Place fields organize along goal trajectory with reinforcement learning

M Ganesh Kumar Cengiz Pehlevan
SEAS, Harvard University, Cambridge, MA 02138 USA
{mganeshkumar, cpehlevan}@seas.harvard.edu

Abstract

When rodents learn goal-directed navigation, a high density of place fields form at reward locations, and the fields increase in width and skew against the movement direction. However, a normative framework to characterize the field distribution during task learning remains elusive. We hypothesize that the observed place field dynamics is a feature of state representation learning that helps policy learning maximize the reinforcement learning objective. We develop an agent that uses Gaussian basis functions to model place fields which directly synapse to a policy network. Each field's center, width and amplitude, and the policy parameters are updated trial by trial to maximize the cumulative discounted reward. When the agent learns to navigate to a goal in a one-dimensional track or two-dimensional environment with obstacles, a higher number of Gaussian fields organize near the goal while the rest of the fields increase in width to tile the goal trajectory. We show that the correlation between the frequency of being in a location and the field density at that location increases with training, as postulated by the efficient coding hypothesis. Additionally, Gaussian fields elongate along the goal trajectory aggregating future positions with similar actions, resembling a successor representation-like map. We further show that this learned map facilitates faster policy convergence, when the number of basis functions is low. To conclude, we develop a normative model that recapitulates several hippocampus place field learning dynamics and unify alternative proposals to offer testable predictions for future experiments.

Keywords: Place field dynamics; State representation learning; Normative navigation model

Introduction

Place fields have been studied extensively given their causal role in goal-directed navigation (Steele & Morris, 1999). Canonical navigation tasks require animals to find a hidden goal either in a 1D track or 2D arena to receive a reward. At the start of learning, place fields are distributed throughout the environment but reorganize with a higher density of fields forming at the goal location (Hok et al., 2007; Lee, Briguglio, Cohen, Romani, & Lee, 2020). Furthermore, field widths and centers increase and skew against the direction of movement respectively (Mehta, Quirk, & Wilson, 2000), although there are other fields that become narrower with experience (Frank, Stanley, & Brown, 2004). Although there are several proposals to characterize place field representations (Ganguli & Simoncelli, 2014; Stachenfeld, Botvinick, & Gershman, 2017), it is unclear why these representations form and how they change during goal-directed learning. Whether this representation results in faster policy learning also remains elusive. Here, we develop a normative navigation model whose Gaussian field parameters and navigation policy are optimized using a reinforcement learning objective to recapitulate several place field phenomena while demonstrating faster policy convergence. The model suggests a unification of disparate proposals and proposes testable predictions for neural experiments.

Methods

We model each place field's firing rate ϕ_i using a Gaussian basis function $\phi_i(x_t) = \alpha_i * \exp\left(-\frac{||x_t-\lambda_i||^2}{2\sigma_i^2}\right)$. The field centers uniformly tile the 1D track or 2D arena with constant widths $\sigma_i = 0.1$ **(Fig. 1A)**. The fields directly synapse to an action selector using $\pi(a_t^j|x_t) = softmax(\sum_{i=0}^{N} W_{ij}^{\top}\phi_i(x_t))$ that stochastically selects one of two (left, right) or four (left, right, up, down) discrete actions in either the track or arena respectively. The discrete actions are converted to a velocity metric so that the actual step taken in the environment is continuous (Kumar, Tan, Libedinsky, Yen, & Tan, 2022) and capped at 0.1. The objective function is the cumulative discounted reward $J(\theta) = \mathbb{E}_{a\sim\pi}\left[\sum_{k=0}^{T}\gamma^k r_{t+1+k}\right]$ and is maximized by optimizing the parameters $\theta = \{W_{ij}, \lambda_i, \sigma_i, \alpha_i\}$ using the Policy Gradient algorithm $\nabla_\theta J(\theta) = \mathbb{E}_{a\sim\pi}\left[\sum_{t=0}^{T}\nabla_\theta \log\pi(a_t^j|x_t)\sum_{k=0}^{T}\gamma^k r_{t+1+k}\right]$ (Sutton, McAllester, Singh, & Mansour, 1999).

Results

We first study how place fields uniformly distributed **(Fig. 1A)** on a continuous 1D track will reorganize when an agent learns to navigate from the start (green line at -0.5) to the goal (orange line at 0.5) with a radius of 0.01 to receive a reward of value 1, after which the trial ends. Since the goal radius is small, the agent has to reduce its velocity at the goal to avoid passing over and receive the reward. After 2000 learning trials, a high number of fields organize after the start and just after the goal while the rest of the fields stretch along the trajectory **(Fig. 1B)**. We consider the field density **(Fig. 1D)** as the summation of the population firing rate at a location $d(x) = \sum_n^N \phi_n(x)$. When uniformly distributed, field density is constant (blue) along the track, and after 500 trials, field density increases before the goal (orange). After 2000 trials, the field density peaks after the start location and ramps down before peaking again after the goal.

