A Framework for Cheap Universal Approximation in Embodied Systems

Guido Montúfar
Max Planck Institute for Mathematics in the Sciences
montufar@mis.mpg.de

Nihat Ay
Max Planck Institute for Mathematics in the Sciences
Department of Mathematics and Computer Science, Leipzig University
Santa Fe Institute
nay@mis.mpg.de

Keyan Ghazi-Zahedi
Max Planck Institute for Mathematics in the Sciences
zahedi@mis.mpg.de

Abstract

We present a framework for designing cheap control architectures for embodied agents. Our derivation is guided by the classical problem of universal approximation. We explore the possibility of exploiting the agent’s embodiment constraints for a new and more efficient universal approximation of behaviors generated by sensorimotor control. This embodied universal approximation is compared with the classical non-embodied universal approximation. To exemplify our approach, we present a detailed quantitative case study for policy models defined in terms of conditional restricted Boltzmann machines. In contrast to non-embodied universal approximation, which requires an exponential number of parameters, in the embodied setting we are able to generate all possible behaviors with a drastically smaller model, thus obtaining cheap universal approximation. We test and corroborate the theory experimentally with a six-legged walking machine.

Keywords: universal approximation, cheap design, embodiment, sensorimotor loop, conditional restricted Boltzmann machine

1 Introduction

The goal of this article is to provide a framework for selecting the complexity of policy models in accordance with the cheap design principle from embodied artificial intelligence [15]. We introduce the notions of embodiment dimension and embodied universal approximation, which quantify the effective dimension of a system that is subject to sensorimotor (embodiment) constraints and formalize the minimal control paradigm of cheap design in the context of the sensorimotor loop (SML). We substantiate these ideas with theoretical results on the representational capabilities of conditional restricted Boltzmann machines (CRBMs) as policy models for embodied systems, giving bounds on the minimal size of CRBMs that meet, in each case, the following requirements: (1) unconstrained universal approximation of policies, and (2) universal approximation of a set of policies that generates all possible behaviors given the embodiment dimension. In order to test our theory, we present an experimental study with a six-legged walking robot, and find a clear corroboration of our theorems.
The causal structure of the SML [7, 1] describes a type of partially observable Markov decision process (POMDP), where an embodied agent chooses actions on the basis of noisy partial observations of its environment. An illustration of this causal structure is given in Fig. 1A. We aim at optimizing the design of policy models for controlling these processes. One aspect of the optimal design problem is addressed by working out the optimal complexity of the policy model. For example, a well known problem in this direction is concerned with universal approximation in the context of neural networks, see e.g., [6], and the minimal number of hidden units required for this purpose. In most realistic scenarios, universal approximation is out of question, since it demands an enormous number of parameters – many more than actually needed. In this paper we reconsider the universal approximation problem by exploiting embodiment constraints and restrictions in the desired behavioral patterns. This all, using the minimal possible number of parameters, is what we refer to as cheap design.

Restricted Boltzmann machines (RBMs) [18, 4] are recurrent neural networks with bipartite interactions between visible and hidden units. They define products of experts probability models and have efficient training algorithms [5]. RBMs are well known in the context of feature learning, where they are used to infer distributed representations of data and to train the layers of deep neural networks. The theory and in particular the expressive power of RBM probability models has been studied in numerous papers, e.g., [9, 12, 14, 11], and is much better understood than that of CRBMs. CRBMs are defined by clamping an input subset of the visible units of an RBM. The right panel of Fig. 1B illustrates the causality diagram. Conditional models of this kind have found a wide range of applications, e.g., in classification, collaborative filtering, and motion modeling, see [8, 16, 19, 20], and have proven useful as policy models in reinforcement learning settings [17]. Although estimating the probability distributions represented by RBMs is hard [10], approximate samples can be generated easily by finite Gibbs sampling. CRBMs can model non-trivial conditional distributions on high dimensional input-output spaces using relatively few parameters, and their complexity can be adjusted by simply increasing or decreasing the number of hidden units. Hence these models provide a well-suited platform for investigating SML control problems.

A CRBM conditional distribution can be viewed as a collection of RBM probability distributions with shared parameters, with one RBM distribution for each possible input value. The sharing of parameters draws a substantial distinction from RBMs. We will discuss both, direct implications from known results on the expressive power of RBMs, and also non-trivial independent results for CRBMs.

Sec. 2 contains definitions around the SML. Sec. 3 presents the notions of embodiment dimension and universal approximation, which we use to quantify and enforce dimensionality reduction. Sec. 4 contains our theoretical analysis of CRBM models, including results on non-embodied and embodied universal approximation. Sec. 5 puts the theory to the test in a realistic robot control problem. Sec. 6 offers our conclusions.

2 The Causal Structure of the Sensorimotor Loop

The SML is a basic structure to describe the causal interaction of an agent with its environment. Typically, the agent has access to (partial) probabilistic information about its environment, based on which it chooses actions. The agent’s actions influence the environment, thereby closing the loop. The loop is specified by Markov transition kernels describing the agent’s observations, selection of actions, and the world’s dynamics. More specifically, these kernels are

\[ \beta: \mathcal{W} \rightarrow \Delta_{\mathcal{S}}, \quad \pi: \mathcal{S} \rightarrow \Delta_{\mathcal{A}}, \quad \gamma: \mathcal{W} \times \mathcal{A} \rightarrow \Delta_{\mathcal{W}}, \]

(1)

where \( \mathcal{W}, \mathcal{S}, \) and \( \mathcal{A} \) denote the state spaces of the world, the sensors, and the actuators, respectively, which are assumed to be finite in what follows. Furthermore, the set of all probability distributions on a finite set \( \mathcal{X} \) is denoted by \( \Delta_{\mathcal{X}} \), and the set of kernels \( \mathcal{W} \rightarrow \Delta_{\mathcal{X}} \) or \( |\mathcal{W}| \times |\mathcal{X}| \) row-stochastic matrices by \( \Delta_{\mathcal{X}} \).

The kernel \( \beta \) describes the probabilities of sensor values from the set \( \mathcal{S} \), given the current state of the world \( \omega \in \mathcal{W} \). The kernel \( \pi \) is the agent’s policy, which consists of probabilities of selecting actuator values (loosely speaking actions) \( a \in \mathcal{A} \), given the current sensor values \( s \in \mathcal{S} \). As an example of a policy, consider a robot whose task is to navigate a territory. If the robot detects an obstacle ahead, then the policy could choose between the actions ‘turn left’ and ‘turn right’.
with probability $\frac{1}{2}$. The kernel $\gamma$ describes the probabilistic state-transitions of the world under the influence of the agent’s actions.

Within this framework, the agent has direct access to $\pi$ (e.g., by learning), but it does not have direct access to $\beta$ nor $\gamma$. The latter two kernels represent sensor processes and physical properties, while the first represents the brain or controller of the agent. Note that, in contrast, in a traditional Markov decision process (MDP) the agent has full access to the state of the world, and $\mathcal{S}$ is identified with $\mathcal{W}$.

In the next section we discuss how the physical constraints of the agent, given by $\beta$ and $\gamma$, lead to the existence of low-dimensional policy spaces which are just as expressive as the full policy space, in terms of the world’s dynamics or the observable agent’s behavior. In the subsequent Sec. 4 we study policies represented in terms of RBMs, and demonstrate the dimensionality reduction in terms of the number of hidden units that suffices for the agent to perform in any achievable or interesting way.

