+++ title = “Successor Representations” subtitle = ""

Add a summary to display on homepage (optional).

summary = ""

date = 2019-02-25T08:32:46+01:00 draft = false

Authors. Comma separated list, e.g. `["Bob Smith", "David Jones"]`.

authors = [“Julien Vitay”]

Tags and categories

For example, use `tags = []` for no tags, or the form `tags = ["A Tag", "Another Tag"]` for one or more tags.

tags = [“Reinforcement Learning”, “Machine Learning”, “Dopamine”] categories = []

Projects (optional).

Associate this post with one or more of your projects.

Simply enter your project’s folder or file name without extension.

E.g. `projects = ["deep-learning"]` references

`content/project/deep-learning/index.md`.

Otherwise, set `projects = []`.

projects = [“dopamine”, “reinforcement-learning”]

math = true

Featured image

To use, add an image named `featured.jpg/png` to your page's folder.

[image] # Caption (optional) caption = ""

Focal point (optional) # Options: Smart, Center, TopLeft, Top, TopRight, Left, Right, BottomLeft, Bottom, BottomRight focal_point = "" +++

Model-free vs. model-based RL

There are two main families of **reinforcement learning** (RL; Sutton and Barto, 2017) algorithms:

- **Model-free** (MF) methods estimate the value of a state $V^\pi(s)$ or of a state-action pair $Q^\pi(s,a)$ by sampling trajectories and average the obtained returns:

$$V^\pi(s\_t) = \mathbb{E}\_\pi[\sum\_k = 0\gamma^k\, r\_t + k + 1] = \mathbb{E}\_\pi[r\_t + 1 + \gamma\, V^\pi(s\_t + 1)]$$

$$Q^\pi(s\_t, a\_t) = \mathbb{E}\_\pi[\sum\_k = 0\gamma^k\, r\_t + k + 1] = \mathbb{E}\_\pi[r\_t + 1 + \gamma\, Q^\pi(s\_t + 1, a\_t + 1)]$$

- **Model-based** (MB) methods use (or learn) a model of the environment - transition probabilities $p(s\_t + 1|s\_t, a\_t)$ and reward probabilities $r(s\_t, a\_t, s\_t + 1)$ - and use it to plan trajectories maximizing the theoretical return, either through some form of forard planning (search tree) or dynamic programming (solving the Bellman equations).

$$\max\_\pi\, p(s\_0)\, \sum\_t = 0^\infty \gamma^t\, p(s\_t + 1|s\_t, a\_t)\, \pi(s\_t, a\_t)\, r(s\_t, a\_t, s\_t + 1)$$

The main advantage of model-free methods is their speed: they *cache* the future of the system into value functions. When having to take a decision at time t, we only need to look at the action with the highest Q-value in the state s_t and take it. If the Q-values are optimal, this is the optimal policy. Oppositely, model-based algorithms have to plan sequentially in the state-action space, what can be very long if the problem has a long horizon.

The main drawback of MF methods is their *inflexibility* when the reward distribution changes. When the reward associated to a transition changes (the source of reward has vanished, or its nature has changed), each action leading to that transition has to be experienced multiple times before the corresponding

values reflect that change. This is due to the use of the **temporal difference** (TD) algorithm, where the **reward prediction error** (RPE) is used to update values:

$$\delta\_t = r\_t + 1 + \gamma\, V^{\pi}(s\_t + 1) - V^{\pi}(s\_t)$$

$$\Delta V^{\pi}(s\_t) = \alpha\, \delta\_t$$

When the reward associated to a transition changes drastically, only the last state (or action) is updated after that experience (unless we use eligibility traces). Only multiple repetitions of the same trajectory would allow to change the initial decisions. This is opposite to MB methods, where a change in the reward distribution would very quickly influence the planning of the optimal trajectory. The reward probabilities can be estimated with:

$$\Delta r(s\_t, a\_t, s\_t + 1) = \alpha\, (r_{t+1} - r(s\_t, a\_t, s\_t + 1))$$

with r_{t+1} being the reward obtained during one sampled transition. The transition probabilities can also be learned from experience using:

$$\Delta p(s'|s\_t, a\_t) = \alpha\, (\mathbb{1}(s_{t+1} = s') - p(s'|s\_t, a\_t))$$

Depending on the learning rate, changes in the environment dynamics can be very quickly learned by MB methods, as updates do not depend on other estimates (bootstrapping).

