LEARNING GROUP STRUCTURE AND DISENTANGLED REPRESENTATIONS OF DYNAMICAL ENVIRONMENTS

Robin Quessard
École Normale Supérieure, Paris, France
indust.ai, Paris, France

Thomas D. Barrett
University of Oxford, Oxford, UK

William R. Clements
indust.ai, Paris, France

ABSTRACT

Representing observational data in a way that both correctly reflects and disentangles the underlying data generation mechanisms plays an important role in causal inference. Inspired by physics, we propose a method, built upon group representation theory, that learns a representation of an environment structured around the transformations that generate its evolution. Experimentally, we learn the structure of explicitly symmetric environments without supervision while disentangling independent data generation mechanisms. We show that the learned representations allow for accurate long-horizon predictions and further demonstrate a correlation between the quality of predictions and disentanglement of the representations.

1 INTRODUCTION

The notion of representation learning occupies a central place in the machine learning literature (Bengio et al. (2013); Ridgeway (2016)). A good representation should ideally reflect and disentangle the underlying data generation mechanisms, and can be used to efficiently predict, classify, or generalize. However, learning interpretable representations of data that explicitly disentangle the underlying mechanisms structuring this data is still a challenge.

To address this, one can begin by drawing a parallel between the pursuit of underlying structure in machine learning and in physics. In machine learning, data is structured by its underlying generative factors, which can be viewed as the causal mechanisms that when intervened on independently and tractably transform the generated data (Peters et al. (2017)). Physics often searches for structure using group representation theory by considering the infinitesimal transformations that generate the symmetry group of a physical environment (Lie (1893); Weinberg (1995)). In both physics and machine learning, one has to find a faithful – and, ideally, interpretable – representation of these generative factors to structure the representation one has of the environment. This connection between representation learning in machine learning and representations in physics was previously highlighted in Higgins et al. (2018). However, although they propose to define representations with respect to the analogy with physics, they do not propose a method for learning these representations from data.

In this work, motivated by the parallel between transformations in physics and in machine learning, we propose a method for learning disentangled representations of dynamical environments. Our method focuses on learning the structure of the symmetry group ruling the environment's transformations, where symmetry transformations are understood as transformations that do not change the nature of the objects. For this purpose we focus on representing dynamical environments through a representation of observations and transformations. We encode observations as elements of a latent space and represent transformations as special orthogonal matrices that act linearly on the latent space. As we consider not only representations of observations but also representations of transformations, in the following we will semantically overload the term of representation to describe the full representation of an environment through its transformations and its observations.

Correspondence to: william.clements@indust.ai

2 RELATED WORK

Different definitions of what constitutes, and how to learn, a disentangled representation have been put forward (Locatello et al. (2018)). Generative Adversarial Networks (GAN) (Goodfellow et al. (2014)) and Variational AutoEncoders (VAE) (Kingma & Welling (2013)) have been used with some success to identify loosely defined data generative factors (Higgins et al. (2017a); Chen et al. (2016); Karras et al. (2019)) in non-interactive datasets. These approaches have also been used for dynamical environments (Burgess et al. (2018)), where they have focused on learning disentangled state representations that can for example be used for domain adaptation (Higgins et al. (2017b)).

However, it was argued by Higgins et al. (2018) that a disentangled representation of an environment should focus not only on its states but also on its transformations. They proposed a more formal definition of disentangled representations based on the physical notion of symmetry transformations. Based upon this definition, Caselles-Dupré et al. (2019) showed that symmetry-based disentangled representation learning requires interaction with environments. However, Higgins et al. (2018) did not propose a method for learning such representations, and the method of Caselles-Dupré et al. (2019), which in contrast to our work uses a type of VAE, requires prior knowledge of the symmetries in the system. Prior work on learning underlying group structure from data (Cohen & Welling (2014a;b)) also assumed prior knowledge of the symmetry group. To the best of our knowledge, our work is the first to learn the underlying group structure of environments and disentangled representations (as defined in Higgins et al. (2018)) without any prior knowledge of the symmetry group.