### 3 Non-Embodied vs. Embodied Universal Approximation

Given the two kernels $\beta$ and $\gamma$, we are interested in the possible behaviors that the system can exhibit for the different choices of the policy. The behavior is determined by the world-state Markov transition kernel

$$ p^\pi(w'|w) = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \beta(s|w)\pi(a|s)\gamma(w'|a, w), \quad \text{for all } w, w' \in \mathcal{W}. \quad (2) $$

We denote by $\Delta^{(\beta, \gamma)} \subseteq \Delta^\mathcal{S}$ the set of all such kernels for fixed $\beta$ and $\gamma$, for all possible choices of the policy $\pi \in \Delta^\mathcal{P}$. This set may be relatively small, since not all world-state transitions may be possible in a single time step and, moreover, several policies may lead to the same transitions. In view of this, we ask the following question related to the notion of universal approximation:

**Problem 1.** What is the smallest model $\mathcal{M} \subseteq \Delta^{\mathcal{M}}$ within a given family $\mathcal{M}$ of policy models, such that each element of $\Delta^{(\beta, \gamma)}$ is realized by a policy from $\mathcal{M}$? Here $\mathcal{M} \subseteq \Delta^\mathcal{S}$ denotes the set of policies that can be approximated arbitrarily well by policies from $\mathcal{M}$.

In Sec. 4 we will consider a family $\mathcal{M}$ given by CRBM policy models. In order to address Problem 1 we need to study the way in which different policies lead to different world-state transition kernels. We develop the necessary formalism in the following.

First, we express Eq. (2) in a vector times matrix form: we view $\pi$ and $p^\pi$ as row vectors (writing the rows of the stochastic matrices one after the other into a long row vector), and view $(\beta(s|w)\gamma(w'|a, w))_{s,a,w,w'}$ as a matrix (with row-index $(s, a)$ and column-index $(w, w')$), which we denote the *embodiment matrix*, given by

$$ E^{(\beta, \gamma)} := \left( \beta^\top \otimes \gamma_1 \cdots \beta^\top \otimes \gamma_\mathcal{S} \right) \in \mathbb{R}^{(\mathcal{P} \times \mathcal{A}) \times (\mathcal{W} \times \mathcal{W})}, \quad (3) $$

where $\beta_w := \beta(\cdot|w)$ is the $w$-th row of $\beta$ and $\gamma_w := \gamma(\cdot|\cdot, w)$ is the $(\mathcal{A} \times \mathcal{W})$ stochastic matrix arising from $\gamma$ by fixing the world-state input to $w$. Here $\top$ denotes vector transposition and $\otimes$ the Kronecker product $(a_{ij})_{i,j} \otimes (b_{kl})_{k,l} := (a_{ij}b_{kl})_{ik,jl}$. With these definitions, we can write Eq. (2) as $p^\pi = \pi \cdot E^{(\beta, \gamma)}$, and

$$ \Delta^{(\beta, \gamma)} = \Delta_{\mathcal{M}} \cdot E^{(\beta, \gamma)}. \quad (4) $$
In words, the set of realizable world-state transition kernels is the projection of the policy polytope to the column-space of the embodiment matrix. We call \( d^{(\beta,\gamma)} := \dim(\Delta^{(\beta,\gamma)}) \) the embodiment dimension; it captures the effective dimension of the system, given the embodiment constraints \( \beta \) and \( \gamma \). The embodiment dimension is upper bounded by the rank of the embodiment matrix.

**Definition 2.** A policy model \( \mathcal{M} \subseteq \Delta_\mathcal{S}^{\mathcal{F}} \) is an embodied universal approximator iff

\[
\mathcal{M} \cdot E^{(\beta,\gamma)} = \Delta^{(\beta,\gamma)} = \Delta_\mathcal{S}^{\mathcal{F}} \cdot E^{(\beta,\gamma)}.
\]  

(5)

In contrast, the policy model is a non-embodied universal approximator iff \( \mathcal{M} = \Delta_\mathcal{S}^{\mathcal{F}} \).

In particular, any non-embodied universal approximator satisfies \( \dim(\mathcal{M}) = \dim(\Delta_\mathcal{S}^{\mathcal{F}}) = |\mathcal{S}| - 1 \). In most cases of practical interest \( \dim(\mathcal{M}) \) is at most equal to the number of parameters of \( \mathcal{M} \), and hence, for large input-output spaces (e.g., \( \{0,1\}^k \) and \( \{0,1\}^n \) for \( k \) binary sensors and \( n \) binary actuators), non-embodied universal approximation requires a prohibitive number of parameters. In the embodied case, the situation can be drastically different. The number of parameters of an embodied universal approximator can be as low as the embodiment dimension. Here is an example:

**Example 3.** Let \( F^1, \ldots, F^d \in \mathbb{R}^{(|\mathcal{S}| \times \mathcal{F})} \) be a basis of the column-space of the embodiment matrix modulo vectors whose values are independent of \( s \) for any fixed \( s \). Then the exponential family of policies, defined by

\[
\pi_\theta(a|s) = \frac{\exp(\sum_{i \in [d]} \theta_i F^i_{(s,a)})}{\sum_{a' \in \mathcal{A}} \exp(\sum_{i \in [d]} \theta_i F^i_{(s,a')})}, \quad \text{for all } a \in \mathcal{A} \text{ and } s \in \mathcal{S}, \quad \text{for all } \theta \in \mathbb{R}^d,
\]  

(6)

is an embodied universal approximator of dimension \( d = d^{(\beta,\gamma)} \) (details can be found in App. A).

The previous example shows that, when the column-span of the embodiment matrix is known, it is possible to construct an embodied universal approximator as a smooth manifold of dimension equal to the embodiment dimension. The problem with such a construction is that it is very specific to the kernels \( \beta \) and \( \gamma \) that define the embodiment matrix, which are not directly accessible to the agent, as mentioned in the previous section. Therefore, we are interested in more generic constructions of small embodied universal approximators, depending only on the embodiment dimension.

Consider a system with embodiment dimension \( d \). Geometrically, the embodiment matrix projects the policy polytope into a ‘squeezed’ polytope of dimension \( d \). This implies that the result of applying any given policy from \( \Delta_\mathcal{S}^{\mathcal{F}} \) can be achieved equally well by applying a policy from a \( d \)-dimensional face of \( \Delta_\mathcal{S}^{\mathcal{F}} \). See Fig. 2B for an illustration. On the other hand, every \( d \)-dimensional face of \( \Delta_\mathcal{S}^{\mathcal{F}} \) consists of policies that have at most \( |\mathcal{S}| + d \) non-zero entries (see App. A). Hence any policy model that contains all policies with \( |\mathcal{S}| + d \) non-zero entries is an embodied universal approximator.

