UNSUPERVISED EMERGENCE OF SPATIAL STRUCTURE FROM SENSORIMOTOR PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite its omnipresence in robotics application, the nature of spatial knowledge and the mechanisms that underlie its emergence in autonomous agents are still poorly understood. Recent theoretical work suggests that the concept of space can be grounded by capturing invariants that space's structure induces in an agent's raw sensorimotor experience. Moreover, it is hypothesized that capturing these invariants is beneficial for a naive agent trying to predict its sensorimotor experience. Under certain exploratory conditions, spatial representations should thus emerge as a byproduct of learning to predict. We propose a simple sensorimotor predictive scheme, apply it to different agents and types of exploration, and evaluate the pertinence of this hypothesis. We show that a naive agent can capture the topology and metric regularity of its spatial configuration without any a priori knowledge, nor extraneous supervision.

1 INTRODUCTION

Space appears to be a pervasive concept in our perception of the world, and as such plays a central role in most artificial perception systems, in particular in computer vision and robotics applications. Yet its fundamental nature and the mechanisms that could lead to its emergence in an artificial system still remain poorly understood (Kant, 1998; Poincaré, 1895; Nicod, 1924). In most cases, the problem is circumvented by implementing prior knowledge in the system regarding the structure of space, and how motor and sensory information convey spatial properties (for instance through a kinematics model (Siciliano & Khatib, 2016), or a sensor model (Cadena et al., 2016)). In more recent years, and with the developments of machine learning techniques, approaches with less hand-engineered priors developed to solve spatial tasks ((Kahn et al., 2017; Quillen et al., 2018; Levine et al., 2018; Smolyanskiy et al., 2017) to name a few). However they tend to rule out the specificity of spatial experiences in favor of a global assessment of the agent's performance, leaving the question of the origin and structure of spatial knowledge largely open.

Nevertheless recent work presented in (Laflaquière et al., 2018), directly addresses the grounding of spatial knowledge, and gives a theoretical perspective on how it could emerge in an unsupervised way. The approache takes inspiration from philosophical ideas formulated more than a century ago by H.Poincaré (Poincaré, 1895), and from the more recent Sensori-Motor Contingencies Theory (SMCT) which puts forward an unsupervised sensorimotor grounding of perception (O'Regan & Noë, 2001). It describes how space naturally induces specific invariants in any situated agent's sensorimotor experience, and how these invariants can be autonomously captured to improve the compactness of internal representations and the counter-factual prediction of sensory experiences.

In this paper, we propose a practical evaluation of the theoretical claims put forward in (Laflaquière et al., 2018). We introduce a simple self-supervised learning scheme in which an agent learns to predict its sensory experience given its motor state. We analyze the encoding of the motor experience produced this way, and in particular how it is impacted by different types of exploration of the environment, and how it relates to the original theory. More precisely, we show that by experiencing consistent sensorimotor transitions in an environment which can move, a naive agent can learn to encode its motor state in such a way as to capture the topological structure and metric regularity of its external position without the need for additional a priori knowledge.

2 RELATED WORK

The problem of building spatial representations of some sort is addressed in a multitude of machine learning papers. Due to space constraints, we cannot give a complete overview, but will only highlight the most relevant to this work.

The emergence of space perception is often conceptualized as a grid cells emergence problem, like in the recent (Cueva & Wei, 2018; Banino et al., 2018). These approaches however rely on extraneous spatial supervision signals, which counters any claim of autonomy. Other work like (Weiller et al., 2010; Jonschkowski & Brock, 2015) build spatial-like internal representation by instead focusing more on the interaction with the environment, but still rely on priors designed to favor the emergence of such representations. Although less explicitly focused on spatial representations, work like (Watter et al., 2015; Thomas et al., 2017) build representations that support the use as a state signal in motor control tasks, which often turn out to correspond to spatial displacements. The controllability constraint however relies on priors which shape the representation.

A large range of other machine learning applications aim at solving intrinsically spatial tasks without explicitly studying or constraining any intermediary spatial representation. For instance, (Eslami et al., 2018) recently presented a network able to generate perspective-conditioned images of 3D scenes. The approach however relies on an explicit spatial perspective input, and no study regarding how the network incorporates this spatial information has been done. The influential work on Atari playing agents (Mnih et al., 2013) is another example of a system which learns to solve spatial tasks without explicitly forming any spatial representation. Some intermediary Reinforcement Learning approaches like (Mirowski et al., 2016) solve spatial tasks in an end-to-end fashion, but still incite spatial consistency in the system via specifically designed auxiliary tasks.

An unsupervised sensorimotor predictive scheme has been proposed in (Ortiz & Laflaquière, 2018) to build representations of sequences of displacements ; a concept which is close to the one of position developed in this paper. Finally, the question of the acquisition of spatial knowledge has also been addressed more fundamentally from the perspective of the SMCT theory in (Laflaquiere et al., 2012; Laflaquière et al., 2015; Terekhov & O'Regan, 2016).

3 PROBLEM SETUP

We borrow the mathematical formalism from (Laflaquière et al., 2018), where a thorough description of the problem of spatial knowledge acquisition is provided. Here we only introduce the core ideas and hypotheses that are relevant for this work. The reader is invited to refer to (Laflaquière et al., 2018) for a more extensive description of the theory. We assume that an agent's sensory and motor states are respectively defined by vectors $\mathbf{s} \in \mathbb{R}^{N_s}$ and $\mathbf{m} \in \mathbb{R}^{N_m}$. No additional assumption is made regarding the way this information is encoded. We denote ϕ_ϵ the unknown mapping induced by the environment between $\mathbf{m}$ and $\mathbf{s}$, such that $\mathbf{s} = \phi_\epsilon(\mathbf{m})$, where ϵ denotes the state of the environment. The mapping ϕ_ϵ can be seen as describing how "the world" transforms changes in motor state to changes in the sensory state. It encapsulates all the unknown properties of the environment which affect the agent's sensorimotor experience, and in particular the structure of space. This structure induces specific sensorimotor invariants.

A first kind of invariant concerns the topology of the spatial configuration of an agent's sensor. Assuming a continuous sensorimotor mapping ϕ_ϵ in a rich enough environment, the topology of the sensory manifold generated by exploring the motor space is identical to the topology of the sensor's spatial configuration in the environment (Laflaquiere et al., 2013). Intuitively this means that small motor changes correspond to small displacements of the sensor, which themselves generate small sensory changes. This property is invariant to the environmental state ϵ. This means that the topology of the sensor's spatial configuration is accessible to the agent via the sensorimotor flow. In particular, in the case of a redundant motor system, the agent can discover that multiple motor configurations always lead to identical sensory states, and can then represent them with a single internal state, effectively reducing the dimension of the original motor space.

