Model-Based and Model-Free Decision-Making

Neural Modelling 2023
Georgy Antonov

Computational Neuroscience department
Max Planck Institute for Biological Cybernetics

Outline

- Model-based and model-free control
- Dyna
- Hippocampal replay
- Exploration
- Assignment: part 1
- Assignment: part 2
- Questions

Model-based and model-free control

Model-based control

- Learns a model of the environment
- Performs prospective evaluation (planning)

Pros:

- Reflective; affords behavioural flexibility

Cons:

- Expensive; slow

Model-free control

- Learns and stores expected outcomes associated with each state-action pair

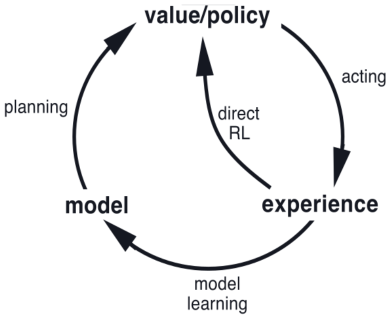
Pros:

- Reflexive; fast
- Computationally cheap

Cons:

- Stubborn; inflexible

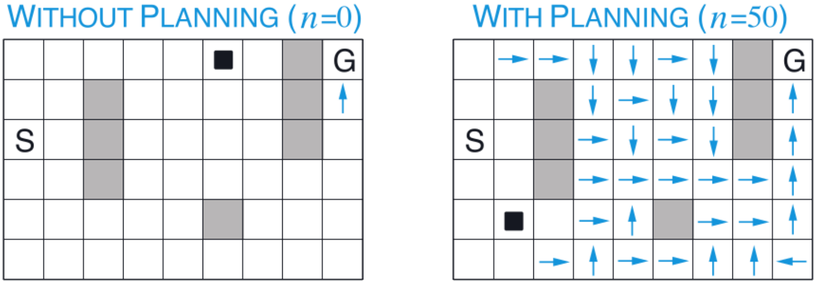
Dyna



Sutton (1990)

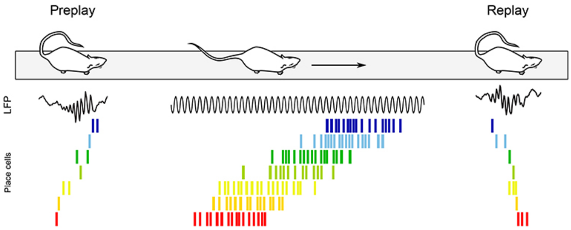
- DYNA is an integrated architecture
- Combines a *reflexive* MF policy and a *reflective* MB system
- MB system is used offline to provide additional training for MF values

Dyna



- Agent discovers prediction erros (e.g., a goal) online
- Model inversion (planning) to additionally train MF values

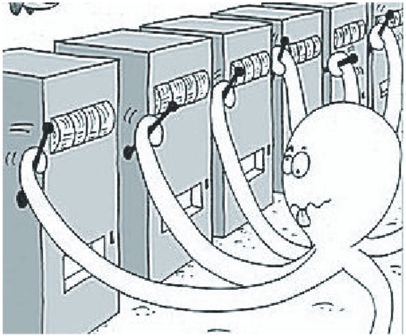
Hippocampal replay



Drieu et al. (2019); Diba et al. (2007)

- Reinstatement of behaviourally-relevant neural activity during periods of quiet wakefullness and sleep [offline periods] (M. A. Wilson et al., 1993)
- The order of the replayed experiences is highly specific
- Forward replay seems to be predictive of the subsequent animal choices (Pfeiffer et al., 2013); reverse replay is highly sensitive to reward (Ambrose et al., 2016)