The notion of learning a disentangled representation centered around transformations of the environment is closely related to the notion of learning independent causal mechanisms (Parascandolo et al. (2018)). However, whereas a causal perspective of disentangled representations as well as disentanglement metrics have been proposed by Suter et al. (2019), they also stop short of proposing a method for learning such representations. By viewing transformations applied to the environment as interventions on the causal mechanisms underlying the data, our work also provides a method for learning representations that are disentangled in the sense of Suter et al. (2019).

Interesting parallels can be drawn between our work and state-space models in machine learning. Learning disentangled representations of dynamical environments is important for state-space models to robustly predict the evolution of complex systems (Miladinović et al. (2019)). Moreover, the state-space model implementation of Fraccaro et al. (2017) is theoretically close to our method as they consider representations of both observations and the dynamics that act linearly on the latent space, however their method does not reveal the group structure of the transformations. In parallel, physics and machine learning have been forging strong ties mostly based upon the Hamiltonian theory and the integration of ordinary differential equations in the latent space to describe the evolution of dynamical systems (Chen et al. (2018), Toth et al. (2019), Greydanus et al. (2019)). Finally, Hamiltonian-based methods have also been used to discover specific symmetries of physical systems (Bondesan & Lamacraft (2019)).

3 DISENTANGLED REPRESENTATIONS

3.1 OVERVIEW

In this section we review the notion of symmetry-based disentangled representations which is based upon the definition provided by Higgins et al. (2018). As we intend to represent observations as elements of a vector space V, the representation (or realization) of the symmetry group we wish to learn has to be a linear representation on V. By focusing on linear disentangled representations as described by group theory (Hall (2015)), we enforce all transformations of the environment to be represented by only linear transformations in the learnt latent space.

The goal of representation learning is to discover useful representations of data (Ridgeway (2016)). Not only must representations be faithful and preserve the information held in the data, they must be explicit and interpretable such that every generative factor is expressed and can be easily identified when looking at the representation. Specifically, in our case, we want to represent observations of a dynamical environment in a latent space that exhibits all those qualities. Learning such representations requires using inductive biases Locatello et al. (2018), which in our case take the form of symmetry transformations that operate on the environment.

3.2 REPRESENTATIONS OF DYNAMICAL ENVIRONMENTS

Consider a dynamical environment from which we can extract observations and a set of transformations that act on the environment. As in physics, we assume that those transformations generate a symmetry group G. We will have learnt a representation of this environment if we can map observations of the environment to elements of a latent space V and map symmetry transformations to linear applications on this latent space such that the group structure of the symmetry group is conserved in the latent space.

Formally, learning a symmetry-based representation requires learning the structure of the symmetry group and a latent representation of the observations. This means finding a homomorphism $\rho : G \to GL(V)$ between the symmetry group G and the general linear group of the latent space $GL(V)$. In order to learn a full representation of an environment, the observation space X shall be mapped to the latent space V with $f : X \to V$ and transformations $g \in G$ shall be represented as matrices acting linearly on the latent space so that the map in Figure 1 is equivariant.

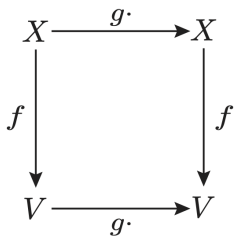


Figure 1: Equivariant map between the space of observations X and the latent space V

For example, consider the observation x_1 of a ball, the transformation g of moving the ball to the left, the observation of the ball after the transformation is $x_2 = g.x_1$. We need to verify $f(g.x_1) = \rho(g).f(x_1)$. We make the common shortcut of forgetting the notation $\rho(g)$ and we use g to denote the transformation in both the observation space and the latent space.

3.3 DISENTANGLEMENT OF REPRESENTATIONS

Another requirement we wish to make concerning the representation to be learned is that it is disentangled in the sense of Higgins et al. (2018). The question of disentanglement makes sense because it ensures the interpretability of the model of the environment we learn. Furthermore, if a robot or any intelligent agent wants to learn a representation of its environment, it should learn a disentangled representation so it can associate simple actions to distinct subspaces of the representation it has of its environment. Then it is simpler for it to perform tasks (Raffin et al. (2019)) and learn complex actions in this representation as they would become combinations of simple disentangled actions.

