Max-Planck-Institut für Mathematik in den Naturwissenschaften Leipzig

Information driven self-organization of complex robotic behaviors

by

Georg Martius, Ralf Der, and Nihat Ay

Preprint no.: 15 2013

Information driven self-organization of complex robotic behaviors

Georg Martius[1], Ralf Der[1], Nihat Ay[1,2]
[1]Max Planck Institute for Mathematics, Leipzig, Germany
[2]Santa Fe Institute, Santa Fe, USA
{martius|ralfder|nay}@mis.mpg.de

January 31, 2013

Abstract

Information theory is a powerful tool to express principles to drive autonomous systems because it is domain invariant and allows for an intuitive interpretation. This paper studies the use of the predictive information (PI), also called excess entropy or effective measure complexity, of the sensorimotor process as a driving force to generate behavior. We study nonlinear and nonstationary systems and introduce the time-local predicting information (TiPI) which allows us to derive exact results together with explicit update rules for the parameters of the controller in the dynamical systems framework. In this way the information principle, formulated at the level of behavior, is translated to the dynamics of the synapses. We underpin our results with a number of case studies with high-dimensional robotic systems. We show the spontaneous cooperativity in a complex physical system with decentralized control. Moreover, a jointly controlled humanoid robot develops a high behavioral variety depending on its physics and the environment it is dynamically embedded into. The behavior can be decomposed into a succession of low-dimensional modes that increasingly explore the behavior space. This is a promising way to avoid the curse of dimensionality which hinders learning systems to scale well.

1 Introduction

Autonomy is a puzzling phenomenon in nature and a major challenge in the world of artifacts. A key feature of autonomy in both natural and artificial systems is seen in the ability for independent exploration [9]. In animals and humans, the ability to modify its own pattern of activity is not only an indispensable trait for adaptation and survival in new situations, it also provides a learning system with novel information for improving its cognitive capabilities, and it is essential for development. Efficient exploration in high-dimensional spaces is a major challenge in building learning systems. The famous exploration-exploitation trade-off was extensively studied in the area of reinforcement learning [60]. In a Bayesian formulation this trade-off can be optimally solved [22], however it is computationally intractable. A more conceptual solution is to provide the agent with an intrinsic motivation [51, 55] for focusing on certain things and thus constraining the exploration to a smaller space. To approach this problem in a more fundamental way we consider mechanisms for goal-free exploration of the dynamical properties of a physical system, e. g. a robot. The obtained sensorimotor coordination may later be used to quickly pursue goals. If the exploration is rooted in the agent in a self-determined way, i. e. as a deterministic function of internal state variables and not via a pseudo-random generator it has the chance to escape the curse of dimensionality. Why? Because specific features of the system such as constrains and other embodiment effects can be exploited to reduce the search space.

The solution for such a general problem needs a core paradigm in order to be relevant for a large class of systems. In recent years, information theory has come into the focus of researchers interested in a number of related issues ranging from quantifying and better understanding autonomous systems [7, 33, 23, 56, 24, 64, 53] to questions of spontaneity in biology and technical systems [10] to the self-organization of robot behavior [2, 65].

A systematic approach requires both a convenient definition of the information measure and a robust, real time algorithm for the maximization of that measure. This paper studies in detail the use of the predictive information (PI) of a robot's sensorimotor process. The predictive information of a process quantifies the total information of past experience that can be used for predicting future events. Technically, it is defined as the mutual information between the past and the future of the time series. It has been argued [8] that predictive information, also termed excess entropy [13] and effective measure complexity [26], is the most natural complexity measure for time series. By definition, predictive information of the sensor process is high if the robot manages to produce a stream of sensor values with high information content (in the Shannon sense) by using actions that lead to predictable consequences. A robot maximizing PI therefore is expected to show a high variety of behavior without becoming chaotic or purely random. In this working regime, somewhere between order and chaos, the robot will explore its behavioral spectrum in a self-determined way in the sense discussed above.

