p$, which coincides with $\gamma f(p_{t+1}) - f(p_t)$ in the limit $\gamma \to 1$ and $|p_{t+1} - p_t| \to 0$.

The predicted ramping behavior is illustrated in Figure 1 using $\gamma = 0.99$ and $\kappa = 1$. For comparison, we also show the results for linear and exponential proximity transformations. Although there is a slight ramping predicted by the exponential transformation (and this ramping can be made stronger by increasing the slope of the exponential transformation), the ramping is always convex, which is inconsistent with the near-linear (and sometimes slightly concave) ramping observed by Howe et al. (2013).

Several assumptions were made in these simulations for convenience rather than mathematical necessity. First, the proximity transformation was configured to monotonically increase from 0 to 1, but ramping will occur for any monotonically increasing convex transformation. Second, we assumed a one-dimensional proximity representation, but this can be generalized: any nonnegative combination of convex functions is convex, and therefore ramping will occur as long the asymptotic weights are nonnegative and each feature is computed by a convex transformation. A corollary of this assumption is that the spatial representation is not a form of table look-up (Sutton & Barto, 1998), since table look-up is incompatible with a graded representation of space. Although earlier TD models of dopamine used a form of table look-up (e.g., Daw, Courville, & Touretzky, 2006; Schultz et al., 1997), more recent models have emphasized the importance of graded, distributed representations (e.g., Gustafson & Daw, 2011; Kurth-Nelson & Redish, 2009; Ludvig, Sutton, & Kehoe, 2008).

Howe et al. (2013) made a number of other observations that are consistent with this model: (1) ramps leading to similar rewards peaked at similar levels despite differences in running speed; (2) ramps leading to large rewards exceeded ramps leading to small rewards; and (3) the ramps dynamically changed when large and small rewards switched locations. The insensitivity to running speed (and hence time until the goal is reached) arises because the prediction errors near the goal will be the same regardless of how long it took to get there. The model is reward sensitive because the asymptotic weight scales with reward, as stipulated by equation 3.1. When rewards switch locations, the corresponding asymptotic weights will switch, leading to the observed ramp dynamics.

5 Conclusion

The gradual ramping of dopamine activity as a rat approaches a goal is consistent with the basic predictions of TD models. The special ingredient is a convex transformation of proximity to the goal. This transformation implies a spatial compression of the value function similar to Weber's law, such that values of locations far from the goal are closer together than values of locations near the goal. Interestingly, landmark-based compression of space has been reported in several species (Cheng, 1990; Cheng, Srinivasan, & Zhang, 1999), as well as in the hippocampal representation of space (O'Keefe & Burgess, 1996). We may speculate that the source of ramping lies in the hippocampal inputs to the striatum, which are thought to provide the features for value functions defined over space (Foster, Morris, & Dayan, 2000; Gustafson & Daw, 2011).

Acknowledgments

This work was supported by a postdoctoral fellowship from the MIT Intelligence Initiative. I thank Yael Niv and Ann Graybiel for helpful discussions.

References

Bradtke, S. J., & Barto, A. G. (1996). Linear least-squares algorithms for temporal difference learning. *Machine Learning*, *22*, 33–57.

Cheng, K. (1990). More psychophysics of the pigeon's use of landmarks. *Journal of Comparative Physiology A*, *166*, 857–863.

Cheng, K., Srinivasan, M., & Zhang, S. (1999). Error is proportional to distance measured by honeybees: Weber's law in the odometer. *Animal Cognition*, *2*, 11–16.

Daw, N. D., Courville, A. C., & Touretzky, D. S. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, *18*, 1637–1677.

Foster, D., Morris, R., & Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, *10*, 1–16.

Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*, 15647–15654.

Gustafson, N. J., & Daw, N. D. (2011). Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Computational Biology*, *7*(10), e1002235.

Howe, M. W., Tierney, P. L., Sandberg, S. G., Phillips, P. E., & Graybiel, A. M. (2013). Prolonged dopamine signalling in striatum signals proximity and value of distant rewards. *Nature*, *500*, 575–579.

Kurth-Nelson, Z., & Redish, A. D. (2009). Temporal-difference reinforcement learning with distributed representations. *PloS One*, *4*, e7362.

Ludvig, E. A., Sutton, R. S., & Kehoe, E. J. (2008). Stimulus representation and the timing of reward-prediction errors in models of the dopamine system. *Neural Computation*, *20*, 3034–3054.

Niv, Y. (2013). Neuroscience: Dopamine ramps up. *Nature*, *500*, 533–535.

Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*, 265–272.

O'Keefe, J., & Burgess, N. (1996). Geometric determinants of the place fields of hippocampal neurons. *Nature*, *381*, 425–428.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Received August 30, 2013; accepted October 3, 2013.