A second kind of invariant concerns the metric regularity of the sensor's spatial configuration. Assuming that the environment can move relative to the agent, the same sensory variations associated with a displacement of the sensor between two parts of the environment can be associated with multiple motor variations, depending on the position of the environment relative to the agent. The

sensory variations are thus invariants to the corresponding set of motor transitions. Intuitively this means that the motor states associated with sensing, for instance, the top and bottom of a bottle change depending on where the bottle is relative to the agent. However, these motor transitions might vary in magnitude depending on where the starting state lies in the motor space (for instance moving a sensor that is attached to a kinematic chain 10 cm up might require a different motor command when it is right in front of the agent, or when it is on its right). It is thus possible for the agent to regularize its internal metric to account for these observed sensory invariants, this way capturing the regularity of the external spatial (Euclidean) metric. Note that following the theoretical developments of (Laflaquière et al., 2018), only translations of the environment are considered in this work. This suggests that the topology and metric regularity of space can be described as sensorimotor invariants. These invariants should be accessible to a naive agent exploring its environment under certain conditions. First, to experience topological invariants, the agent should experience *consistent* sensorimotor transitions such that the state of the environment does not change during the transition. This ensures that the sensory change experienced during the transition is consistent with the motor change produced. Second, to experience metric invariants, the agent should experience *displacements of the environment* in-between some sensorimotor transitions. This ensures that the agent has the opportunity to experience the same sensory changes from different starting motor states. Moreover, it has been hypothesized in (Laflaquière et al., 2018) that capturing such sensorimotor invariants should be beneficial to a naive agent which learns to predict its own sensorimotor experience. In other words, building a representation of the motor states which complies with the topological and metric invariants should act as a prior over sensorimotor transitions not yet experienced, improving this way the predictive capacity of the learning system. As a consequence, a representation capturing the topology and metric regularity of space should naturally emerge in a sensorimotor predictive system. This includes any representation related to the ground truth position of the agent via an arbitrary non-singular affine transformation which preserves ratio of distances. It is this emergence hypothesis that we test in the following sections.

4 EXPERIMENTS

4.1 SENSORIMOTOR PREDICTIVE NETWORK ARCHITECTURE

We propose a simple sensorimotor predictive architecture to test the validity of the hypotheses laid out in Sec. 3. The architecture is made of two types of modules: i) Net_{enc}, a Multi-Layer Perceptron (MLP) taking a motor state $\mathbf{m}_t$ as input and outputting a motor representation $\mathbf{h}_t$ of dimension dim_H, and ii) Net_{pred}, a MLP taking as input the concatenation of a current representation $\mathbf{h}_t$, a future representation $\mathbf{h}_{t+1}$, and a current sensory state $\mathbf{s}_t$, and outputting a prediction for the future sensory state $\tilde{\mathbf{s}}_{t+1}$. As illustrated in Fig. 1, the overall network architecture connects a predictive module Net_{pred} to two siamese copies of a Net_{enc} module, ensuring that both motor states $\mathbf{m}_t$ and $\mathbf{m}_{t+1}$ are consistently encoded using the same mapping. The self-supervised sensorimotor predictive task of the network consists in minimizing the Mean Square Error between the prediction $\tilde{\mathbf{s}}_{t+1}$ and the ground truth $\mathbf{s}_{t+1}$. No explicit component is added to the loss regarding the representation $\mathbf{h}$. Unless stated otherwise, the dimension dim_H of the representation is arbitrarily set to 3 for the sake of visualization. A thorough description of the network is proposed in Appendix A, alongside a description of the training procedure.

4.2 TESTING HYPOTHESIS

We laid out in Sec. 3 hypotheses about the impact of different exploration strategies on the internal representation of motor states. To confirm or disprove them, we introduce a measure of how well $\mathbf{h}$ approximates the ground truth spatial configuration of the agent, denoted $\mathbf{p}$. Unfortunately the two cannot be compared directly, as there might exists an affine transformation between the two (see Sec.3). Yet, for a given set of representations $H = \{\mathbf{h}_i\}$ and their ground-truth counterparts $P = \{\mathbf{p}_i\}$, one can perform a linear regression to estimate and compensate for the affine transformation between P and H (and similarly between H and P). Distances between pairs of points in P and H after this affine compensation can be compared to evaluate how much the structure of the two sets differ. We thus define two measures of dissimilarity between the representations of motor states and

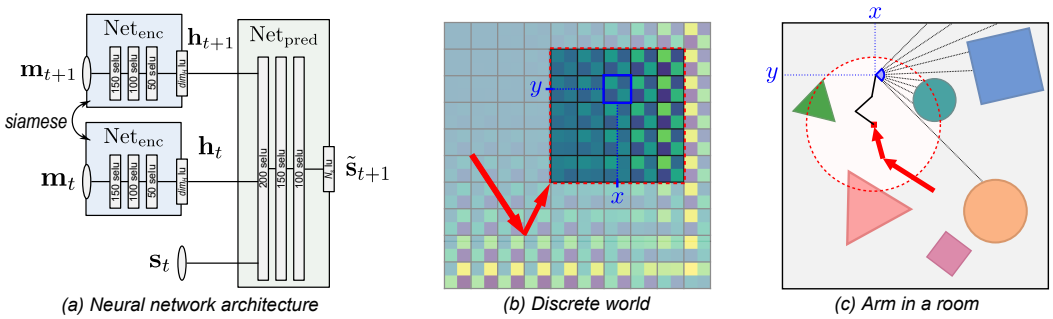

Figure 1: (a) The neural network architecture featuring two siamese instances of the Net$_{enc}$ module and the Net$_{pred}$ module. (b-c) Illustrations of the discrete world and arm in a room simulations. The part of the environment accessible to the agent for the current translation of the environment is framed in dotted red. The agent's current position $[x, y]$ in the section frame is displayed in blue. Red arrows illustrate the type of translation the environment can undergo (equivalent to a displacement of the agent's base in a static environment). (Best seen in color)

the ground-truth positions as:

$$Q_p = \frac{1}{N^2} \sum_i^N \sum_j^N \frac{|D_{ij}^{H\rfloor_P} - D_{ij}^P|}{\max_{kl}(D_{kl}^P)} \ , \ Q_h = \frac{1}{N^2} \sum_i^N \sum_j^N \frac{|D_{ij}^{P\rfloor_H} - D_{ij}^H|}{\max_{kl}(D_{kl}^H)}, \tag{1}$$

where N is the number of samples, D_{ij}^X denotes the distance between samples i and j in the set X, $H\rfloor_P$ denotes the set H after projection in the space of P by affine compensation, and $P\rfloor_H$ the set P after a similar projection in the space of H. The errors are normalized by the maximal distance in the projection space in order to avoid undesired scaling effects with scaling of the representation. In order to rigorously compare dissimilarity measures between experiments, we always generate evaluation data points in H and P by sampling the agent's motor space in a fixed and regular fashion.

In order to test our hypotheses, we consider 3 types of exploration of the environment by the agent:

- **Inconsistent transitions in a moving environment**: A baseline scenario in which the spatiotemporality of sensorimotor experiences is not respected, making the data akin to what a typical passive and non-situated (not in a continuous interaction with the world) agent would get. Pairs of states $(\mathbf{m}_t, \mathbf{s}_t)$ and $(\mathbf{m}_{t+1}, \mathbf{s}_{t+1})$ are randomly generated by exploring a single environment which translates randomly between t and $t+1$. In the following we refer to it as **MTM** (Motor-Translation-Motor) for the sake of compactness.
- **Consistent transitions in a static environment**: A first situated scenario in which the agent can experience the spatiotemporal consistency of its sensorimotor transitions. Pairs of states $(\mathbf{m}_t, \mathbf{s}_t)$ and $(\mathbf{m}_{t+1}, \mathbf{s}_{t+1})$ are randomly generated by exploring a single static environment. In the following we refer to it as **MM** (Motor-Motor) for the sake of compactness.
- **Consistent transitions in a moving environment**: A second situated scenario in which the agent can experience consistent transitions *and* displacements of the environment. Pairs of states $(\mathbf{m}_t, \mathbf{s}_t)$ and $(\mathbf{m}_{t+1}, \mathbf{s}_{t+1})$ are randomly generated by exploring a single environment which can translate randomly after each transition. In the following we refer to it as **MMT** (Motor-Motor-Translation) for the sake of compactness.