Let us elaborate this a bit further. Embodiment or also behavioral constraints generically lead to a situation where only a relatively small subset \( \mathcal{S} \subseteq \mathcal{S} \) of sensor values is observed. In such a situation, we only need to model the policy for the sensor values \( \mathcal{S} \). Assume that a certain interesting behavior takes place within the restricted set of world states \( \mathcal{W}_S := \{ w \in \mathcal{W} : \text{supp}(\beta(\cdot|w)) \subseteq \mathcal{S} \} \). For these world states, the measurement by \( \beta \) always produces sensor values in \( \mathcal{S} \), and the policy rows with indices not in \( \mathcal{S} \) do not play any role. Let \( E^{(\beta,\gamma),S} \in \mathbb{R}^{(|\mathcal{S}| \times \mathcal{F}) \times (\mathcal{W}_S \times \mathcal{W})} \) denote the restriction of the embodiment matrix to the columns with \( w \in \mathcal{W}_S \). We define an embodied universal approximator on \( \mathcal{S} \) as a model \( \mathcal{M} \subseteq \Delta_\mathcal{S}^{\mathcal{F}} \) for which \( \mathcal{M} \cdot E^{(\beta,\gamma),S} = \Delta_\mathcal{S}^{\mathcal{F}} \cdot E^{(\beta,\gamma),S} \). This means that the model is powerful enough to control any behavior on \( \mathcal{W}_S \) just as well as the entire policy polytope. We write \( d^{(\beta,\gamma),S} := \dim(\Delta_\mathcal{S}^{\mathcal{F}} \cdot E^{(\beta,\gamma),S}) \) for the corresponding embodiment dimension.

**Theorem 4.** Any model \( \mathcal{M} \subseteq \Delta_\mathcal{S}^{\mathcal{F}} \) with the following property is an embodied universal approximator on \( \mathcal{S} \): for every policy \( \pi \in \Delta_\mathcal{S}^{\mathcal{F}} \) whose \( \mathcal{S} \)-rows have a total of \( |\mathcal{S}| + d^{(\beta,\gamma),S} \) or less non-zero entries, there exists a policy \( \pi^* \in \mathcal{M} \) with \( \pi^*(\cdot|s) = \pi^*(\cdot|s) \) for all \( s \in \mathcal{S} \).

This theorem states that for embodied universal approximation it suffices to approximate policies which, for a relevant set of sensor values, assign positive probability only to a limited number of actions.
Before we move on, let us briefly comment on why we can expect the embodiment dimensions to be relatively small, possibly much smaller than the dimension of the policy polytope. The embodiment matrix $E^{(\beta, \gamma)}$ is defined as the block-wise Kronecker product of the block-matrices $(\beta_i^T \cdot \ldots \cdot \beta_i^{|W|})$ and $(\gamma_1 \cdot \ldots \cdot \gamma_1^{|W|})$. Unfortunately, there is no formula for the rank of such a product in terms of the block ranks. A simple hard bound can be given as $\text{rank}(E^{(\beta, \gamma)} \cdot S) \leq \text{rank}(E^{(\beta, \gamma)}) \leq \text{rank}(\beta_i^T \cdot \ldots \cdot \beta_i^{|W|}) \cdot \text{rank}(\gamma_1 \cdot \ldots \cdot \gamma_1^{|W|})$. Let us discuss a few reasons why this rank can be very low in embodied systems. One natural assumption is that the sensors are insensitive to a large number of variations of the world-state $w$ and, therefore, that the matrix $\beta$ has many repeated rows. It is also natural to assume that the sensor values contain a certain degree of redundancy, such that the rows of $\beta$ have repeated entries. For the kernel $\gamma$, a natural assumption is that several actions produce the same outcome, such that each matrix $\gamma_w$ has repeated rows. Also, we can assume that for any given world-state $w$, the number of states $w'$ with positive transition probability is very small. For example, a robot in a maze cannot teleportate from a given position to any other, but it can only move to the contiguous positions.

4 A Case Study with CRBMs

A Boltzmann machine (BM) is an undirected stochastic network with binary units, some of which may be hidden. It defines probabilities for the joint states of its visible units, given by the relative frequencies at which these states are observed, asymptotically, depending on the network parameters. At each time $t \in \mathbb{N}$, this machine selects a unit at random, say unit $i$, and updates its state according to a Bernoulli draw with success probability $\text{sigm}(\sum_j W_{ji} x_j + b_i)$, where $\text{sigm}(c) := \frac{1}{1 + \exp(-c)}$. $x_j \in \{0, 1\}$ is the current state of unit $j$, $W_{ji} \in \mathbb{R}$ is an interaction weight attached to the unit-pair $(j, i)$, and $b_i \in \mathbb{R}$ is a bias weight attached to unit $i$. In the limit of infinite time, each joint state $x = (x_V, x_H)$ of the network’s visible and hidden units occurs with a relative frequency described by the Gibbs-Boltzmann distribution $p(x) = \frac{1}{Z} \exp(-\mathcal{H}(x))$ with energy function $\mathcal{H}(x) = \sum_{i,j} x_i W_{ij} x_j + \sum_i b_i x_i$ and normalization $Z = \sum_{x'} \exp(-\mathcal{H}(x'))$. The probabilities of the visible states are given by marginalizing out the states of the hidden units, i.e., by $p(x_V) = \sum_{x_H} p(x_V, x_H)$.

Now, an RBM is a BM with the restriction that there are no interactions between the visible units nor between the hidden units, such that $W_{ij} \neq 0$ only when unit $i$ is visible and unit $j$ is hidden. The probabilities of the visible states are given by $p(x_V) = \sum_{x_H} \frac{1}{Z} \exp(\sum_{i \in V, j \in H} W_{ij} x_i x_j + \sum_{i \in V} b_i x_i + \sum_{j \in H} c_j x_j)$. 
We will use the shorthand notations $\Delta_n := \Delta_{\{0,1\}^n}$ and $\Delta_n^k := \Delta_{\{0,1\}^k}$. For example, $\Delta_n^k$ is a $2^k$-cube. As any multivariate model of probability distributions, RBMs define models of conditional distributions:

**Definition 5.** The conditional restricted Boltzmann machine (CRBM) model with $k$ input, $n$ output, and $m$ hidden units, denoted $\text{RBM}_{n,m}^k$, is the set of all conditional distributions in $\Delta_n^k$ that can be written as
\[
p(x|y) = \frac{1}{Z(W, b, V + c)} \sum_{z \in \{0,1\}^m} \exp(z^TWx+z^TVy+b^Tx+c^Tz), \quad \forall x \in \{0,1\}^n, y \in \{0,1\}^k,
\]
where $Z(W, b, V + c)$ normalizes the probability distribution $p(\cdot|y)$. Here, $y$, $x$, and $z$ are state vectors of the input, output, and hidden units, respectively.

In practice one does not compute the full conditional distributions, as this involves intractable sums with exponentially many terms. Instead, one uses the CRBM to sample outputs given inputs in the following way. (1) Take an input state $y$. (2) Update the state $z$ of all hidden units fixing the state $y$ of the input units and the current state $x$ of the output units. (3) Update the state $x$ of all output units fixing the state $y$ of the input units and the current state $z$ of the hidden units. (4) Repeat the last two steps as many times as desired. (5) Output the current state $x$ of the output units. The state updates in this procedure are cheap, as they involve only Bernoulli draws with tractable success probabilities. Also note that the bipartite interaction structure of the network allows to update the states of all visible or hidden units in parallel. The same sampling mechanism can be used to approximate expectation values and CRBM policy gradients.

The conditional model $\text{RBM}_{n,m}^k$ has $(n+k)m+n+m$ parameters. A bias term $a^Ty$ for the input units does not appear in the definition, as it would cancel out with the normalization function $Z$. When there are no input units, i.e., $k = 0$, the conditional model $\text{RBM}_{n,m}^k$ reduces to the restricted Boltzmann machine probability model with $n$ visible and $m$ hidden binary units, which we denote by $\text{RBM}_{n,m}$.