An agent's visit frequency $f(x)$ is the proportion of time spent at specific locations of the environment, and is aggregated over 5 trials and plotted as histograms in **(Fig. 1C)**. In

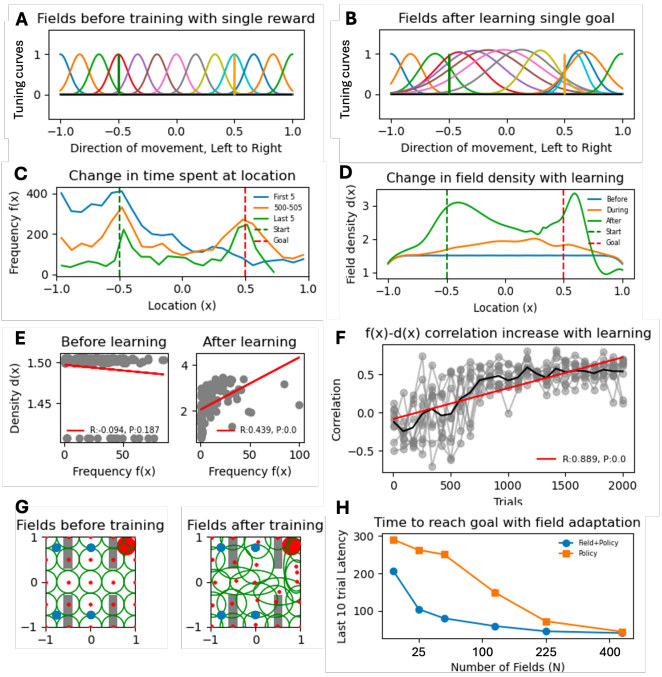

Figure 1: **(a)** 13 uniformly distributed fields on a 1D track. Agent navigates from the start (green line) to a goal (orange line). **(b)** After learning, fields stretch along the trajectory and organize after the goal. **(c)** Frequency of time spent at the start location shifts towards the goal location with learning. Histogram aggregated over 10 iterations. **(d)** Before learning, field density is uniform along the track (blue), but after learning, the density peaks at the start and ramps down to the goal before peaking again after the goal (green). Average density over 10 iterations. **(e)** Correlation between frequency and density is zero (left) and positive (right) before and after learning respectively. Each colors corresponds to 10 iterations. **(f)** Frequency-density correlation monotonically increases with training. Grey and black lines indicate correlation for 10 iterations and the average. **(g)** (Left) 25 uniformly distributed fields (left) in a 2D arena with obstacles (grey). Agent navigates to the goal (red) from the start (blue). (Right) Fields along the goal trajectory elongate and overlap while being narrow to avoid obstacles. **(h)** When the number of fields is low, field adaptation facilitates faster policy convergence (blue) compared to when only the policy is optimized (orange).

the first 5 trials, the agent spends a higher frequency of time at the start location given its random policy (blue). As learning progressed, the agent's frequency distribution shifted to the goal (orange) such that by the last 5 trials, the agent spent a higher frequency at the goal, with a slight peak at the start (green). There is no correlation (right, $R = -0.094, P = 0.187$) between the frequency of visit $f(x)$ and the field density $d(x)$ at the start of learning **(Fig. 1E)**, but correlation increased monotonically as learning progressed **(Fig. 1F)** to become significantly positive (left, $R = 0.439, P < 1e^{-5}$), suggesting that field adaptation when maximizing the cumulative discounted reward gradually tiles the trajectory distribution.

Next, we study the hypothesis that the learned place field representation facilitates faster policy learning in a continuous 2D arena with obstacles. The agent has to navigate from one of 4 start locations (blue circles) to the goal (red circle) **(Fig. 1G)**. We evaluate two condition, where we only optimize the synapses W_{ij} from the fields to the policy (policy learning) or we optimize both the field and policy parameters θ (state representation + policy learning). After learning, The field widths (green circle) are stretched along the goal trajectory and are narrow along regions with obstacles while field centers aggregate before the goal (red dots), replicating the field organization in the 1D track. When the number of fields is small **(Fig. 1h)**, policy learning alone (orange) does not converge after 2000 trials but optimizing both the field and policy leads to behavior convergence (blue). As the number of fields increase to 441, policy learning alone converges to a similar navigation performance as when fields are also optimized.