These exploration scenarios are designed to evaluate the impact of consistent sensorimotor transitions and displacements of the environment on the structure of the motor representation built by the agent (see Sec. 5). More precisely, the MTM exploration should induce no spatial sensorimotor invariants, the MM exploration should induce topological invariants, and the MMT exploration should induce both topological and metric invariants.

4.3 Agent-Environment Setups

We propose two different simulated agent-environment setups to test our hypotheses:

- *Discrete world*: An artificial setup designed to provide a simple and fully controlled sensorimotor interaction. As illustrated in Fig. 1, the environment consists in a grid world of size 10×10, in

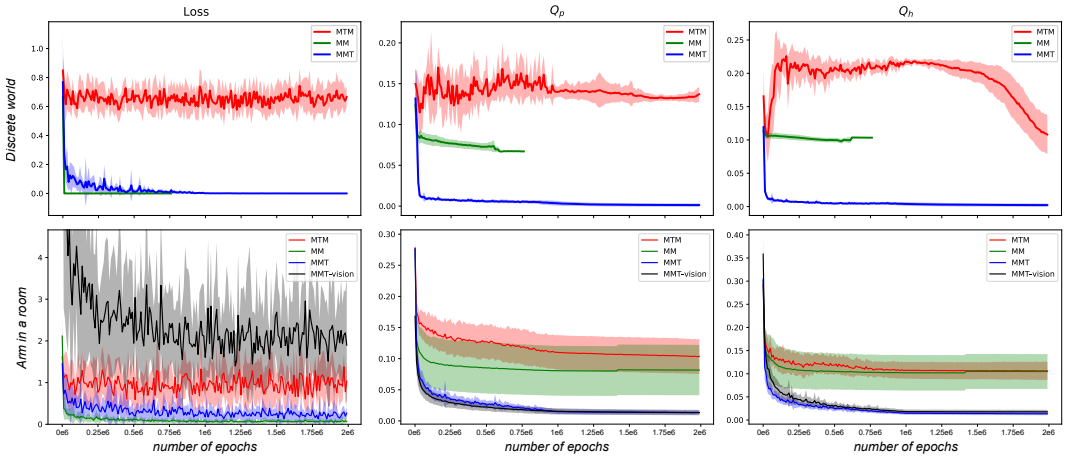

Figure 2: Evolution of the loss and the dissimilarity measures Q_p and Q_h during training for both setups, and for the three types of exploration. The displayed means and standard deviations are computed over 10 independent trials. (Best seen in color)

which each cell is associated with a sensory state $\mathbf{s}$ of dimension $N_s = 4$. Each of these 4 sensory components changes continuously as a random order 3 polynomial function of the position in the grid, in order to approximate a continuous sensory experience. The agent can freely explore a section of the environment of size 5×5. It can generate a 3D motor state $\mathbf{m} = [m_1, m_2, m_3]$ to reach each 2D position $\mathbf{p} = [x, y]$ in this sub-grid, and to receive the corresponding sensory input $\mathbf{s}$. The mapping from motor state to position in the grid is arbitrarily defined as: $\mathbf{p} = [x, y] = [\sqrt{m_1}, \sqrt{m_2}]$. It is purposefully non-linear, and presents a superfluous motor command m_3, which will be useful to study how the representation might capture the topology and metric of $\mathbf{p}$. Because of this non-linearity and the discrete nature of the positions $\mathbf{p}$, the agent can only sample its motor space in a non-linear fashion (see Fig. 3). The environment can translate in 2D, effectively changing the section that the agent can explore. Finally, the environment is setup to act as a torus to avoid border effects; which means that the agent appears on the other side of the environment when exploring beyond its limits.

- *Arm in a room*: A more complex scenario of a simple arm exploring a room, implemented using the Flatland simulator (Caselles-Dupré et al., 2018). As illustrated in Fig. 1, the environment is a 2D square room of width 12 units with walls, randomly filled with simple geometric objects (squares, circles, triangles) with random properties (number, size, position, color). The agent is a three-segment arm, each of length 1 unit, with one motor at each joint to control the relative orientation of the segment in $[-\pi, \pi]$ radians. The arm's tip is equipped with an array of 10 distance sensors. The orientation of the sensor in the world is fixed, as we only consider the experience of sensor and environment translations in this work (see Sec. 3 and (Laflaquière et al., 2018)). The environment can translate relatively to the arm's base, and the agent's sensor cannot explore beyond the walls.

5 Results

We evaluate the two experimental setups on the three types of exploration. Each simulation is run 10 times, with all random parameters drawn independently on each trial. During training, the dissimilarity measures Q_p and Q_h are evaluated on a fixed regular sampling of the motor space, so that they can be compared consistently between epochs. The evolution of the loss, Q_p, and Q_h during training is displayed in Fig. 2. Additionally Fig. 3 shows the final representation of the same regular motor sampling, for one randomly selected trial of each simulation. The corresponding ground truth positions, as well as the affine compensations from the representation space to the position space, and vice versa, are also displayed. Below we present a qualitative analysis of the results and interpret them with regard to the initial hypotheses laid out in Sec.3. A more quantitative and detailed analysis of the results is provided in Appendix B.

5.1 MTM EXPLORATION

We first analyze the impact of the MTM type of exploration on the motor state representation built by the agent in both experiment setups. We can see in Fig. 2 that the loss, and the measures Q_p and Q_h stay at relatively high values throughout the training, for both the discrete world and arm in a room simulations. Such a high loss indicates that the agent is unable to learn a good sensorimotor predictive mapping. This is expected as the network is fed with inconsistent sensorimotor transitions, while the environment moves between sensorimotor states. Therefore the current pair $(\mathbf{m}_t, \mathbf{s}_t)$ is not informative to predict $(\mathbf{s}_{t+1})$. Similarly, the high final values of Q_p and Q_h seem to indicate that the structure of the representation $\mathbf{h}$ significantly differs from the ground truth position $\mathbf{p}$ of the agent's sensor, even after affine compensations. This is confirmed in Fig. 3 in which we can see that the structure displayed by the representation does not match the one of the ground truth position metric-wise or even topology-wise. In particular, redundant motor states corresponding to the same spatial configuration of the sensor are represented differently.
When experiencing inconsistent sensorimotor transitions in a moving environment, the representation of the motor state built by the predictive network thus converges to an arbitrary structure which does not capture topological or metric properties of the spatial configuration of the sensor. It is due to the fact that the naive agent has no opportunity to experience the sensorimotor invariants that a consistent spatial exploration of the world would induce in its sensorimotor experience.