Formally, if there exists a subgroup decomposition of G such that $G = G_1 \times G_2 ... \times G_n$, we would like to decompose the representation (ρ, V) in subrepresentations $V = V_1 \oplus V_2 ... \oplus V_n$ such that the restricted subrepresentations $(\rho_{|G_i}, V_i)_i$ are non-trivial and the restricted subrepresentations $(\rho_{|G_i}, V_j)_{j \neq i}$ are trivial (we recall that a trivial representation of G is equal to the identity for every element of the group G).

4 METHODS

4.1 PARAMETERIZATION

Our goal is to learn a disentangled representation of the symmetry group with no prior knowledge of the actual symmetries of the environment so that any transformation of the environment can be represented by a linear operator in the latent space. Because we are looking for a matrix representation of an *a priori* unknown symmetry group G, we need to use a parameterization of a group of matrices large enough to potentially contain a subgroup that is a representation of G. We will also restrain ourselves to real representations.

We assume that G can be represented by a group of matrices belonging to $SO(n)$ which is the set of orthogonal $n \times n$ matrices with unit determinant. Given its prevalence in physics and in the natural world we can expect $SO(n)$ to be broadly expressive of the type of symmetries we are most likely to want to learn. As orthogonal matrices conserve the norm of vectors they act on, this corresponds to encoding observations in a unit-norm spherical latent space (Davidson et al. (2018); Connor & Rozell (2019)).

We parameterize the n-dimensional representation of any transformation g as the product of $n(n-1)/2$ rotations (Pinchon & Siohan (2016); Clements et al. (2016)) :

$$g(\theta_{1,2}, \theta_{1,3}.., \theta_{n-1,n}) = \prod_{i=1}^{n-1} \prod_{j=i+1}^{n} R_{i,j}(\theta_{i,j}) \tag{1}$$

Where $R_{i,j}$ denotes the rotation in the i, j plane embedded in the n-dimensional representation. For example, in a 3-dimensional representation, one of the rotations $R_{1,3}$ is :

$$R_{1,3}(\theta_{1,3}) = \begin{pmatrix} \cos\theta_{1,3} & 0 & \sin\theta_{1,3} \\ 0 & 1 & 0 \\ -\sin\theta_{1,3} & 0 & \cos\theta_{1,3} \end{pmatrix} \tag{2}$$

And any transformation g in a 3-dimensional representation has 3 learnable parameters such that :

$$g = \begin{pmatrix} \cos\theta_{1,2} & \sin\theta_{1,2} & 0 \\ -\sin\theta_{1,2} & \cos\theta_{1,2} & 0 \\ 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} \cos\theta_{1,3} & 0 & \sin\theta_{1,3} \\ 0 & 1 & 0 \\ -\sin\theta_{1,3} & 0 & \cos\theta_{1,3} \end{pmatrix} . \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_{2,3} & \sin\theta_{2,3} \\ 0 & -\sin\theta_{2,3} & \cos\theta_{2,3} \end{pmatrix} \tag{3}$$

These parameters, θ, are learnt jointly with the parameters of an encoder f_ϕ mapping the observations to the n-dimensional latent space and a decoder d_ψ mapping the latent space to observations. The training procedure is such that we encode a random observation x_0 with f_ϕ then we transform in parallel the environment and the latent vector $f_\phi(x_0)$ using m random transformations and their representation matrices $\{g_k\}_{k=1,...,m}$. The result of those linear transformations in the latent space is decoded with d_ψ and yields $\hat{x}_m$:

$$\hat{x}_m(\phi, \psi, \theta) = d_\psi(g_m(\theta).g_{m-1}(\theta)....g_1(\theta).f_\phi(x_0)) \tag{4}$$

The training objective is the minimization of the reconstruction loss $L_{\text{rec}}(\phi, \psi, \theta)$ (in the following we use a binary cross entropy) between the true observations x_m obtained after the successive transformations in the environment and the reconstructed observations $\hat{x}_m$ obtained after the successive linear transformations using the representations $g_k(\theta)$ on the latent space.

4.2 DISENTANGLEMENT

As explained in section 3.3, for a representation to be disentangled, each subgroup of the symmetry group should act on a specific subspace of the latent space. We want to impose, without supervision, this disentanglement constraint on the set of transformations $\{g_a\}_a$ that act on the environment. In order to do so without any prior knowledge on the structure of the symmetry group, our intuition is that if each transformation g_a acts on a minimum of dimensions of the latent space, then the representation can naturally disentangle itself.