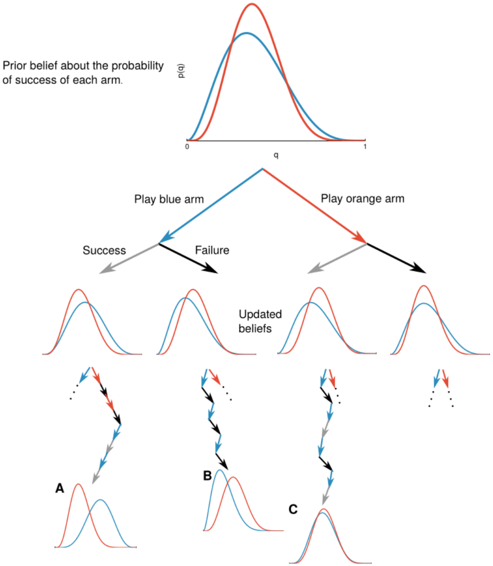
Exploration



Source: link

- Multi-arm bandit is the classic problem for studying the exploration-exploitation tradeoff
- The objective is to maximise discounted expected reward
- Payoff probabilities are unknown
- One of the few problems for which an optimal solution is possible to compute: the Gittins index (Gittins, 1979)
- Some animals explore near-optimally (Krebs et al., 1978)

Optimal exploration



- Optimal exploration amounts to performing optimal control in belief space
- Belief spaces are continuous so forget about tractability in most problems more complex that simple bandits
- Good approximations exist, such as for instance BAMCP (Guez et al., 2012)

Heuristic exploration

- Undirected
  - ϵ-gredy
  - Softmax (Boltzmann)
- Directed, 'optimism in the face of uncertainty'
  - upper confidence bound (Auer, 2002)

$$a = \arg\max_{a} \left[Q_t(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}} \right]$$

Exploration bonus

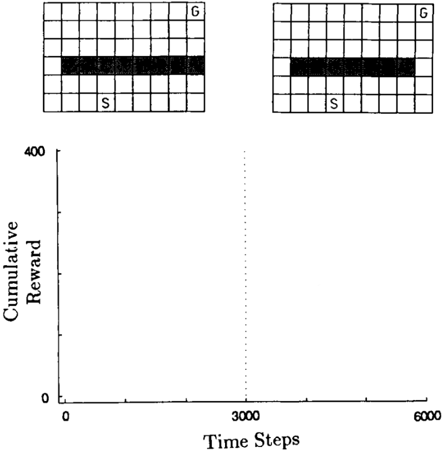
Exploration of humans and other animals is sometimes random, or undirected (Daw, O'Doherty, et al., 2006), sometimes directed (R. C. Wilson et al., 2021)

Exploration signals in the brain

Neuromodulatory systems are thought to broadcast different types of uncertainty throughout the brain (Yu et al., 2005)

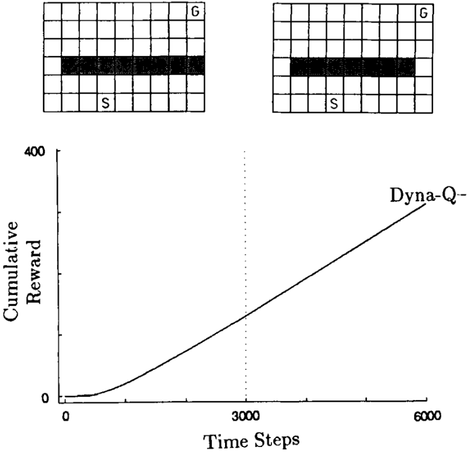
- Acetylcholine (ACh) signals *expected* uncertainty
- Norepinephrine (NE) signals *unexpected* uncertainty
  - Pupil diameter is correlated with NE release. Exploratory choices in humans during an equivalent of multi-arm bandit task were preceded by pupil dilation (Jepma et al., 2011)
  - Administering atomoxetine (NE transporter blocker) increased random exploration in human subjects (Warren et al., 2017)

Dyna exploration



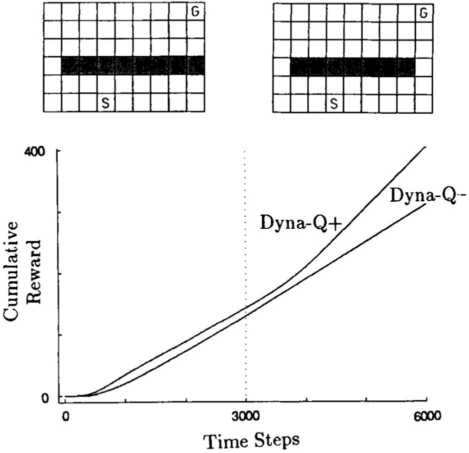
- Sutton (1990)'s changing world example
- Will a 'naive' Dyna agent which performs *Q*-learning updates discover the shortcut?