5.2 MM EXPLORATION

We now analyze the impact of the MM type of exploration on the motor state representation built by the agent in both experiment setups. For both setups, we can see in Fig. 2 that the loss quickly drops to very small values, while the values of Q_p and Q_h rapidly decrease before stabilizing around high but significantly smaller values than in the MTM exploration. This sharp drop of the loss indicates that the network is able to learn an accurate sensorimotor predictive mapping. This is expected as the agent consistently explores an environment which always stays in the same position. It can thus easily map each motor state $\mathbf{m}_{t+1}$ to its corresponding sensory state $\mathbf{s}_{t+1}$. On the other hand, the still high but significantly lower values of Q_p and Q_h seem to indicate that the motor representation built by the network displays a structure which is more similar to the one of the ground truth sensor position. In Fig. 3, the corresponding plots show that, after affine compensation, the representation indeed displays the same topology as the position (best seen in the ground-truth position space). In particular, redundant motor states corresponding to the same spatial configuration of the sensor are represented by the same state in the representational space, this way effectively reducing the dimension of the motor manifold (3D) to match the dimension of the position manifold (2D). However, we can also see that the representation manifold is not necessarily flat, and thus does not match perfectly the flat affine projection of the positions in the representational space. This phenomenon is discussed in more details in Appendix B, but leads to Q_h being less useful than Q_p to assess the quality of the representation. This is particularly visible in Fig. 2 where Q_h evolves towards the same value at the end of the training in the MTM and MM cases. It shows that Q_h does not capture the structural difference induced by the two types of exploration that we observe in the representation, whereas Q_p does.
When experiencing consistent sensorimotor transitions in a static environment, the representation of the motor state built by the predictive network thus captures the topological structure of the spatial configuration of the sensor. This is due to the fact that in this static environment the agent can experience the topological properties of the sensory manifold which inform about the topology of the sensor position itself. In particular, the agent discovers that some redundant motor states produce the same sensory output, and the network converges to a representational structure which represent them with a single point $\mathbf{h}$. Similarly, the agent discovers that small sensory variations are due to small motor variations, and the network preserves this topological property in the representational space. The experience of consistent sensorimotor transitions thus greatly impacts the motor state representation.

5.3 MMT EXPLORATION

Finally we analyze the impact of the MMT type of exploration on the motor state representation built by the agent in both experiment setups. For both setups, we can see in Fig. 2 that the loss, and the measures Q_p and Q_h decrease and stabilize around very small values. This sharp drop of the loss

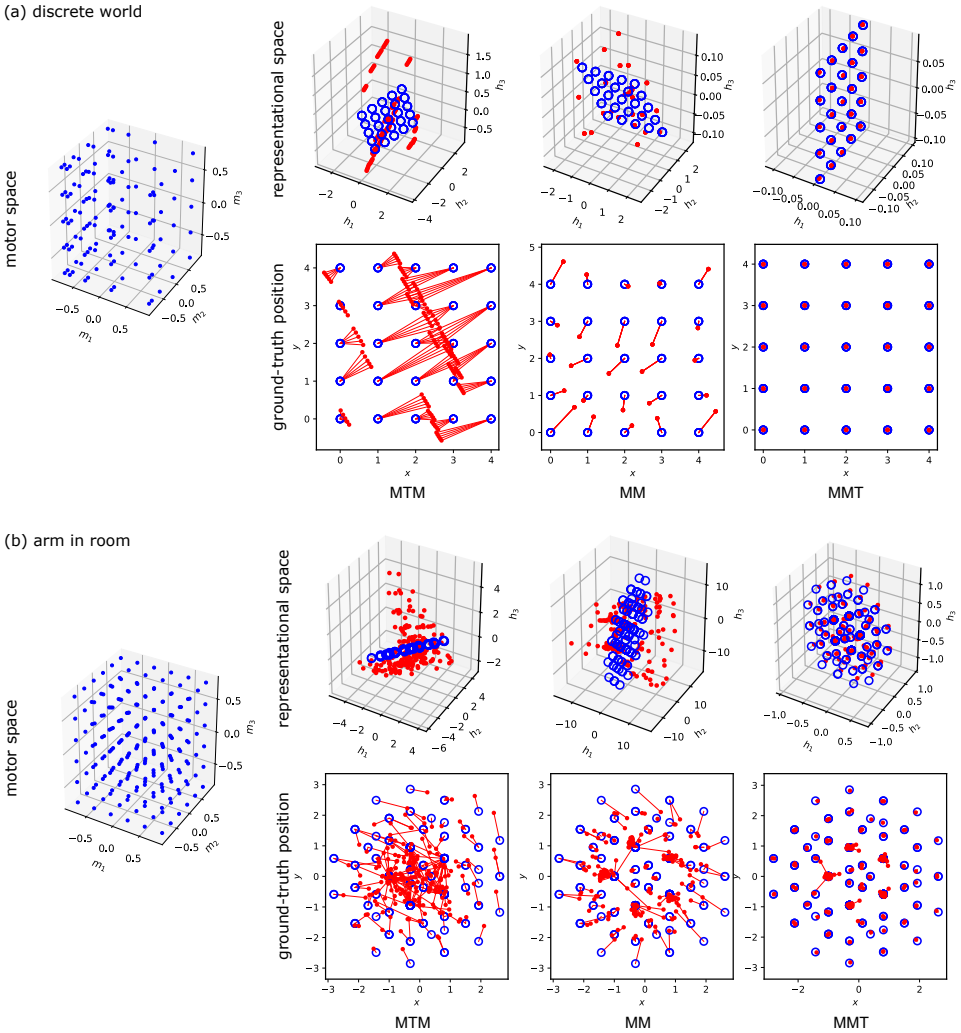

Figure 3: Visualization of the (normalized) regular motor space sampling, its encoding in the representational space (red dots), and the corresponding ground-truth positions (blue circles) for the three types of exploration, for the discrete world setup (a) and the arm in a room setup (b). Both the representations and the positions are also projected into each other's space via an optimized affine transformation. Lines have been added in the ground-truth position space to visualize the distance between the projection of the representations and their ground-truth counterparts. They are the distances considered when computing Q_p. The equivalent is true in the representation space to compute Q_h (not displayed). (Best seen in color)

indicates once again that the network is able to learn an accurate sensorimotor predictive mapping. This is expected as the agent experiences consistent sensorimotor transitions and the environment moves only between transitions. Knowing $\mathbf{m}_t$, $\mathbf{s}_t$, and $\mathbf{m}_{t+1}$, it is thus possible to predict the next sensory state $\mathbf{s}_{t+1}$. Moreover the small final values of the dissimilarity measures indicate that the motor representation built by the network displays a structure which is very similar to the one of the ground truth sensor position. In Fig. 3, the corresponding plots indeed show that the representation of the motor space almost perfectly matches the ground truth position after affine compensation (best seen in the ground-truth position space), for both setups. It displays the same topology and metric regularity as the position, which shows that there exists a simple affine transformation between the motor state representation built by the network and the position of the sensor. Contrarily to the representation built in the MM type of exploration, the manifold in the representational space even tends towards a flat surface. This phenomenon is discussed in Appendix B.