We formalize this notion of entanglement into a metric L_{ent} proper to our parameterization. We choose a metric that quantifies sparsity and interpretability through the number of rotations (each of which is parameterized by $\theta^a_{i,j}$) involved in the transformation matrices $g_a(\theta^a_{i,j})$. The smallest non-trivial transformation matrix involves a single rotation, so L_{ent} measures the use of any additional rotations :

$$L_{\text{ent}}(\theta) = \sum_a \sum_{(i,j)\neq(\alpha,\beta)} |\theta^a_{i,j}|^2 \qquad \text{with} \quad \theta^a_{\alpha,\beta} = \max_{i,j}(|\theta^a_{i,j}|) \tag{5}$$

The higher L_{ent}, the higher the entanglement of the representation of the set of transformations. Minimizing this metric L_{ent} makes sure that for each transformation $g_a \in \{g_a\}_a$, most of the parameters appearing in the representation of this transformation go to 0, which implies that the transformation acts on a minimum of dimensions of the latent space. If there is only one non-zero parameter in the parameterization of a transformation, then it only acts on 2 dimensions of the latent space.

5 EXPERIMENTS

The code to reproduce these experiments is available at `https://github.com/IndustAI/learning-group-structure`.

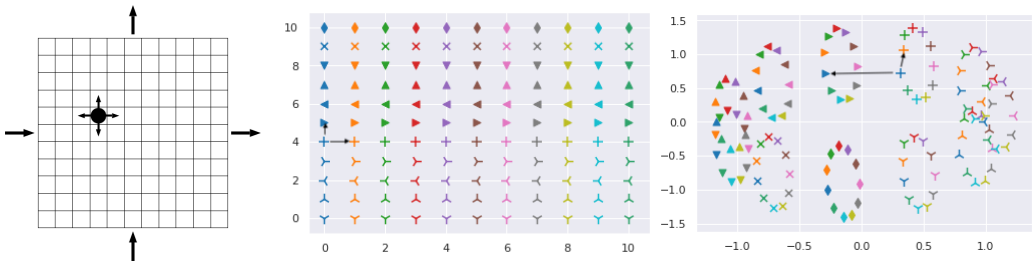
5.1 LEARNING THE LATENT STRUCTURE OF A TORUS-WORLD

Our first goal is to show that the parameterization and the training method described in 4.1 allows us to extract useful information about the structure of an environment from looking at the topology of the learnt latent space. We use a simple environment similar to Higgins et al. (2018), consisting of a ball evolving in a 2-dimensional grid-world of size $n \times n$ with periodic boundary conditions. At each timestep, the ball can move one step left, right, up, and down and observations are returned as $n \times n$ matrices with value 1 at the position of the ball and 0 elsewhere. A 2-dimensional plane with periodic boundary conditions is topologically equivalent to a torus. It is this topology that we aim to learn from the dynamics of the environment.

Concretely, the symmetry group of this environment is the finite group $G = C_n \times C_n$ where C_n denotes the cyclic group of order n (also called $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}_n$) and is a finite subgroup of $SO(2) \times SO(2)$. In order to learn a representation of this environment, we need to learn an encoder, a decoder and the representation matrices for the 4 transformations $g_{\text{up}}, g_{\text{down}}, g_{\text{left}}$ and g_{right} that generate G.

To learn this group structure from data, our only assumption is that it can be represented with 4-dimensional orthogonal matrices $g \in SO(4)$. This choice is motivated from the fact that real representations of cyclic groups can be seen as rotations in planes. Since the symmetry group consists of the direct product of 2 cyclic groups, we need 2 planes so 4 dimensions to learn a real representation of it. We recall that a matrix of $SO(n)$ has $n(n-1)/2$ degrees of freedom so the matrices of this 4-dimensional representation each have 6 parameters.

We use neural networks for the encoder and the decoder. The encoder $f_\phi : \mathbb{R}^{n^2} \to \mathbb{R}^4$ has normalized outputs so that it always maps observations to unit-norm latent vectors. We learn jointly the encoder parameters ϕ, the decoder parameters ψ and the $6 \times 4 = 24$ parameters of the 4 transformation matrices $g_{\text{up}}, g_{\text{down}}, g_{\text{left}}$ and g_{right} which we denote as θ.