Dyna exploration



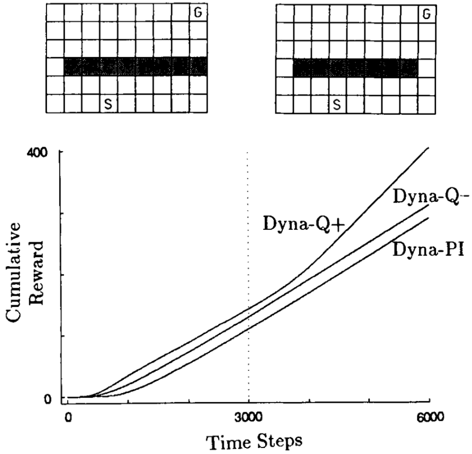
- Sutton (1990)'s changing world example
- Will a 'naive' Dyna agent which performs Q-learning updates discover the shortcut? No
- What if we encourage exploration?

Dyna exploration



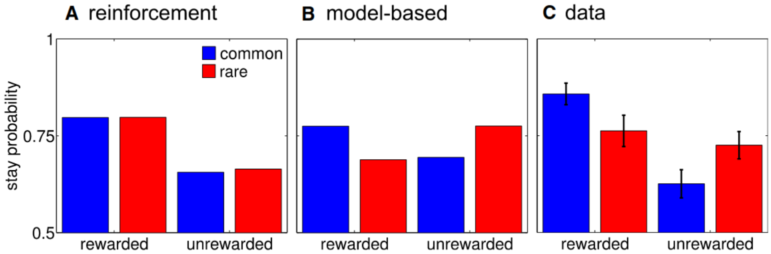
- Sutton (1990)'s changing world example
- Will a 'naive' Dyna agent which performs Q-learning updates discover the shortcut? No
- What if we encourage exploration? Yes

Assignment: part 1. Dyna exploration



- One of the original intensions of Dyna was to improve exploration efficiency
- By incorporating an exploration bonus into the planning updates, uncertainty can propagate to distal states and therefore encourage exploration
- Your task is to reproduce this figure; focus only on Dyna-$Q+$ and Dyna-Q-

Assignment: part 2. Two-step task



- The iconic RL task (Daw, Gershman, et al., 2011) to probe the relative contributions of MB and MF control to subjects' choices
- In this part of the assignment, your task is to reproduce the above figure

Questions?

- You will find the assignment and all the necessary details in my github repository:

  https://github.com/geoant1/GTC_Neural_Modelling_Tutorial

- For part 1 the code is already written for you; the task is to fill in the missing implementation
- For part 2 you have to write most of the code yourself

References I

Ambrose, R. Ellen, Brad E. Pfeiffer, and David J. Foster (Sept. 2016). “Reverse Replay of Hippocampal Place Cells Is Uniquely Modulated by Changing Reward”. In: *Neuron* 91.5, pp. 1124–1136. ISSN: 08966273. DOI: 10.1016/j.neuron.2016.07.047. URL: https://linkinghub.elsevier.com/retrieve/pii/S0896627316304639 (visited on 12/08/2021).

Auer, Peter (2002). “Using Confidence Bounds for Exploitation-Exploration Trade-offs”. In: p. 26.

Daw, Nathaniel D., Samuel J. Gershman, et al. (Mar. 24, 2011). “Model-Based Influences on Humans’ Choices and Striatal Prediction Errors”. In: *Neuron* 69.6, pp. 1204–1215. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2011.02.027. pmid: 21435563. URL: https://www.cell.com/neuron/abstract/S0896-6273(11)00125-5 (visited on 07/23/2023).

Daw, Nathaniel D., John P. O’Doherty, et al. (June 2006). “Cortical Substrates for Exploratory Decisions in Humans”. In: *Nature* 441.7095 (7095), pp. 876–879. ISSN: 1476-4687. DOI: 10.1038/nature04766. URL: https://www.nature.com/articles/nature04766 (visited on 08/16/2022).