When experiencing consistent sensorimotor transitions and movements of the environment, the rep-

resentation of the motor state built by the predictive network thus captures the topological structure and metric regularity of the spatial configuration of the sensor. This is due to the fact that the agent can experience both the topological properties of the sensory manifold, and the sensory invariants corresponding to different motor transitions generating the same sensory transition for different positions of the environment (see Sec. 3). The experience of consistent sensorimotor transitions in a moving environment thus greatly impacts the motor state representation built by the predictive network.

In order to evaluate the robustness of this result with regard to the sensorimotor mapping complexity and the dimension of the representational space, an additional experiment is performed in which the arm is equipped with a RGB camera of resolution 16 pixels and dim_H is set to 25. The results, displayed in black in Fig. 2 are qualitatively identical to the initial setup. This indicates that the results regarding the motor state representation built by the predictive network are insensitive to the complexity of the sensory input and to the dimensionality of the representational space. These results are discussed at length in Appendix B.3.

6 Conclusion

We have proposed in this work a practical evaluation of the hypotheses put forward in (Laflaquière et al., 2018) regarding the unsupervised grounding of spatial topology and metric regularity. We proposed a simple neural network architecture, including a module to encode an agent's motor state into an internal representation, which learns to solve a self-supervised task of sensorimotor prediction. We studied the impact of different data collection strategies on the internal representation of the motor states, ranging from spatiotemporally discontinuous sensorimotor pairs, to consistent sensorimotor transitions in a moving environment. As previously hypothesized, we showed that the experience of consistent sensorimotor transitions in an environment leads a naive agent to build an internal representation of its motor states, which captures the topology of its sensor's spatial configuration. This is because such a type of exploration induces sensorimotor invariants which are captured by the learning system (in this case a neural network) during training. They drive the system to bring closer representations of motor states which lead to close sensory predictions. Similarly we showed that the experience of a moving environment further regularizes the internal representation such as to capture the metric regularity of the sensor's spatial configuration. This is because such a type of exploration induces additional sensorimotor invariants which are captured by the learning system. They drive it to represent with the same vector in the representational space all motor transitions which can lead to the same sensory transition (depending on the position of the environment).

Importantly these properties of the representation emerge in an unsupervised way, by virtue of the benefit that capturing these invariants represents for the predictive model. Note however that, in theory, the predictive module Net_{pred} should be able to learn an appropriate mapping regardless of the structure of the motor representation (assuming no loss of information). Understanding the mechanisms which favor a representation capturing sensorimotor invariants and a less complex predictive mapping during training is thus an interesting question to investigate. Nonetheless, the results presented in this work suggest that no particular prior about space, nor about the properties of its sensorimotor apparatus is required for an agent to acquire spatial knowledge. It can be achieved by having the agent autonomously explore its environment, and capture invariants that space induces in its sensorimotor experience. This approach also suggests that action is essential in such an endeavor. It is of course necessary to explore the environment, but more fundamentally to discover spatial invariants. The results presented in this work would indeed be impossible to achieved without priors and by only considering the sensory flow. The spatial-like representations even originate from the motor space itself, whose internal structure is reshaped by sensorimotor experiences. Indirectly, this approach thus casts some doubts on purely passive and observational approaches of the problem of spatial knowledge acquisition.

Yet, many problems need to be addressed before such an approach can be applied to more complex systems. So far, only translations of the sensor and the environment have been considered, and the impact of adding rotations in the exploration still has to be studied. Understanding how multiple spatial representations associated with multiple independent sensors could be merge to form a global representation of the agent's configuration in space also requires additional theoretical work. Similarly, the approach needs to be extended to characterize the environment's spatial configuration too, especially if it contains independent objects. Finally, the impact of exploring a succession of different environments on the spatial representation also has to be investigated.

REFERENCES

Andrea Banino, Caswell Barry, Benigno Uria, Charles Blundell, Timothy Lillicrap, Piotr Mirowski, Alexander Pritzel, Martin J Chadwick, Thomas Degris, Joseph Modayil, et al. Vector-based navigation using grid-like representations in artificial agents. *Nature*, 557(7705):429, 2018.

Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on Robotics*, 32(6):1309–1332, 2016.

Hugo Caselles-Dupré, Louis Annabi, Oksana Hagen, Michael Garcia-Ortiz, and David Filliat. Flatland: a lightweight first-person 2-d environment for reinforcement learning. *arXiv preprint arXiv:1809.00510*, 2018.

Christopher J Cueva and Xue-Xin Wei. Emergence of grid-like representations by training recurrent neural networks to perform spatial localization. *arXiv preprint arXiv:1803.07770*, 2018.

SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 315–323, 2011.

Rico Jonschkowski and Oliver Brock. Learning state representations with robotic priors. *Autonomous Robots*, 39(3):407–428, 2015.

Gregory Kahn, Adam Villaflor, Bosen Ding, Pieter Abbeel, and Sergey Levine. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. *arXiv preprint arXiv:1709.10489*, 2017.

Immanuel Kant. *Critique of pure reason*. Cambridge University Press, 1998.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, pp. 971–980, 2017.

Alban Laflaquiere, Sylvain Argentieri, Olivia Breysse, Stéphane Genet, and Bruno Gas. A non-linear approach to space dimension perception by a naive agent. In *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*, pp. 3253–3259. IEEE, 2012.

Alban Laflaquiere, Alexander V Terekhov, Bruno Gas, and J Kevin O'Regan. Learning an internal representation of the end-effector configuration space. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pp. 1230–1235. IEEE, 2013.

Alban Laflaquière, J Kevin O'Regan, Sylvain Argentieri, Bruno Gas, and Alexander V Terekhov. Learning agent's spatial configuration from sensorimotor invariants. *Robotics and Autonomous Systems*, 71:49–59, 2015.

Alban Laflaquière, J Kevin O'Regan, Bruno Gas, and Alexander Terekhov. Discovering space-grounding spatial topology and metric regularity in a naive agent's sensorimotor experience. *Neural Networks*, 2018.

Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018.

Piotr Mirowski, Razvan Pascanu, Fabio Viola, Hubert Soyer, Andrew J Ballard, Andrea Banino, Misha Denil, Ross Goroshin, Laurent Sifre, Koray Kavukcuoglu, et al. Learning to navigate in complex environments. *arXiv preprint arXiv:1611.03673*, 2016.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Jean Nicod. *La géométrie dans le monde sensible*. Presses universitaires de France, 1924.

J Kevin O'Regan and Alva Noë. A sensorimotor account of vision and visual consciousness. *Behavioral and brain sciences*, 24(5):939–973, 2001.

Michael Garcia Ortiz and Alban Laflaquière. Learning representations of spatial displacement through sensorimotor prediction. *arXiv preprint arXiv:1805.06250*, 2018.

Henri Poincaré. L'espace et la géométrie. *Revue de métaphysique et de morale*, 3(6):631–646, 1895.

Deirdre Quillen, Eric Jang, Ofir Nachum, Chelsea Finn, Julian Ibarz, and Sergey Levine. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. *arXiv preprint arXiv:1802.10264*, 2018.

Bruno Siciliano and Oussama Khatib. *Springer handbook of robotics*. Springer, 2016.

Nikolai Smolyanskiy, Alexey Kamenev, Jeffrey Smith, and Stan Birchfield. Toward low-flying autonomous mav trail navigation using deep neural networks for environmental awareness. *arXiv preprint arXiv:1705.02550*, 2017.