(a) Considered environment (b) Position of the ball on the grid (c) 2D projection of the latent space

Figure 2: Correspondence between the observation space (b) and the latent space (c) in the case where $n = 11$. Using a random projection, we can see the toroidal structure of the latent space, characteristic of $SO(2) \times SO(2)$.

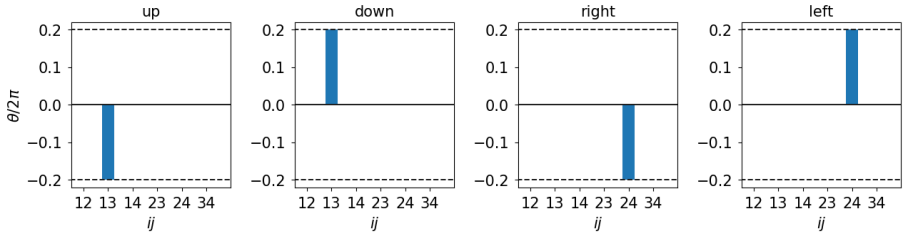
Our results are shown in Figure 2 in which we see that we learn a 4-dimensional representation of the finite symmetry group $C_n \times C_n$. The explicit toroidal structure of the latent space thus respects the structure of the symmetry group. Therefore, we are able to learn without supervision the underlying symmetry structure of the environment and an equivariant map between the observation space and the latent space.

5.2 CONTROLLING THE ENTANGLEMENT OF THE REPRESENTATION

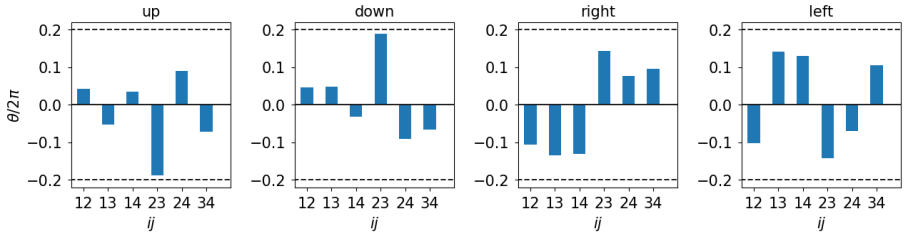
We have shown that we could learn a representation of this environment, now we show that we can control its entanglement using the metric introduced in 4.2. Indeed, many 4-dimensional representations of $C_n \times C_n$ exist, and most of them are entangled. Since the transformation matrices are parameterized as products of 6 rotations, if most of the 6 parameters are non-zero, then the transformation is poorly interpretable because all dimensions of the latent space are mixed after acting on it with this representation.

Using the entanglement metric from section 5 as a regularization term, we are able to control the entanglement of the learnt represenation. Figure 3 compares the learnt transformations between a

regularization minimizing the entanglement ($\lambda > 0$, Figure 3a) and a regularization maximizing the entanglement ($\lambda < 0$, Figure 3b) in the $n = 5$ case. Even though both representations encode $C_5 \times C_5$ and exhibit the corresponding toroidal structure, the maximally disentangled representation is much more interpretable. The up/down transformations rotate in a single plane (dimensions 1 and 3) by $\pm 2\pi/5$, whereas the left/right transformations act equivalently in an orthogonal space (dimensions 2 and 4). This is the most intuitive 4-dimensional disentangled representation of the symmetry group $C_5 \times C_5$ we could have learnt.



(a) Minimization of the entanglement. Value of entanglement : $L_{\text{ent}} = 4.8 \times 10^{-7}$

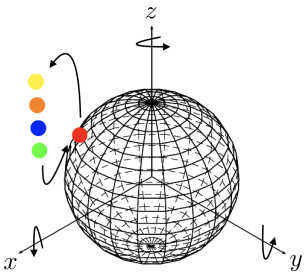


(b) Maximization of the entanglement. Value of entanglement : $L_{\text{ent}} = 1.60$

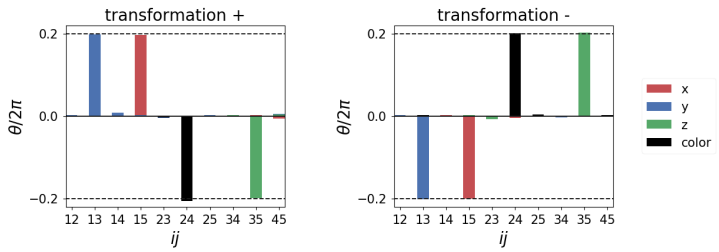
Figure 3: Learnt transformations for a grid-world with $n = 5$. Values of the angles $\theta_{i,j}/2\pi$ learned for representations with regularization on the loss to minimize the entanglement (a) and with regularization to maximize the entanglement (c).