The model $\text{RBM}_{n,m}^k$ can be interpreted as a collection of $2^k$ RBMs with shared parameters. For each input value $y$, the distribution $p(\cdot|y)$ is the distribution represented by $\text{RBM}_{n,m}$ for the parameters $W$, $b$, $(V + c)$. In particular, all $p(\cdot|y)$ are distributions from the model $\text{RBM}_{n,m}$ with the same interaction weights $W$, the same bias weights $b$ for the visible units, but with different bias weights $(V + c)$ for the hidden units. The joint behavior of these RBMs with shared parameters is not trivial. The direct interpretation of $\text{RBM}_{n,m}^k$ is that it represents block-wise normalized versions of the probability distributions represented by $\text{RBM}_{n+k,m}$. Namely, each probability distribution $p \in \text{RBM}_{n+k,m}$ defines a tuple of $2^k$ probability distributions $p(\cdot|y)$, for all possible states $y$ of $k$ of its visible units, and this tuple is an element of $\text{RBM}_{n,m}^k$.

### 4.1 Non-Embodied Universal Approximation

In this section we ask for the minimal number of hidden units $m$ for which the model $\text{RBM}_{n,m}^k$ can approximate every conditional distribution from $\Delta_n^k$ arbitrarily well. Note that each conditional distribution $p(x|y)$ can be identified with the set of joint distributions of the form $r(x, y) = q(y)p(x|y)$, with strictly positive marginals $q(y)$. By fixing $q(y)$ equal to the uniform distribution over $\mathcal{X}$, we obtain an identification of $\Delta_{\mathcal{X}^k}$ with $\mathcal{X} \Delta_{\mathcal{Y} \times \mathcal{Z}}$. See Fig. 2A. In particular, we have that universal approximators of joint probability distributions define universal approximators of conditional distributions.

The preceding observation allows us to translate results on the expressive power of RBMs to corresponding results for CRBMs. For example, we know that $\text{RBM}_{n+k,m}$ is a universal approximator whenever $m \geq \frac{1}{2}2^{k+n} - 1$, see [12], and therefore:

**Proposition 6.** The model $\text{RBM}_{n,m}^k$ can approximate every conditional distribution from $\Delta_n^k$ arbitrarily well whenever $m \geq \frac{1}{2}2^{k+n} - 1$.

This improves previous a previous result that appeared in [21]. On the other hand, since conditional models do not need to model the input-state distributions, in principle it is possible that $\text{RBM}_{n,m}^k$ is
a universal approximator of conditional distributions even if \( \text{RBM}_{n+k,m} \) is not a universal approximator of probability distributions. Therefore, we also consider an improvement of Proposition 6 that does not follow from corresponding results for RBM probability models:

**Theorem 7.** The model \( \text{RBM}^k_{n,m} \) can approximate every conditional distribution from \( \Delta^k_n \) arbitrarily well whenever \( m \geq \frac{1}{2} 2^{k(n-1)} = \frac{1}{2} 2^{k+n} - \frac{1}{2} 2^k \).

The full statement and proof of the theorem are quite technical, and thus we refer the interested reader to [13]. At this point let it suffice to say that the bound on \( m \) decreases with increasing \( k \), such that approximately the prefactor \( \frac{1}{2} \) decreases to \( \frac{1}{4} \) when \( k \) is large enough.

Theorem 7 represents a substantial improvement of Proposition 6 in that it reflects the structure of the policy polytope \( \Delta^k_n \) as a \( 2^k \)-fold product of the \( (2^n - 1) \)-simplex \( \Delta_n \), in contrast to the proposition’s bound, which rather reflects the structure of the \( (2^{n+k} - 1) \)-dimensional joint probability simplex \( \Delta_{k+n} \).

As expected, the asymptotic behavior of this result is exponential in the number of input and output units. We believe that the result is reasonably tight, although some improvements may still be possible. A crude lower bound can be obtained by comparing the number of parameters with the dimension of the policy polytopes (details in [13]):

**Proposition 8.** If the model \( \text{RBM}^k_{n,m} \) can approximate every policy from \( \Delta^k_n \) arbitrarily well, then necessarily \( m \geq \frac{1}{n+k+1} (2^k (2^n - 1) - n) \).

### 4.2 Embodied Universal Approximation

By Theorem 4, we can achieve embodied universal approximation by approximating only policies with a limited number of non-zero entries. Furthermore, as mentioned earlier, if we only care about a subset \( S \subseteq \mathcal{S} \) of sensor values, due to behavioral or embodiment constraints, we can restrict the policy space \( \Delta^S_{\mathcal{S}} \) to \( \Delta^S_{\mathcal{S}} \). This means that the number of policy entries that we need to model is given by the number of interesting sensor values plus the corresponding embodiment dimension. On the other hand, we can use each hidden unit of a CRBM to model each relevant non-zero entry of the policy (details in App. B).

**Theorem 9.** The model \( \text{RBM}^k_{n,m} \) is an embodied universal approximator on \( S \) whenever \( m \geq |S| + d^{(\beta, \gamma),S} - 1 \).

This theorem corresponds to Theorem 4. It gives a bound for the number of hidden units of CRBMs that suffices to obtain embodied universal approximation, depending on the embodiment and behavioral constraints of the system. In general, this will be much smaller than the exponential bound from Theorem 7. We will test this bound in the context of particular behavioral constraint in the next section.

### 5 Experiments with a Hexapod

In the previous sections we have derived a theoretical bound for the complexity of a CRBM based policy. In this section, we want to evaluate this bound experimentally. We chose a six-legged walking machine (hexapod) as our experimental platform (see Fig. 3A), because it is a well-studied morphology in the context of artificial intelligence, with one of its first appearances as Ghengis in [2]. The purpose of this section is *not* to develop an optimal walking strategy for this system based on a CRBM. Contrary, this morphology was chosen, because the *tripod gait* is known to be one of the optimal locomotion behaviors, which can be implemented efficiently in various ways. This said, learning a control for a hexapod is not trivial, and hence, a good test bed to evaluate the previously derived bounds for the complexity of CRBM based policies.

The experiments in this section were conducted with YARS [22], which is a mobile robot simulator based on the bullet physics engine [3]. The detailed description of the hexapod can be found in App. C. The learning of the CRBMs requires a data set of sensor and actuator values. For this purpose, the hexapod was controlled for \( 10^6 \) time steps with an open loop controller that applies phase shifted sinus oscillations to the actuators (see Fig. 3C). Each leg has two actuators and two
sensors, resulting in a system with 12 degrees of freedom (DOFs). In the following we explain how the data was processed and then we discuss the results.

Figs. 3B/C show the recorded data for the sensors \( \bar{S}_i^t \) and the actuators \( \bar{A}_i^t \) for the first second (sampling rate is 100Hz). The continuous sensors and actuators were each binned equidistantly into 16 bins corresponding to four binary units. This gives two combined random variables \( S_i^t, A_i^t \), each with a total of \( 16^{12} \) possible values corresponding to 48 input and 48 output binary units. The resulting (relabeled) data is shown in Fig. 3D. Let us briefly recall the theoretical bounds for the cardinality of \( S \) divided by Mathematica 9. For the conservative estimation we obtained \( \text{MatrixRank} = 159 \) plus the embodiment dimension \( d = d(\delta, \gamma) \cdot |S| \). Here we approximated \( d \) by the rank of the internal world model \( \delta(s'|a, s) \). This is simplistic, in general, but reasonable in the case of a tripod gait on a plane ground. For a validation, the reader is referred to App. C.3. By Theorem 9, a sufficient number of hidden units is given by \( m = |S| + d - 1 \).