Alexander V Terekhov and J Kevin O'Regan. Space as an invention of active agents. *Frontiers in Robotics and AI*, 3:4, 2016.

Valentin Thomas, Jules Pondard, Emmanuel Bengio, Marc Sarfati, Philippe Beaudoin, Marie-Jean Meurs, Joelle Pineau, Doina Precup, and Yoshua Bengio. Independently controllable features. *arXiv preprint arXiv:1708.01289*, 2017.

Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, pp. 2746–2754, 2015.

Daniel Weiller, Robert Märtin, Sven Dähne, Andreas K Engel, and Peter König. Involving motor capabilities in the formation of sensory space representations. *PloS one*, 5(4):e10377, 2010.

A NEURAL NETWORK ARCHITECTURE AND TRAINING

The sensorimotor predictive architecture is composed of two types two types of modules: Net_{enc} and Net_{pref}. The first type of module, Net_{enc}, projects a motor state $\mathbf{m}_t$ onto a representation $\mathbf{h}_t$ in a representational space of dimension dim_H. It is a fully connected MLP with three hidden layers of size $(150, 100, 50)$ neurons with SeLu activation functions (Klambauer et al., 2017). A final output of size dim_H neurons corresponding to the representational space is connected to the last hidden layer and has linear activation functions. The second type of module, Net_{pred}, takes as input the concatenation $(\mathbf{h}_t, \mathbf{h}_{t+1}, \mathbf{s}_t)$ of a current motor state representation, a future motor state representation, and a current sensory state, and outputs a prediction $\tilde{\mathbf{s}}_{t+1}$ for the future sensory state. It is a fully connected MLP with three hidden layers of size $(200, 150, 100)$ neurons with SeLu activation functions. A final output of size N_s neurons corresponding to the sensory prediction is connected to the last hidden layer and has linear activation functions. As illustrated in Fig. 1, the overall network architecture connects a predictive module Net_{pred} to two siamese copies of a Net_{enc} module, ensuring that both motor states $\mathbf{m}_t$ and $\mathbf{m}_{t+1}$ are consistently encoded using the same mapping.

We define a simple self-supervised sensorimotor predictive objective for the network. The loss function to minimize is defined as the Mean Square Error between the prediction $\tilde{\mathbf{s}}_{t+1}$ and the ground truth $\mathbf{s}_{t+1}$. No particular component is added to the loss function regarding the representation $\mathbf{h}$ built by the agent. We use an ADAM optimizer to minimize the loss, with a learning rate linearly decreasing from 10^{-3} to 10^{-5} in 10^{-6} epochs, and a batch size of 100 sensorimotor transitions (Kingma & Ba, 2014). The optimization is stopped after $2e^6$ epochs, or when the training loss falls under a threshold set to $1e^{-8}$ on a batch.

Training data are generated by having the simulated agent explore its environment and collect sensorimotor pairs $(\mathbf{m}_t, \mathbf{s}_t)$. A total of $3e^6$ sensorimotor pairs are collected for each simulation by randomly sampling (uniform distribution) motor states in the agent's motor space and getting the corresponding sensory input. If the motor state corresponds to an impossible spatial configuration of the sensor (sensor lying outside of the environment, or inside an object) no sensory input is received and the experience is discarded. The environment can randomly translate (uniform distribution) during the data collection, with a maximal amplitude equal to the size of the environment. In the MTM type of exploration, the environment translates after each collected pair $(\mathbf{m}_t, \mathbf{s}_t)$. In the MM type of exploration, the environment never translates. Finally, in the MMT type of exploration, the environment translates each time 100 pairs $(\mathbf{m}_t, \mathbf{s}_t)$ have been collected.

After the sensorimotor data has been collected, it is normalized such that each component of the motor and sensory states is in $[-0.8, 0.8]$. The network optimization is then done using a batch size of 100 sensorimotor transitions $(\mathbf{m}_t, \mathbf{s}_t) \rightarrow (\mathbf{m}_{t+1}, \mathbf{s}_{t+1})$. The transitions are generated by independently selecting the two sensorimotor pairs $(\mathbf{m}_t, \mathbf{s}_t)$ and $(\mathbf{m}_{t+1}, \mathbf{s}_{t+1})$ in the whole database, except in the MMT type of exploration for which $(\mathbf{m}_t, \mathbf{s}_t)$ is randomly selected in the database but $(\mathbf{m}_{t+1}, \mathbf{s}_{t+1})$ is randomly selected only in the group of 100 sensorimotor pairs which were generated with the same translation of the environment as for $(\mathbf{m}_t, \mathbf{s}_t)$. This ensures that the sensorimotor transitions fed to the network are consistent.

Note that the overall predictive architecture and training procedure have been kept simple. No particular heuristics have been added to improve convergence, generalization, or any other property of the network such as sparsity. Moreover, no testing database has been generated to evaluate the network generalization, although the number of sensorimotor pairs in the database has been purposefully made large so that the probability of experiencing the same sensorimotor transition twice during training is insignificant (of the order of $1e^{-5}$). This approximates a real continuous exploration of the environment with no redundant experiences, and means that the training loss is constantly evaluated on yet unseen data at each epoch. Similarly, the architecture's meta-parameters have not been optimized beyond simply checking that the network was expressive enough to learn the expected mappings. This is because this work does not aim at designing an optimized architecture to solve a task, but rather at studying how sensorimotor invariants induced by a spatial exploration can lead a neural network to build a spatial representation without the need for additional priors.

B DETAILED ANALYSIS OF THE RESULTS

In addition to the qualitative description of the results proposed in Sec. 5, we develop below a more thorough analysis of each curve. As a reminder, each simulation is run 10 times, with all random parameters drawn independently on each trial. The evolution of the mean and standard deviation of the loss and of the measures Q_p and Q_h evaluated on a fixed regular sampling of the motor space is displayed in Fig. 2. Additionally Fig. 3 shows the final representation of the same fixed motor sampling, for one randomly selected trial of each simulation. The corresponding ground truth positions, as well as the affine compensations from the representation space to the position space, and vice versa, are also displayed.

B.1 DISCRETE WORLD

We first analyze the impact of the exploration on the representation in the discrete world setup.

MTM exploration: We can see in Fig. 2 that the loss rapidly drops, due to the network finding an adequate average value and scale for $\tilde{\mathbf{s}}_{t+1}$, but then stagnates around a high value (≈ 0.7). This shows that the network struggles to learn an accurate prediction. This is expected as the random transitions of the environment between sensorimotor pairs $(\mathbf{m}_t, \mathbf{s}_t)$ in the database make the sensory experiences inconsistent. The dissimilarity measure Q_p stays high (≈ 0.13) during the learning and slightly oscillates. This oscillation is reduced in the second half of the learning due to the smaller learning rate. On the other hand, the dissimilarity measure Q_h also stays high (> 0.12) during the learning. It however goes through an initial phase of significantly greater values (≈ 0.2) before progressively decreasing to its base value (≈ 0.1). This is better illustrated in Fig. 4 where the training has been pursued for $4e^6$ epochs. The same figure also shows a comparison with a network using ReLu units (Glorot et al., 2011) in place of the SeLu units. It suggests that the initial overshooting of Q_h seems to be attributed to the SeLu units initialization and self-normalizing mechanism, and not to a particular property of the sensorimotor experience. (Note also that no qualitative difference was observed in the representation when using ReLu units).
In agreement with these observations, Fig. 3 shows that the representation $\mathbf{h}$ of the motor states $\mathbf{m}$ does not display a particular structure compared to the regular one that the ground-truth positions exhibit. Intuitively, because sensorimotor transitions are inconsistent and the environment moves, the motor states fed to the network are not informative for the prediction of the future sensory state. As a consequence, the network learns to output an average sensory state which statistically minimizes prediction error while not taking into account the motor states. The intermediary representation thus converges to an arbitrary manifold which is disregarded by the predictive module.