5.3 Learning more complex structure

We now show that we are able to learn disentangled representations of environments with more complex symmetry group structures to prove the robustness of our method. We consider a colored point evolving on a 3-dimensional sphere. The transformations that we consider are discrete rotations around the sphere and periodic discrete changes of color of the point. We wish to disentangle two factors of variation: the spatial rotations and the changes of color.



(a) Considered environment



(b) Representations learned for the 8 transformations. The 3D rotations of $\pm 2\pi/5$ are associated to the dimensions 1, 3 and 5 of the latent space. The change of color is associated to dimensions 2 and 4 of the latent space.

Figure 4

We use a set of 5 colors that we visualize on a periodic line and the transformations corresponding to the periodic changes of color are denoted as *color+* and *color-*. We also learn a set of 3-dimensional rotations around the axes of the sphere of radius r that the colored point lives on. These 6 trans-

formations are denoted *x+, x-, y+, y-, z+* and *z-* and they respectively correspond to rotations of an angle $2\pi/5$ and $-2\pi/5$ around each axis. The symmetry group generated by those transformations, and that we aim to learn, therefore lies in $SO(2)_{\text{color}} \times SO(3)_{\text{rotation}}$.

As explained in Higgins et al. (2018), learning a disentangled representation of 3D rotations directly questions the definition we give of a disentangled representation. Indeed, $SO(3)$ cannot be written as a non-trivial direct product of subgroups, therefore, we cannot find a representation in which two rotations around different axes would act on two different subspaces of the latent space. We can still satisfy ourselves with a disentangled representation in which rotations around the x, y and z axes each act on a minimum of dimensions of the latent space, a definition of entanglement aligned with the metric we introduced in 4.2.

We choose to learn a 5-dimensional representation of this environment because an interpretable disentangled representation would associate a 3-dimensional subspace to the space transformations and a 2-dimensional subspace to the color transformations. Nevertheless, using a higher-dimensional latent space does not change the learnt representation as the disentanglement objective makes unnecessary dimensions impactless and present in none of the transformation matrices.

Figure 4b shows that, when also minimizing the entanglement metric, we effectively learn this way a 5-dimensional disentangled representation of the environment that conserves the symmetry group structure and such that the spatial transformations act on a distinct subspace from the subspace on which the color transformations act on.

5.4 Learning disentangled representations of Lie groups

As a final step for learning disentangled representations of symmetry groups, we now show that we are able to learn continuous groups corresponding to infinite sets of transformations and are therefore not limited to discrete environments. We consider a point evolving on a sphere under the continuous set of rotations around the 3 axes x, y and z in the interval $[-\pi, \pi]$.

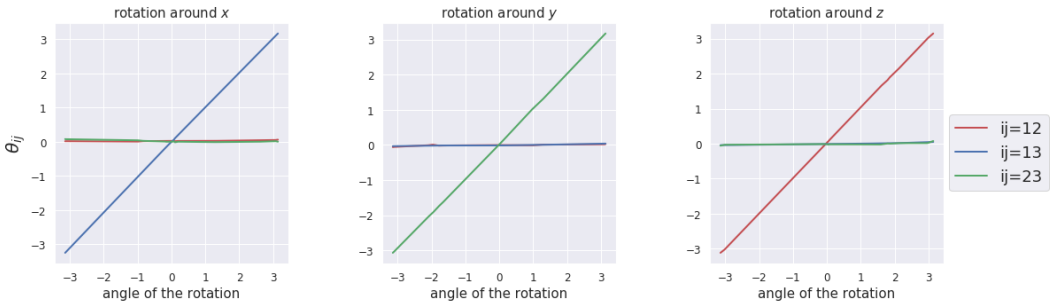


Figure 5: Value of the parameters $\theta_{1,2}, \theta_{1,3}$ and $\theta_{2,3}$ of the transformation matrices, which are the outputs of the neural network ρ_σ after training, as a function of the angle of the rotation in the environment for each rotation axis x, y and z. For any angle and each rotation axis, there is only one non-zero parameter (for example $\theta_{2,3}$ for rotations around the y axis) which makes this representation a well-disentangled and interpretable representation.