In order to estimate the cardinality of \( S \), we counted the number of occurrences for each bin and removed all rows of the data for which the sensor bin count was smaller than 20 or 140. This corresponds to removing about 10% and 15% of the data, which gives the conservative and progressive estimates \( |S| = 159 \) and \( |S| = 59 \), respectively. Fig. 4A shows one of the resulting histograms. We estimated the world model \( \delta(s'|a, s) \) by counting the occurrences of \( S^{t+1} \), \( S^t \), and \( A^t \) in the pruned binned data for both cases. The matrix rank was calculated with the function \text{MatrixRank} \text{ provided by Mathematica 9. For the conservative estimation we obtained} \( d \leq 119 = \text{rank}(\delta_{|S|=159}) \) and for the progressive estimation \( d \leq 59 = \text{rank}(\delta_{|S|=59}) \).

As a result, the hexapod should be able to reproduce the walking behavior with the following CRBM configuration: \( n = k = 48 \) and \( m \) between 59+59−1 ≈ 120 and 159+119−1 ≈ 280. To evaluate the quality of this estimation, we trained CRBMs with \( m = 1, 10, 20, 30, \ldots, 1000 \) hidden units. We used contrastive divergence [5] with 15 up-down iterations (CD_1) and meta-parameters set to: epochs = 50000, batch = 500, alpha = 1.0, momentum = 0.1, perturbation = 0.1. These values were determined empirically before the scan over \( m \) was performed. Each experiment was repeated 30 times with different (randomized) parameter initializations. Fig. 4B shows the resulting plot for the smoothed average over the best five CRBMs for each \( m \). The smoothing is generated by \( y_{t+1} = y_t + \zeta(x_{t+1} - y_t), \) where \( x \) is the data vector, \( y_0 = x_0, \) and \( \zeta = 0.2 \). The embedded plot shows the raw curve and the standard deviation. The plot shows that the performance of the CRBMs
has converged for $m = 280$ and is already close to optimal for $m = 120$. Hence, the theoretical bound derived in the previous section is tight enough to be used in practical applications.

Given that Theorem 9 does not depend on the particular behavior but only on its support, we had expected that the theoretically determined number of hidden units $m$ would significantly exceed the number of hidden units actually required to control the hexapod. Therefore, we find it remarkable that the experimentally estimated bound aligns so well with the theoretical bound. This indicates the sharpness of our theoretical bound. Moreover, we conjecture that most behaviors will require the number of hidden units determined by Theorem 9.

6 Conclusions

We present an approach for implementing cheap design principles within the embodied artificial intelligence paradigm. Our framework focusses on the observable behavior of an agent in the SML, and allows us to exploit embodiment for dimensionality reduction. In this way, we are able to define low dimensional policy spaces that can generate all possible observable behaviors, thus achieving cheap universal approximation. This approach aligns with the idea that interesting data lies on a low dimensional manifold, and that, therefore, universal approximation is not important, but, instead, what matters is the approximation of the interesting low dimensional manifold. This is precisely the meaning of embodied universal approximation in the context of the SML. We take CRBMs as a platform of study, for which we obtain non-trivial universal approximation results in both the non-embodied and the embodied settings. While the former requires an enormous number of hidden units (exponentially many in the number of input and output units), the latter can be achieved using essentially only as many hidden units as the effective dimension of the system (the number of interesting sensor values plus the embodiment dimension), independently of the specific embodiment constraints and the number of input and output units. Experiments conducted on a walking machine demonstrate the practical utility of our theoretical analysis. To the best of our knowledge, the formalism and results thus presented are amongst the first quantitative contributions to cheap design.

Acknowledgment

We would like to acknowledge support for this project from the DFG Priority Program Autonomous Learning (DFG-SPP 1527). G. M. and K. G.-Z. would like to thank the Santa Fe Institute for hosting them during the initial work on this article.

References

[1] N. Ay and K. Zahedi. On the causal structure of the sensorimotor loop. In M. Prokopenko, editor, Guided Self-Organization: Inception. Springer, 2014.
[2] R. A. Brooks. A robot that walks; emergent behaviors from a carefully evolved network. Technical report, Cambridge, MA, USA, 1989.

[3] E. Coumans. Bullet physic sdk manual. www.bulletphysics.org, 2012.

[4] Y. Freund and D. Haussler. Unsupervised Learning of Distributions of Binary Vectors Using Two Layer Networks. Technical report. Computer Research Laboratory, University of California, Santa Cruz, 1994.

[5] G. E. Hinton. Training products of experts by minimizing contrastive divergence. Neural Computation, 14(8):1771–1800, 2002.

[6] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. Neural Networks, 2:359–366, 1989.

[7] A. S. Klyubin, D. Polani, and C. L. Nehaniv. Tracking information flow through the environment: Simple cases of stigmergy. In J. Pollack, editor, Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems. MIT Press, 2004.

[8] H. Larochelle and Y. Bengio. Classification using discriminative restricted Boltzmann machines. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, Proceedings of the 25th International Conference on Machine Learning (ICML 2008), volume 307, pages 536–543, 2008.

[9] N. Le Roux and Y. Bengio. Representational power of restricted Boltzmann machines and deep belief networks. Neural Computation, 20(6):1631–1649, June 2008.

[10] P. M. Long and R. A. Servedio. Restricted Boltzmann machines are hard to approximately evaluate or simulate. In J. Fürnkranz and T. Joachims, editors, Proceedings of the 27th International Conference on Machine Learning (ICML 2010), pages 703–710. Omnipress, 2010.

[11] J. Martens, A. Chattopadhya, T. Pitassi, and R. Zemel. On the expressive power of restricted Boltzmann machines. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, NIPS 26, pages 2877–2885, 2013.

[12] G. Montúfar and N. Ay. Refinements of universal approximation results for deep belief networks and restricted Boltzmann machines. Neural Computation, 23(5):1306–1319, 2011.

[13] G. Montúfar, N. Ay, and K. Ghazi-Zahedi. Expressive power of conditional restricted Boltzmann machines. arXiv preprint arXiv:1402.3346, 2014.

[14] G. Montúfar, J. Rauh, and N. Ay. Expressive power and approximation errors of restricted Boltzmann machines. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger, editors, NIPS 24, pages 415–423, 2011.

[15] R. Pfeifer and J. C. Bongard. How the Body Shapes the Way We Think: A New View of Intelligence. The MIT Press (Bradford Books), Cambridge, MA, 2006.

[16] R. Salakhutdinov, A. Mnih, and G. E. Hinton. Restricted Boltzmann machines for collaborative filtering. In Proceedings of the 24th International Conference on Machine Learning (ICML 2007), pages 791–798, 2007.

[17] B. Sallans and G. E. Hinton. Reinforcement learning with factored states and actions. J. Mach. Learn. Res., 5:1063–1088, 2004.

[18] P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. In Symposium on Parallel and Distributed Processing, 1986.

[19] I. Sutskever and G. E. Hinton. Learning multilevel distributed representations for high-dimensional sequences. Proceeding of the 11th International Conference on Artificial Intelligence and Statistics, 2007.

[20] G. W. Taylor, G. E. Hinton, and S. Roweis. Modeling human motion using binary latent variables. In NIPS 19, pages 1345–1352. MIT Press, 2007.