The model-free RPE has become a very influential model of dopaminergic (DA) activation in the ventral tegmental area (VTA) and substantia nigra pars compacta (SNc). At the beginning of classical Pavlovian conditioning, DA cells react phasically to unconditioned stimuli (US, rewards). After enough conditioning trials, DA cells only react to conditioned stimuli (CS), i.e. stimuli which predict the delivery of a reward. Moreover, if the reward is omitted, DA cells exhibit a pause in firing. This pattern of activation corresponds to the RPE: DA cells respond to unexpected reward event, either positively when more reward than expected is received, or negatively when less reward is delivered. The simplicity of this model has made RPE a successful model of DA activity (but see Vitay and Hamker, 2014).

A similar but not identical functional dichotomy opposes deliberative **goal-directed** behavior and reflexive **habits** (Dickinson and Balleine, 2002). Goal-directed behavior is sensitive to reward devaluation: if an outcome was previously rewarding but ceases to be (for example, an unpleasant product is injected into some food), goal-directed behavior would quickly avoid that outcome, while

habitual behavior will continue to seek for it. Overtraining can transform goal-directed behavior into habits (Corbit and Balleine, 2011). Habits are usually considered as a model-free learning behavior, while goal-directed behavior implies the use of a world model.

Both forms of behavior are thought to happen concurrently in the brain, with model-based / goal-directed behavior classically assigned to the prefrontal cortex and the hippocampus and model-free / habitual behavior mapped to the ventral basal ganglia and the dopaminergic system. However, recent results and theories suggest that these two systems are largely overlapping and that even dopamine firing might reflect model-based processes (Doll, Simon and Daw, 2012; Miller et al., 2018). It is yet to be understood how these two extreme mechanisms of the RL spectrum might coexist in the brain: successor representations might provide us with useful insights into the functioning of the brain.

Successor representations

The original formulation of **successor representations** (SR) is actually not recent (Dayan, 1993), but it is subject to a revival since a couple of years with the work of Samuel J. Gershman (e.g. Gershman et al., 2012).

The SR algorithm learns two quantities:

1. The average immediate reward received after each state:

$$r(s) = \mathbb{E}_\pi[r_{t+1}|s_t = s]$$

2. The expected discounted future state occupancy (the **SR** itself):

$$M(s, s') = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k \, 1(s_{t+k+1} = s')|s_t = s]$$

References

Corbit, L. H., and Balleine, B. W. (2011). The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. The Journal of neuroscience 31, 11786–94. doi:10.1523/JNEUROSCI.2711-11.2011.

Dayan, P. (1993). Improving Generalization for Temporal Difference Learning: The Successor Representation. Neural Computation 5, 613–624. doi:10.1162/neco.1993.5.4.613.

Dickinson, A., and Balleine, B. (2002). The role of learning in the operation of motivational systems. In: Gallistel CR, editor. Steven's handbook of experimental psychology: learning, motivation and emotion . 3rd ed.New York: John Wiley & Sons, 497–534.

Doll, B. B., Simon, D. A., and Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. Current Opinion in Neurobiology 22, 1075–1081. doi:10.1016/j.conb.2012.08.003.

Gershman, S.J., Moore, C.D:, Todd, M.T., Norman, K.A., and Sederberg, P.B. (2012). The successor representation and temporal context. Neural Computation, 24(6):1553–1568, 2012.

Miller, K., Ludvig, E. A., Pezzulo, G., and Shenhav, A. (2018). Re-aligning models of habitual and goal-directed decision-making. In Goal-Directed Decision Making : Computations and Neural Circuits, eds. A. Bornstein, R. W. Morris, and A. Shenhav (Academic Press). Available at: https://www.elsevier.com/books/goal-directed-decision-making/morris/978-0-12-812098-9.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J Neurophysiol 80, 1–27.

Sutton, R. S., and Barto, A. G. (2017). Reinforcement Learning: An Introduction. 2nd ed. Cambridge, MA: MIT Press. Available at: http://incompleteideas.net/book/the-book-2nd.html.

Vitay, J., and Hamker, F. H. (2014). Timing and expectation of reward: A neurocomputational model of the afferents to the ventral tegmental area. Frontiers in Neurorobotics 8. doi:10.3389/fnbot.2014.00004.