MM exploration: The experience of consistent sensorimotor transitions greatly impacts the loss and the dissimilarity measures. The environment being static and unambiguous by construction, the predictive task is now trivial, and the loss rapidly drops to very small values of the order of 10^{-8}. Because of the stopping criterion used in the simulations, the 10 independent trials end before the maximal number of epochs set to $2e^6$ epochs. The mean and standard deviation values displayed in Fig. 2 are thus progressively computed over less and less trials, as some of them have already converged and been stopped. The last simulation reaches the threshold after $7.6e^5$ epochs. The final drop observed in the Q_p and Q_h curves at epoch $6.1e^5$ corresponds to the penultimate trial being stopped, the mean collapsing to the measures of the last still running trial, and the standard deviation being undefined. More importantly, Q_p and Q_h undergo a rapid decrease before stabilizing around values which are significantly lower than in the previous scenario for Q_p (≈ 0.07) and similar to the previous scenario for Q_h (≈ 0.1). Both curves also exhibit a significantly lower standard deviation. This correlates with the representation displayed in Fig. 3 in which we can see that the affine projection of $\mathbf{h}$ in the position space gives rise to a point cloud with the same topology as the ground truth positions. In particular, all redundant motor states $\{\mathbf{m} = [m_1, m_2, m_3], \forall m_3\}$ are projected onto the same point $\mathbf{h}$. In the 3D representational space we can however see that the representations do not lie on a flat 2D manifold but are instead spread in the whole space. Indeed, the constraints induced by the topological invariants while exploring the static environment drive the motor representation to copy the topology of the sensory space. Nevertheless the representation can perfectly respect these constraints by forming a curved 2D manifold in the representational space. During training, the representation can thus converge to an infinity of such solutions, of which a flat manifold is only a negligible portion.

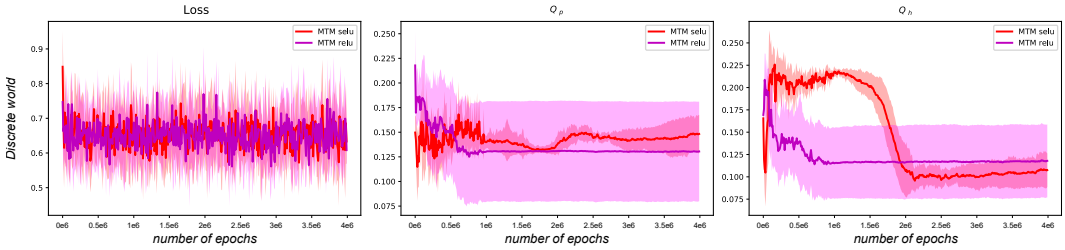

Figure 4: Evolution of the loss and the dissimilarity measures Q_p and Q_h during a longer training session for the discrete world and the MTM type of exploration. Two networks using SeLu units and ReLu units are compared. The displayed means and standard deviations are computed over 10 independent trials. (Best seen in color)

This analysis also leads to the conclusion that both Q_p and Q_h are not ideal criteria to compare the topology of the representation with the one of the ground-truth position as they implicitly assume a linear mapping between the two. They act as an upper bound, as a topology comparison method taking into account the non-linearity between the two manifolds would necessarily produce a lower dissimilarity. Performing such a non-linear comparison would however require a significantly more complex comparison of the two point clouds. Although not ideal, Q_p and Q_h are sufficient to evaluate the topological evolution of the representation's topology between the different types of exploration in this work. Moreover, Q_p appears to be a more adequate criterion than Q_h, as Q_h fails to capture the structural difference between the MTM and MM scenarios (see Fig. 2).

MMT exploration: The addition of translations of the environment between sensorimotor transitions greatly impacts the dissimilarity measures. As in previous case, the loss rapidly drops to very small values ($\approx 3e^{-6}$). Despite the sensorimotor predictive mapping being more complex than in the static case, the network is thus able to produce accurate sensory predictions. Moreover Q_p and Q_h also drop to very small values ($\approx 2e^{-3}$ and $\approx 3e^{-3}$ respectively). This indicates that the representation built by the agent and the ground-truth position have similar structures, and that there exists a simple affine transformation between the two. This is confirmed in Fig. 3, where we can see that both the positions and representations align perfectly after affine compensation in the position space. Therefore the agent has been able to capture both the topology and the metric regularity of its position in the environment.
Note that this time the representation $\mathbf{h}$ even approximates a flat 2D manifold in the 3D representational space. However this is not necessarily the case in all trials, as shown in Fig. 5. Indeed in the network, the representation $\mathbf{h}$ is fed to a fully connected layer in which each neuron performs a linear projection of $\mathbf{h}$ before passing it through its activation function. As a consequence, the network has two equivalent options to respect the sensory invariants induced by the MMT exploration: i) flatten the 2D representation manifold such that all the metric equalities between the motor states are consistent, or ii) adapt the weights of the connection to the next layer such that each neuron performs a projection which disregards all dimensions of the representational space, except two. In both cases, the input received by the predictive module Net_{enc} would be identical. Depending on the trial, it seems that the network converges arbitrarily to one of those two solutions, or rather a mix of the two, often displaying a slight curvature of the 2D manifold in the representational space. This explains the slightly greater standard deviation of Q_h compared to Q_p observed in Fig. 2. Note that even when the network converges to a curved manifold, the projective nature of the second option ensures that the affine projection of the representation in the ground-truth position space preserves the metric constraints induced by the sensory invariants. This explains why the seemingly spread cloud of points $\mathbf{h}$ observed in Fig. 5 does actually correspond to a topologically organized grid when projected in the position space via an affine transformation.

B.2 Arm in a room

We now perform the same analysis on the more complex arm in a room setup.