Diba, Kamran and György Buzsáki (Oct. 2007). “Forward and Reverse Hippocampal Place-Cell Sequences during Ripples”. In: *Nature Neuroscience* 10.10, pp. 1241–1242. ISSN: 1097-6256, 1546-1726. DOI: 10.1038/nn1961. URL: http://www.nature.com/articles/nn1961 (visited on 12/07/2021).

Drieu, Céline and Michaël Zugaro (2019). “Hippocampal Sequences During Exploration: Mechanisms and Functions”. In: *Frontiers in Cellular Neuroscience* 13. ISSN: 1662-5102. URL: https://www.frontiersin.org/article/10.3389/fncel.2019.00232 (visited on 03/07/2022).

References II

Gittins, J. C. (Jan. 1979). “Bandit Processes and Dynamic Allocation Indices”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 41.2, pp. 148–164. ISSN: 00359246. DOI: 10.1111/j.2517-6161.1979.tb01068.x. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1979.tb01068.x (visited on 12/07/2021).

Guez, Arthur, David Silver, and Peter Dayan (2012). “Efficient Bayes-Adaptive Reinforcement Learning Using Sample-Based Search”. In: *Advances in Neural Information Processing Systems*. Vol. 25. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2012/hash/35051070e572e47d2c26c241ab88307f-Abstract.html (visited on 02/09/2022).

Jepma, Marieke and Sander Nieuwenhuis (July 1, 2011). “Pupil Diameter Predicts Changes in the Exploration–Exploitation Trade-off: Evidence for the Adaptive Gain Theory”. In: *Journal of Cognitive Neuroscience* 23.7, pp. 1587–1596. ISSN: 0898-929X. DOI: 10.1162/jocn.2010.21548. URL: https://doi.org/10.1162/jocn.2010.21548 (visited on 06/26/2023).

Krebs, John R., Alejandro Kacelnik, and Peter Taylor (Sept. 1978). “Test of Optimal Sampling by Foraging Great Tits”. In: *Nature* 275.5675 (5675), pp. 27–31. ISSN: 1476-4687. DOI: 10.1038/275027a0. URL: https://www.nature.com/articles/275027a0 (visited on 06/19/2023).

Pfeiffer, Brad E. and David J. Foster (May 2013). “Hippocampal Place-Cell Sequences Depict Future Paths to Remembered Goals”. In: *Nature* 497.7447, pp. 74–79. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature12112. URL: http://www.nature.com/articles/nature12112 (visited on 12/07/2021).

References III

- Sutton, Richard S. (1990). "Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming". In: *Machine Learning Proceedings 1990*. Elsevier, pp. 216–224. ISBN: 978-1-55860-141-3. DOI: 10.1016/B978-1-55860-141-3.50030-4. URL: https://linkinghub.elsevier.com/retrieve/pii/B9781558601413500304 (visited on 12/07/2021).

- Warren, Christopher M. et al. (Apr. 26, 2017). "The Effect of Atomoxetine on Random and Directed Exploration in Humans". In: *PLOS ONE* 12.4, e0176034. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0176034. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176034 (visited on 06/25/2023).

- Wilson, Matthew A. and Bruce L. McNaughton (Aug. 20, 1993). "Dynamics of the Hippocampal Ensemble Code for Space". In: *Science* 261.5124, pp. 1055–1058. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.8351520. URL: https://www.science.org/doi/10.1126/science.8351520 (visited on 12/07/2021).

- Wilson, Robert C et al. (Apr. 2021). "Balancing Exploration and Exploitation with Information and Randomization". In: *Current Opinion in Behavioral Sciences* 38, pp. 49–56. ISSN: 23521546. DOI: 10.1016/j.cobeha.2020.10.001. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352154620301467 (visited on 12/07/2021).

- Yu, Angela J. and Peter Dayan (May 19, 2005). "Uncertainty, Neuromodulation, and Attention". In: *Neuron* 46.4, pp. 681–692. ISSN: 0896-6273. DOI: 10.1016/j.neuron.2005.04.026. URL: https://www.sciencedirect.com/science/article/pii/S0896627305003624 (visited on 06/26/2023).