[21] L. van der Maaten. Discriminative restricted Boltzmann machines are universal approximators for discrete data. Technical Report EWI-PRB TR 2011001, Delft University of Technology, 2011.

[22] K. Zahedi, A. von Twickel, and F. Pasemann. Yars: A physical 3d simulator for evolving controllers for real robots. In S. Carpin, I. Noda, E. Pagello, M. Reggiani, and O. von Stryk, editors, SIMPAR 2008, LNAI 5325, pages 71–82. Springer, 2008.
Appendix

A Embodied Universal Approximation

Here we expand on the definitions made in Sec. 3.

Consider the embodied universal approximation Problem 1. Since we are more used to solve linear equations written in terms of vector-matrix products than equations written in terms of matrix-matrix products, we first rewrite the embodied universal approximation problem in terms of a vector-matrix product.

Let $\mathcal{W} = \{1, \ldots, |\mathcal{W}|\}$, $\mathcal{S} = \{1, \ldots, |\mathcal{S}|\}$. Consider the row vectors $\beta_w := \beta(\cdot|w)$, $\pi_s := \pi(\cdot|s)$, and the matrices $\gamma_w := \gamma(\cdot|w)$ with rows $\gamma_w^a := \gamma(\cdot|a, w)$. For any policy $\pi \in \Delta_\mathcal{S}$, the resulting world-state transition kernel is given by $p^\pi(w'|w) := \sum_s \sum_a \beta(s|w)\pi(a|s)\gamma(\cdot|a, w) \in \Delta_\mathcal{W}$. This is a stochastic matrix with rows $p^\pi_w := p^\pi(\cdot|w)$ for all $w \in \mathcal{W}$.

We can write the world-state transition kernel (stretched into a row vector) as the result of multiplying the policy $\pi$ (also stretched into a row vector) times the embodiment matrix, as

$$ (p^\pi_1, \ldots, p^\pi_{|\mathcal{W}|}) = (\pi_1, \ldots, \pi_{|\mathcal{S}|}) \cdot E^{(\beta, \gamma)}. $$

This shows that the set of kernels $p^\pi$ realizable for all possible $\pi$ is equal to the projection of the policy polytope $\Delta_\mathcal{S}$ by the embodiment matrix. Linear projections of polytopes are again polytopes. Hence the set of realizable kernels $\Delta^{(\beta, \gamma)}$ is a polytope.

**Embodiment matrix.** Let us have a look at the embodiment matrix

$$ E^{(\beta, \gamma)} := \left( \beta_1^\top \otimes \gamma_1 \right) \cdots \left( \beta_{|\mathcal{W}|}^\top \otimes \gamma_{|\mathcal{W}|} \right). $$

We note that this can be equivalently written as

$$ E^{(\beta, \gamma)} = \beta^\top \otimes \gamma, $$

where $\otimes$ denotes the Khatri-Rao product, which is defined as the block-wise Kronecker-product of two block matrices with the same number of blocks.

**Embodiment dimension.** Let us make a few comments in relation to the embodiment dimension and the rank of the embodiment matrix. Since policies have row sums equal to one, we can consider an ‘affine’ version of Eq. (7), and write

$$ (\pi_1, \ldots, \pi_{|\mathcal{S}|}) \cdot (\beta^\top \otimes \gamma) = (\tilde{\pi}_1, \ldots, \tilde{\pi}_{|\mathcal{S}|}) \cdot (\beta^\top \otimes \gamma) + C, $$

where $C := (\gamma_1, \ldots, \gamma_{|\mathcal{W}|})$ is a vector that does not depend on the policy; furthermore, $\tilde{\pi}_s = (\pi(a|s))_{a=1,\ldots,|\mathcal{A}|-1}$ denotes $\pi_s$ with the last entry removed, and

$$ \tilde{\gamma}_w = \begin{pmatrix} \gamma_w^1 - \gamma_w^{|\mathcal{A}|} \\ \vdots \\ \gamma_w^{|\mathcal{A}|-1} - \gamma_w^{|\mathcal{A}|} \end{pmatrix} $$

corresponds to removing $C$ from $\gamma_w$.

Unlike $\pi_s$, the entries of $\tilde{\pi}_s$ do not have to add to any particular value. In other words, the entries of the vector $(\tilde{\pi}_1, \ldots, \tilde{\pi}_{|\mathcal{A}|})$ are independent of each other. This shows that the embodiment dimension $d^{(\beta, \gamma)} := \text{dim}(\Delta_\mathcal{S} \cdot E^{(\beta, \gamma)})$ is equal to the rank of $\beta^\top \otimes \gamma$, which is upper bounded by the rank of $\beta^\top \otimes \gamma$, but can be strictly smaller.

**Details to Example 3.** Consider the exponential family $\mathcal{E}_{\mathcal{S} \times \mathcal{A}}$ of probability distributions on $\mathcal{S} \times \mathcal{A}$, whose sufficient statistics matrix $F \in \mathbb{R}((\mathcal{W} \times \mathcal{W}) \times |\mathcal{S}| \times |\mathcal{A}|)$ is composed of two row blocks; the first block $E$, given by the transpose of the embodiment matrix, and the second block $I$, given
by the set of indicators $\mathbb{1}_s$ for all $s \in \mathcal{S}$, defined by $\mathbb{1}_s(s', a') = 1$ for all $(s', a')$ with $s' = s$ and $\mathbb{1}_s(s', a') = 0$ otherwise. This exponential family consists of all distributions of the form
\[
p_s(s, a) = \frac{\exp\left(\sum_i \theta_i F_i(s, a)\right)}{\sum_{s', a'} \exp\left(\sum_i \theta_i F_i(s', a')\right)}, \quad \text{for all } s \in \mathcal{S} \text{ and } a \in \mathcal{A}, \text{ for all } \theta \in \mathbb{R}^{(|\mathcal{W} \times \mathcal{W}| + |\mathcal{S}|)}.
\] (12)

The moment map $F: p \mapsto F \cdot p$ maps the closure of this exponential family bijectively to the convex support polytope (given by the convex hull of the columns of the sufficient statistics matrix), with
\[
F \cdot \mathcal{E}_{\mathcal{S} \times \mathcal{A}} = F \cdot \Delta_{\mathcal{S} \times \mathcal{A}}.
\] (13)

We denote by $\mathcal{E}_{\mathcal{S}}$ the set of $s$-marginals and by $\mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+$ the set of conditionals of joint distributions in $\mathcal{E}_{\mathcal{S} \times \mathcal{A}}$. By the way in which we chose the sufficient statistics matrix (including $I$, for any $p \in \mathcal{E}_{\mathcal{S}}$ we have that $p \cdot \mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+ \subseteq \mathcal{E}_{\mathcal{S} \times \mathcal{A}}$, where the product is understood as computing the joint distributions from the marginal and the conditionals.