To learn a continuous group of symmetry transformations, also known as a Lie group, we approximate the group homomorphism $\rho : G \to GL(V)$ using a neural network, ρ_σ. The network takes as input the transformation in the environment: a concatenation of a scalar denoting the rotation axis (0 for x, 1 for y and 2 for z) and the value of the angle of the rotation around this axis. It outputs $n(n-1)/2$ scalars parameterizing the n-dimensional representation of this transformation.

In this case, we aim to learn the 3D rotations around the axes of a sphere so we use a 3-dimensional latent space to represent the symmetry group $SO(3)$. We recall that, in our parameterization, 3-dimensional representations have 3 parameters corresponding to rotations $R_{1,2}(\theta_{1,2})$, $R_{1,3}(\theta_{1,3})$ and $R_{2,3}(\theta_{2,3})$. For example, the representation g of a rotation of $\pi/4$ around the y axis is parameterized

as a product of 3 rotations such that :

$$g(\theta_{1,2}, \theta_{1,3}, \theta_{2,3}) = R_{1,2}(\theta_{1,2}).R_{1,3}(\theta_{1,3}).R_{2,3}(\theta_{2,3}) \quad \text{with} \quad \begin{pmatrix} \theta_{1,2} \\ \theta_{1,3} \\ \theta_{2,3} \end{pmatrix} = \rho_\sigma(\begin{pmatrix} 1 \\ \pi/4 \end{pmatrix}) \tag{6}$$

As in previous sections, the training objective is the minimization of a loss combining a reconstruction loss and an entanglement regularization $L(\phi, \psi, \sigma) = L_{\text{rec}}(\phi, \psi, \sigma) + \lambda L_{\text{ent}}(\sigma)$.

Our results are shown in Figure 5, proving that we effectively learn a 3-dimensional representation of $SO(3)$, where rotation about each axis acts only within a single plane of the latent space. With this 3-dimensional disentangled representation of a continuous group of symmetry transformations in a 3-dimensional space, we show that we learn a perfectly interpretable representation of an environment with continuous dynamical transformations.

5.5 Multi-step predictions with disentangled representations

We now show that the learnt representations are capable of excellent long-term predictions. We go back to the torus-world environment of a ball evolving on a $n \times n$ grid with periodic boundary conditions. Using our entanglement metric L_{ent}, we control the entanglement of the representations to reach a target entanglement T_{ent}. We train the parameters of the encoder f_ϕ, the decoder d_ψ and the representation matrices parameters θ to minimize the objective : $L(\phi, \psi, \theta) = L_{\text{rec}}(\phi, \psi, \theta) + \lambda |L_{\text{ent}}(\theta) - T_{\text{ent}}|$.

During training, we minimize this objective on sequences of 10 successive random transformations sampled among *up, down, left*, and *right*. After training, we use the learnt representations to predict the state of the environment over 500 random transformation steps in the latent space and we measure the reconstruction error between the decoded observation and the true one.

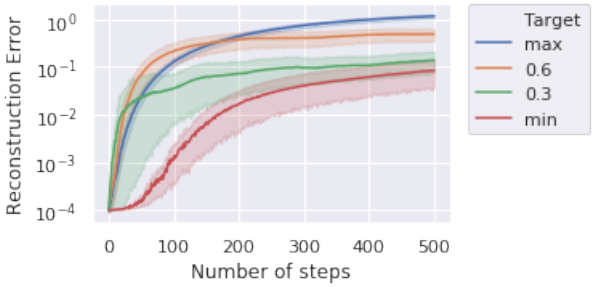


Figure 6: Value of the reconstruction error between the true observation and the reconstruction as a function of the number of transformations in the latent space for several target entanglements. Shaded areas correspond to 95% confidence intervals of the mean, measured over 100 training seeds.