MTM exploration: Like in the discrete world setup, the predictive task is made difficult by the translation of the environment occurring during sensorimotor transitions. In Fig. 2, the loss stagnates

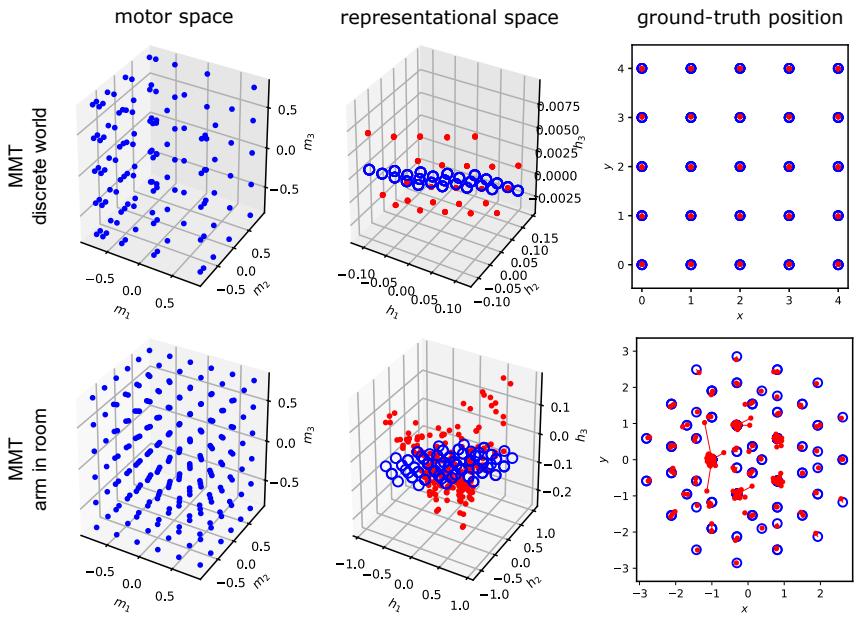

Figure 5: Visualization of the (normalized) regular motor space sampling, its encoding in the representational space (red dots), and the corresponding ground-truth positions (blue circles) for the MMT type of exploration, for different trials of the discrete world setup (a) and of the arm in a room setup (b) then in Fig. 3. Both the representations and the positions are also projected into each other's space via an optimized affine transformation. (Best seen in color)

around a high value ($\approx$ 1.0), after a quick drop due to the network finding an adequate average value and scale for $\tilde{\mathbf{s}}_{t+1}$. Additionally, both Q_p and Q_h progressively decrease during the learning, before reaching a plateau at relatively high values ($\approx$ 0.12 each). This progressive decrease was not observed in the discrete world setup. We hypothesize it is due to a border effect related to the environment having walls. Indeed because the world is not a torus in this setup, the motor states corresponding for instance to the sensor being to the left of the agent's base rarely observe the sensory states associated with positions of the sensor on the far right of the environment. Similarly, different positions of the sensor statistically observe slightly different sensory distributions. Consequently this drives the network to capture an approximation of the topology of the sensor's position where, in particular, redundant motor states tend to be represented by the same point (as they are associated with the same sensory distribution). This explains why the dissimilarity between the representation and the position measured by Q_p and Q_h decreases during training. Note however that this mechanism is not sufficient to capture the true topology of the sensor position as it is based on the statistics of the sensory states, and not on their absolute values (different sensorimotor mappings could lead to similar sensory statistics without necessarily inducing the topological invariants described in Sec. 3). The representation displayed in Fig. 3 indeed shows that the representation built by the network seems less arbitrary than in the discrete world setup, as motor states corresponding to close sensor positions tend to appear closer in the representation than in the discrete world setup. Yet the representation still does not capture the topology of the sensor's position.

MM exploration: In the static environment, the loss quickly drops to small values ($\approx$ 0.07). This shows that the sensorimotor predictive task is relatively easy due to the environment not moving, despite the sensorimotor mapping being more complex that in the discrete world setup. The measures Q_p and Q_h also rapidly decrease, before reaching a plateau ($\approx$ 0.09 and $\approx$ 0.12 respectively). This seems to indicate that the representation exhibits a structure more similar to the ground truth position as in the MTM exploration. This is confirmed in Fig. 3, where we can see in the position space that the representation topologically organizes as the star-shaped grid of position associated with the regular sampling of the motor space (despite some outliers). In particular, the multiple redundant motor configurations which lead to the same external position of the sensor are clustered together (see for instance the inner corners of the star). In the representation space, the representations appear spread in the 3D space, for the same reason as described for the discrete world setup. Moreover,

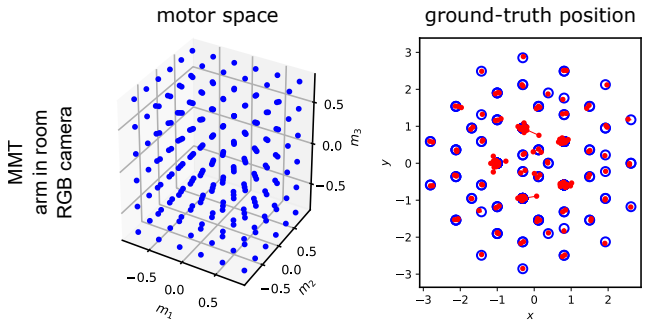

Figure 6: Visualization of the (normalized) regular motor space sampling, the corresponding ground-truth positions (blue circles), and the affine transformation of the representation into the ground-truth space of positions (red dots) for one trial of the arm in a room simulation with a RGB camera of 16 pixels, a representational space of dimension $dim_H = 25$ and a MMT type of exploration. (Best seen in color)

these results confirm that Q_h is not the best measure to analyze the structure of the representation. Indeed, in Fig. 2, Q_h fails to capture the structural difference between the MTM and MM scenarios, whereas Q_p does.

MMT exploration: With consistent sensorimotor transitions in a moving environment, the loss rapidly decreases and reaches a plateau around a small value (≈ 0.25). The network is thus able to learn a relatively good sensorimotor predictive model. The residual error is however greater than in the discrete world setup, in part because the sensory experience in this environment can be ambiguous. For instance, when sensing only the right wall, the environment could be in many different positions, and the agent thus cannot predict the future sensory state with certainty. More importantly, Q_p and Q_h also drop to very small values (≈ 0.014). This indicates a high structural similarity between the representation and the ground-truth position. This is confirmed in the ground-truth position space of Fig. 3, where we can see that the representation almost perfectly matches the star-shaped grid of positions induced by the motor sampling after affine compensation. Therefore the agent has been able to capture both the topology and the metric regularity of its position in the environment. Note that in this trial the representation $\mathbf{h}$ once again approximates a flat 2D manifold in the 3D representational space. This is not necessarily the case in all trials, as shown in Fig. 5, for the same reason as discussed in the discrete case.

B.3 Additional experiment

We propose a last additional experiment in order to assess the robustness of the approach to more complex sensory inputs and representational space of higher dimension. The arm is now equipped with a 1D RGB camera (in the 2D room) with a resolution of 16 pixels. Additionally dim_H is set to 25 instead of 3. The resulting learning curves are displayed in black in Fig. 2, for the MMT exploration case. Like with the initial arm in a room setup, the loss progressively decreases and reaches a plateau (≈ 2.1). The predictive loss thus exhibits a similar behavior, despite its overall greater magnitude that is due to the MSE being computed in a sensory space of dimension 48 instead of 10. Likewise the evolution of the errors for Q_p and Q_h is qualitatively identical to the case of the simpler arm, despite the largely greater dimension of the representational space. The measures even converge to identical values. This seems to indicate that the network built a representation which is structurally similar to the ground truth positions. This in confirmed in Fig. 6 where the affine projection of the representation matches the ground-truth position of the sensor (the representational space is now of too high dimension to be straight-forwardly visualized). The approach thus seems insensitive to the complexity of the sensory input and to the dimension of the representational space[1].

[1]The RGB camera experiment was also run for $dim_H = 3$ and generated the same qualitative results.