Note that
\[
\begin{bmatrix} E \\ I \end{bmatrix} \cdot p \cdot \mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+ = \begin{bmatrix} E \cdot p \cdot \mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+ \\ p \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} E \\ I \end{bmatrix} \cdot p \cdot \Delta_{\mathcal{S} \times \mathcal{A}} = \begin{bmatrix} E \cdot p \cdot \Delta_{\mathcal{S} \times \mathcal{A}} \\ p \end{bmatrix}.
\] (14)

Since the moment map is a bijection on $\mathcal{E}_{\mathcal{S} \times \mathcal{A}}$, using Eq. (13) we obtain $E \cdot p \cdot \mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+ = E \cdot p \cdot \Delta_{\mathcal{S} \times \mathcal{A}}^+$ for all $p$, which in particular implies that
\[
E \cdot \mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+ = E \cdot \Delta_{\mathcal{S} \times \mathcal{A}}^+.
\] (15)

This means that $\mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+$ is an embodied universal approximator. Now, the conditional exponential family $\mathcal{E}_{\mathcal{S} \times \mathcal{A}}^+$ only depends on the row span of the sufficient statistics matrix $F$ up to addition of vectors whose entries are independent of the index $a$ given the index $s$ (due to the row wise normalization of stochastic matrices). This means that Eq. (15) also holds for the conditional exponential family given in Example 3.

We conclude this section with some details about the faces of the policy polytope:

**The faces of $\Delta_{\mathcal{S} \times \mathcal{A}}$.** The policy polytope is a product of simplices:
\[
\Delta_{\mathcal{S} \times \mathcal{A}} = \times_{s \in \mathcal{S}} \Delta_{\mathcal{A}_s},
\] (16)

where each factor corresponds to the set of all possible choices of the probability distributions $\pi(\cdot|s)$ for some $s \in \mathcal{S}$. Each face of a product of simplices is also a product of simplices. The faces of $\Delta_{\mathcal{S} \times \mathcal{A}}$ are the products of the form
\[
\times_{s \in \mathcal{S}} \Delta_{\mathcal{A}_s}, \quad \text{where } \mathcal{A}_s \subseteq \mathcal{A} \text{ for all } s \in \mathcal{S}.
\] (17)

In other words, each face of $\Delta_{\mathcal{S} \times \mathcal{A}}$ corresponds to a choice of positions $\mathcal{A}_s \subseteq \mathcal{A}$ of the non-zero entries of $\pi(\cdot|s)$ for all $s \in \mathcal{S}$. The policies in a given face are the policies which have non-zero entries contained in these positions. The $d$-dimensional faces are those for which $\sum_{s \in \mathcal{S}}(|\mathcal{A}_s| - 1) = d$, meaning that they consist of policies which have at most $\sum_{s \in \mathcal{S}}|\mathcal{A}_s| = |\mathcal{S}| + d$ non-zero entries.

**B** CRBM Embodied Universal Approximation (Proof of Theorem 9)

We will use the following lemma.

**Lemma 10.** Let $\pi$ be a conditional distribution from $\Delta_{\mathcal{S}}^k$ with $N$ non-zero entries, $2^k \leq N \leq 2^{k+n}$. Then $\pi$ can be approximated arbitrarily well by $\text{RBM}^k_{n,m}$ whenever $m \geq N - 1$.

**Proof.** This is because the joint probability model $\text{RBM}^k_{n+k,m}$ can approximate any probability distribution with support of cardinality $m + 1$ arbitrarily well, see [12]. Such distributions intersect the interior of the equivalence class $\mathcal{P}_\pi$ of joint distributions with conditionals $\pi$. Hence $\pi$ is approximated arbitrarily well by $\text{RBM}^k_{n,m}$. \qed
The previous lemma implies the following preliminary version of Theorem 9.

**Proposition 11.** For a system with embodiment dimension \(d^{(\beta, \gamma)}\), the model \(\text{RBM}^{k}_{n, m}\) is an embodied universal approximator whenever \(m \geq 2^{k} + d^{(\beta, \gamma)} - 1\).

**Proof.** By the discussion from Sec. 3, we can achieve embodied universal approximation by approximating only policies with \(|\mathcal{F}| + d^{(\beta, \gamma)}\) non-zero entries. By Lemma 10, this is possible with the claimed number of hidden units. \(\square\)

With these tools at hand, now we provide the proof of the theorem.

**Proof of Theorem 9.** The projection of the policy polytope by the \(S\) embodiment matrix can be regarded as a composition which first projects \(\Delta_{\mathcal{F}}^{S}\) to \(\Delta_{\mathcal{F}}^{S_{A}}\) and then projects \(\Delta_{\mathcal{F}}^{S_{A}}\) by \(E^{(\beta, \gamma)}_{\mathcal{F}}S\). Now we can use the same arguments used in Theorem 11 with the difference that now only need to represent a \(d^{(\beta, \gamma)}S\)-dimensional face of the smaller polytope \(\Delta_{\mathcal{F}}^{S_{A}}\). \(\square\)

## C Details about the Hexapod Experiment

This section gives additional information to the Sec. 5 of the main document (Experiments with a Hexapod). It will begin with a description of the simulated morphology, which is then followed by an explanation how the data was recorded and analyzed.

### C.1 Simulation

The hexapod was simulated with YARS [22], which is a mobile robot simulator based on the bullet physics engine [3]. Each segment of the hexapod is defined by its physical properties (dimension, weight, etc.) and each actuator is defined by its force, velocity and its angular range. In the case of the hexapod shown in Fig. 3A, the main body’s dimension (bounding box) and weight is: 4.4m length, 0.7m width, 0.5 height, 2kg. Each leg consists of three segments (femur, tarsus, tibia), of which the two lower segments (tarus, tibia) are connected by a fixed actuator. The leg segments were freely modeled with respect to the dimensions of an insect leg. The actuator which connects the femur and tarsus (knee actuator) only allows rotations around the local \(y\)-axis of the femur segment (see Fig. 3 of the main document). The maximal deviation for the femur-tarsus actuator limited to \(\omega_{\text{ma-fe}} \in [-\frac{15}{\pi}, \frac{25}{\pi}]\). For actuators, which connect the main body with the femur (ma-fe), the maximal deviation is limited to \(\omega_{\text{ma-fe}} \in [-\frac{35}{\pi}, \frac{35}{\pi}]\). The rotation axis of the ma-fe actuator is limited to the local \(z\)-axis of the main body. In bullet, an actuator is defined by its impulse (set to 1Ns for both actuators) and its maximal velocity (set to 0.75 rad/s \(\approx 42^\circ/s\) for both actuators). It must be noted here, that the sensors and actuators are all mapped onto the interval \([-1, 1]\), which means that a sensor value of \(\tilde{S}_i = 1\) refers to the maximal current deviation of the corresponding knee or shoulder. In the same sense, an actuator value \(\tilde{A}_j = 1\) refers to the motor command to deviate the corresponding knee or shoulder to its maximal position.

### C.2 Estimation of the Support and Embodiment Dimension

In order to calculate the required number of hidden units \(m\) based on Theorem 9, we need to determine the support set of the sensor states and the rank of the internal world model \(\delta\) (for a justification, see App. C.3 below). The first step is to record the sensors \(\tilde{S}_i\) and actuators \(\tilde{A}_j\) for the behavior of interest, which is the tripod gait in this case. Fig. 3B and Fig. 3C show the recorded data of four sensors and actuators for the first second (100 data points). A total of \(10^6\) time steps was recorded for the estimation of the support set and world model’s rank. The sensors and actuators, which values are in the domain \([-1, 1]\) (see App. C.1 above), are individually binned into 16 equidistant bins. All 12 sensors are combined into a single random variable \(S\), and all 12 actuators are combined into a single random variable \(A\), which can both have a total number of \(16^{12}\) values. The first row of Fig. 5 shows the histograms for the sensor state \(S\) (see Fig. 5A) and the actuator state \(A\) (see Fig. 5B). For simplicity, the sensor and actuator bins that have no occurrences in the data are omitted, and the remaining bins are relabeled, which leads to the histograms in Fig. 5 and in the time series shown in Fig. 3D.
Figure 5: Data pruning for the support and rank estimation. Histograms A and B show the original data, C and D the data after removing sensor bins with less than 20 counts, and finally, E and F the data after removing sensor bins with less than 140 counts. An actuator state \( a^t \) was removed from the data, if its corresponding sensor state \( S^t \) had less than 20 or 140 counts (histograms D and F).