Discussion

We have shown that when maximizing a normative goal such as the cumulative discounted reward, place fields organize with a high density at goals while the rest stretch along the goal trajectory. Importantly, we show that the normative objective optimizes field density to become correlated with the agent's trajectory distribution with training $d(x) \propto f(x)$. Furthermore, we show that adapting field parameters improves policy learning, when the number of fields is low.

The field density optimized through reinforcement learning supports the efficient coding hypothesis proposal which specifies that the optimal field density has to be proportional to the stimulus distribution to improve discriminability (Ganguli & Simoncelli, 2014), which in this case is to find the small goal in a specific location. Based on the learned policy (not shown), we propose that the increase in field width is a mechanism to aggregate continuous information requiring the same action into a single discrete state to speed up policy learning. Hence, we postulate that the skewed place fields (Mehta et al., 2000; Frank et al., 2004) and the successor representation proposal (Stachenfeld et al., 2017) is an emergent outcome when learning goal-directed navigation.

Using this normative navigation model, we also observe representational drift after policy convergence (Qin et al., 2023) and predict that fields near the goal drift at a higher rate. Additionally, maximizing the reward prediction error (Schulman, Moritz, Levine, Jordan, & Abbeel, 2016) instead of the cumulative discounted reward as a normative goal recapitulates the remapping dynamics observed when reward expectancy is low (Krishnan, Heer, Cherian, & Sheffield, 2022). Future works include developing an analytical solution to determine the optimal field density when maximizing the reward objective function and a biologically plausible learning algorithm to adapt the field and policy parameters to model task learning (Kumar, Tan, Libedinsky, Yen, & Tan, 2021).

Acknowledgments

This research was supported in part by grants NSF PHY-1748958 and PHY-2309135 to the Kavli Institute for Theoretical Physics (KITP). MGK and CP is supported by NSF Award DMS-2134157. CP is further supported by NSF CAREER Award IIS-2239780, and a Sloan Research Fellowship. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

Frank, L. M., Stanley, G. B., & Brown, E. N. (2004). Hippocampal plasticity across multiple days of exposure to novel environments. *Journal of Neuroscience*, *24*(35), 7681–7689.

Ganguli, D., & Simoncelli, E. P. (2014). Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural computation*, *26*(10), 2103–2134.

Hok, V., Lenck-Santini, P.-P., Roux, S., Save, E., Muller, R. U., & Poucet, B. (2007). Goal-related activity in hippocampal place cells. *Journal of Neuroscience*, *27*(3), 472–482.

Krishnan, S., Heer, C., Cherian, C., & Sheffield, M. E. (2022). Reward expectation extinction restructures and degrades ca1 spatial maps through loss of a dopaminergic reward proximity signal. *Nature Communications*, *13*(1), 6662.

Kumar, M. G., Tan, C., Libedinsky, C., Yen, S.-C., & Tan, A. Y. (2022). A nonlinear hidden layer enables actor–critic agents to learn multiple paired association navigation. *Cerebral Cortex*, *32*(18), 3917–3936.

Kumar, M. G., Tan, C., Libedinsky, C., Yen, S.-C., & Tan, A. Y.-Y. (2021). One-shot learning of paired association navigation with biologically plausible schemas. *arXiv preprint arXiv:2106.03580*.

Lee, J. S., Briguglio, J. J., Cohen, J. D., Romani, S., & Lee, A. K. (2020). The statistical structure of the hippocampal code for space as a function of time, context, and value. *Cell*, *183*(3), 620–635.

Mehta, M. R., Quirk, M. C., & Wilson, M. A. (2000). Experience-dependent asymmetric shape of hippocampal receptive fields. *Neuron*, *25*(3), 707–715.

Qin, S., Farashahi, S., Lipshutz, D., Sengupta, A. M., Chklovskii, D. B., & Pehlevan, C. (2023). Coordinated drift of receptive fields in hebbian/anti-hebbian network models during noisy representation learning. *Nature Neuroscience*, *26*(2), 339–349.

Schulman, J., Moritz, P., Levine, S., Jordan, M., & Abbeel, P. (2016). High-dimensional continuous control using generalized advantage estimation. *International Conference on Learning Representations*.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature neuroscience*, *20*(11), 1643–1653.

Steele, R., & Morris, R. (1999). Delay-dependent impairment of a matching-to-place task with chronic and intrahippocampal infusion of the nmda-antagonist d-ap5. *Hippocampus*, *9*(2), 118–136.

Sutton, R. S., McAllester, D., Singh, S., & Mansour, Y. (1999). Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, *12*.