Figure 6 shows that the lower the entanglement of the learnt representation, the better the long term predictions. The label "min" stands for minimization of the entanglement ($T_{\text{ent}} = 0$ and $\lambda > 0$) and "max" for maximization of the entanglement ($T_{\text{ent}} = 0$ and $\lambda < 0$). For "0.3" and "0.6", the label is the value of T_{ent} and $\lambda > 0$. This result is in agreement with the widespread notion that disentangled representations make better predictions (Bengio et al. (2013); Ridgeway (2016); Higgins et al. (2017b)).

6 Conclusion

In this work, we have opened the possibility of applying representation theory to the problem of learning disentangled representations of dynamical environments. We have exhibited the faithful, explicit and interpretable structure of the latent representations learnt with this method for simple symmetrical environments with a specific parameterization. With this method, the structure of the latent space naturally respects the structure of the symmetry group without imposing any constraint on the latent space during training. Adding a very general regularization on the parameters of the transformation matrices makes the representations easily interpretable, yielding a representation very similar to a representation that physicists would derive when formulating their conception of the symmetries of the environment. Whilst performance in complex real world environments remains untested we think that this further evidences the benefits of applying physics-based biases to representation learning.

REFERENCES

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Roberto Bondesan and Austen Lamacraft. Learning symmetries of classical integrable systems. *arXiv preprint arXiv:1906.04645*, 2019.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Symmetry-Based Disentangled Representation Learning requires Interaction with Environments. *arXiv preprint arXiv:1904.00243*, 2019.

Tian Qi Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pp. 6571–6583, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

William R Clements, Peter C Humphreys, Benjamin J Metcalf, W Steven Kolthammer, and Ian A Walmsley. Optimal design for universal multiport interferometers. *Optica*, 3(12):1460–1465, 2016.

Taco Cohen and Max Welling. Learning the irreducible representations of commutative lie groups. In *International Conference on Machine Learning*, pp. 1755–1763, 2014a.

Taco S Cohen and Max Welling. Transformation properties of learned visual representations. *arXiv preprint arXiv:1412.7659*, 2014b.

Marissa Connor and Christopher Rozell. Representing closed transformation paths in encoded network latent space. *arXiv preprint arXiv:1912.02644*, 2019.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Marco Fraccaro, Simon Kamronn, Ulrich Paquet, and Ole Winther. A disentangled recognition and nonlinear dynamics model for unsupervised learning. In *Advances in Neural Information Processing Systems*, pp. 3601–3610, 2017.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Sam Greydanus, Misko Dzamba, and Jason Yosinski. Hamiltonian Neural Networks. *arXiv preprint arXiv:1906.01563*, 2019.

Brian Hall. *Lie groups, Lie algebras, and representations: an elementary introduction*, volume 222. Springer, 2015.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *ICLR*, 2(5):6, 2017a.

Irina Higgins, Arka Pal, Andrei Rusu, Loic Matthey, Christopher Burgess, Alexander Pritzel, Matthew Botvinick, Charles Blundell, and Alexander Lerchner. Darla: Improving zero-shot transfer in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1480–1490. JMLR. org, 2017b.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019.

Diederik P Kingma and Max Welling. Auto-encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sophus Lie. *Vorlesungen über continuierliche Gruppen mit geometrischen und anderen Anwendungen*. BG Teubner, 1893.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Đorđe Miladinović, Muhammad Waleed Gondal, Bernhard Schölkopf, Joachim M Buhmann, and Stefan Bauer. Disentangled State Space Representations. *arXiv preprint arXiv:1906.03255*, 2019.

Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pp. 4036–4044, 2018.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT press, 2017.

Didier Pinchon and Pierre Siohan. Angular parametrization of rectangular paraunitary matrices. 2016.

Antonin Raffin, Ashley Hill, Kalifou René Traoré, Timothée Lesort, Natalia Díaz-Rodríguez, and David Filliat. Decoupling feature extraction from policy learning: assessing benefits of state representation learning in goal based robotics. *arXiv preprint arXiv:1901.08651*, 2019.

Karl Ridgeway. A survey of inductive biases for factorial representation-learning. *arXiv preprint arXiv:1612.05299*, 2016.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pp. 6056–6065, 2019.

Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. Hamiltonian Generative Networks. *arXiv preprint arXiv:1909.13789*, 2019.

Steven Weinberg. *The quantum theory of fields. Vol. 1: Foundations*. Cambridge University Press, 1995.