The histogram of the sensor values \( s \in \mathcal{S} \) (see Fig. 5A) shows that the estimation of the support and rank will benefit from a pruning as most sensor values have a very low count and can, therefore, be considered to be noise. Fig. 5C shows the histogram for all sensor states \( s \in S \) which have a count of at least 20, and Fig. 5E shows the histogram for all sensor states \( s \in S \) which have a count of at least 140. The corresponding actuator state histograms are shown in Fig. 5D and Fig. 5F. It must be noted that an actuator state \( A^t \) was removed from the data if its sensor state \( S^t \) had less than 20 (Fig. 5D) or 140 (Fig. 5F) occurrences in the recorded data. This is why the frequency of the actuator values \( a \in \mathcal{A} \) changes (compare Fig. 5B with Fig. 5D), but not the number of observed actuator values as in the case of the observed sensor values \( s \in \mathcal{S} \) (compare Fig. 5A with Fig. 5C). The estimation of \( |S| \) is based on the data that is displayed in Fig. 5C/E and it shows that the estimation of the support of the sensor distribution is to some extent arbitrary and worth a deeper investigation. We decided to pick a progressive and conservative value, which lead to \( |S| = 159 \) and \( |S| = 59 \).

The second step was to estimate the rank of the internal world model (see Equation 18), \( \text{rank}(\delta_{|S|=159}) \) and \( \text{rank}(\delta_{|S|=59}) \). For this purpose, the world models \( \delta_{|S|=159} \) and \( \delta_{|S|=59} \) were estimated empirically by counting the occurrences of \( s', s, \) and \( a \) in the data displayed in Figures 5C/D and Figures 5E/F. The MatrixRank provided by Mathematica 9 was used on the resulting matrices to determine the rank, which lead to the values of \( \text{rank}(\delta_{|S|=159}) = 119 \) and \( \text{rank}(\delta_{|S|=59}) = 59 \).
The embodiment dimensions is upper bounded by the rank of the internal world model, which is why we estimated the number of hidden units for both cases by $m_{|S|=159} = 159 + 119 - 1 \approx 280$ and $m_{|S|=59} = 59 + 59 - 1 \approx 120$.

C.3 Embodiment Dimension and Internal World Model

In many situations, the embodiment dimension is not available from a perspective that is intrinsic to the agent, as it has no direct access to the sensor kernel $\beta$ and the world kernel $\gamma$. From that perspective, only an internal version of the world model is accessible, which we refer to as internal world model. It is defined as a kernel $\delta \in \Delta S \times A$, assigning to each sensor state $s$ with positive probability and each actuator state $a$ the next sensor state $s'$, that is

$$
\delta(s'|s, a) = \sum_{w, w'} p(w|s) \gamma(w'|w, a) \beta(s'|w) = \sum_{w, w'} \frac{p(w) \beta(s|w)}{\sum_{w''} p(w'') \beta(s|w'')} \gamma(w'|w, a) \beta(s'|w') \cdot (18)
$$

Note that the internal world model is not completely determined by $\beta$ and $\gamma$. It also depends on the distribution $p(w)$ of the world states $w$. The extent to which $\delta$ is not a good replacement for $\gamma$ depends on how much the agent can “see” from the world with its sensors. If the agent has direct access to the world state, that is $W_t = S_t$, then both kernels coincide. However, this is not very realistic. Generically, only partial observation of the world is possible. Now the question arises whether it is possible to determine the embodiment dimension $d(\beta, \gamma)$ in terms of $\delta$ even in cases where the agent has only partial access to the world state. This is indeed possible under specific conditions which are met in our experimental setup. We first present the theoretical conditions which have to be met, before they are related to our experiment at the end of this section. In the main text, we have used the rank of $\delta$ as an upper bound of the embodiment dimension $d(\beta, \gamma)$. In what follows, we justify this estimate. We begin with the assumptions for this justification. Consider the world state $w$ as being split into two parts $s, r$, where $s$ is directly accessible to the agent and $r$ denotes the remaining part of the world, which is hidden to the agent. The situation is illustrated in Fig. 6.

$$
\gamma(s', r'|s, r, a) = \gamma^S(s'|s, a) \cdot \gamma^R(r'|r, s, s'). \quad (19)
$$

In this interpretation of the world state, the sensor kernel $\beta$ is simply the identity map $s \mapsto s$. Furthermore, this interpretation also sets structural constraints on the world transition kernel $\gamma$, which assigns a probability for the next world state $w' = (s', r')$ given the current world state $w = (s, r)$ and an actuator value $a$. As $r$ is assumed to be hidden to the agent, $s'$ should not depend on $r$. This leads to the following natural factorization of $\gamma$:

$$
\gamma(s', r'|s, r, a) = \gamma^S(s'|s, a) \cdot \gamma^R(r'|r, s, s'). \quad (19)
$$

With this assumption, we obtain as internal world model

$$
\delta(s'|s, a) = \gamma^S(s'|s, a), \quad \text{whenever } p(s) > 0, \quad (20)
$$

Figure 6: Special causal structure of the sensorimotor loop. The dashed arrows are the ones that we omit within our assumptions.
and the following transition probabilities from a world state \( w = (s, r) \) with positive probability to the world state \( w' = (s', r') \):

\[
p^\pi(s', r'|s, r) = \sum_{s'', a} \beta(s''|s, r) \pi(a|s'') \gamma(s', r'|s, r, a)
= \sum_{a} \pi(a|s) \gamma^S(s'|s, a) \gamma^R(r'|r, s, s')
= \gamma^R(r'|r, s, s') \sum_{a} \pi(a|s) \gamma^S(s'|s, a).
\]

This shows that \( p^\pi = p^\pi^* \) if and only if \( q^\pi = q^\pi^* \), and therefore the embodiment dimension is given by the dimension of the image of \( \pi \mapsto q^\pi \). This is upper bounded by the rank of the kernel \( \gamma^S \) which coincides with the rank of \( \delta \).

This applies to our hexapod experiment discussed in Sec. 5 of the main document for the following reason. In the special case of the tripod gait of a hexapod on an even and otherwise featureless plane, the next joint angles \( S^{t+1} \) are only determined by the current joint angles \( S^t \) and the current action \( A^t \). The rest of the world, here denoted by \( R \) contains information such as the contact points of the legs with the ground. This information is carried from one time step to the next, as it determines how the hexapod walks along the plane. Nevertheless, the contact points of the legs do not influence the joint angles. Hence, in our experiment, \( S' \) is conditionally independent of \( R \) given \( S \) and \( A \). Furthermore, \( R' \) is conditionally independent of \( A \) given \( R \), \( S \) and \( S' \) as the contact points of the legs with the ground are only determined by the relative joint angles, and not by the current action. Therefore, we can estimate the embodiment dimension by the rank of the internal world model.