This paper studies the control of robots by simple neural networks whose parameters (synaptic strengths and threshold values) are adapted on-line to maximize (a modified) PI of the sensor process. These rules define a mechanism for behavioral variability as a deterministic function formulated at the synaptic level. For linear systems a number of features of the PI maximization method have been demonstrated [2]. In particular, it could be shown that the principle makes the system to explore its behavior space in a systematic manner. In a specific case, the PI maximization caused the controller of a stochastic oscillator system to sweep through the space of available frequencies. More importantly, if the world is hosting a latent oscillation, the controller will learn by PI maximization to go into resonance with this inherent mode of the world. This is encouraging, since maximizing the PI means (at least in this simple example) to recognize and amplify the latent modes of the robotic system.

The present paper is devoted to the extension of the above mentioned method to nonlinear systems with nonstationary dynamics. This leads to a number of novel elements in the present approach. Commonly information theoretic measures are optimized in the stationary state. This is not adequate for a robot in a self-determined process of behavioral development. This paper develops a more appropriate measure for this purpose called the time-local predictive information (TiPI) for general nonstationary processes by using a specific windowing technique and conditioning. Moreover, the application of information theoretic measures in robotics is often restricted to the case of a finite state-action space with discrete actions and sensor values. Also these restrictions are overcome in this paper so that it can be used immediately in physical robots with high dimensional state-action space. This will be demonstrated by examples with two robots in a physically realistic simulation. The approach is seen to work from scratch, i. e. without any knowledge about the robot, so that everything has to be inferred from the sensor values alone. In contrast to the linear case the nonlinearities and the nonstationarity introduce a number of new phenomena, for instance the self-switching dynamics in a simple hysteresis system and the spontaneous cooperation of physically coupled systems. In high-dimensional systems we observe behavioral patterns of reduced dimensionality that are dependent on the body and the environment of the robot.

1.1 Relation to other work

Finding general mechanisms that help robots and other systems to more autonomy, is the topic of intensive recent research. The approaches are widely scattered and follow many different routes so that we give in the following just a few examples.

1.1.1 Information theoretic measures

Information theory has been used recently in a number of approaches in robotics in order (i) to understand how input information is structured by the behavior [34, 33] and (ii) to quantify the nature of information flows inside the brain [23, 56, 24] and in behaving robots [53, 64]. An interesting information measure is the empowerment, quantifying the amount of Shannon information that an agent can "inject into" its sensor through the environment, affecting future actions and future perceptions. Recently, empowerment has been demonstrated to be a viable objective for the self-determined development of behavior in the pole balancer problem and other agents in continuous domains [28].

Driving exploration by maximizing PI can also be considered as an alternative to the principle of homeokinesis as introduced in Der and Liebscher [15], Der [14] that has been applied successfully to a large number of complex robotic systems, see Der et al. [19, 18, 20], Der and Martius [16] and the recent book Der and Martius [17]. Moreover, this principle has also been extended to form a basis for a guided self-organization of behavior [39, 17].

1.1.2 Intrinsic motivation

As mentioned above, the self-determined and self-directed exploration for embodied autonomous agents is closely related to many recent efforts to equip the robot with a motivation system producing internal reward signals for reinforcement learning a pre-specified task. Pioneering work has been done by Schmidhuber using the prediction error as a reward signal in order to make the robot curious for new experiences [50, 58, 52]. Related ideas have been put forward in the so called play ground experiment [29, 43] by using the learning progress as a reward signal. There have been also a few proposals to autonomously form a hierarchy of competencies using the prediction error of skill models [3] or more abstractly to balance skills and challenges [57]. Predictive information can also be used as an intrinsic motivation in reinforcement learning [66] or additional fitness in evolutionary robotics [46].

1.1.3 Embodiment

The past two decades in robotics have seen the emergence of a new trend of control in robotics which is rooted more deeply in the dynamical systems approach to robotics using continuous sensor and action variables. This approach yields more natural movements of the robots and allows to exploit embodiment effects in an effective way, see Pfeifer and Bongard [44], Pfeifer et al. [45] for an excellent survey. The approach described in the present paper is tightly coupled to the ideas of exploiting the embodiment, since the development of behavioral modes is entire dependent on the dynamical coupling of the body, brain, and its environment.

1.1.4 Spontaneity

We would like to briefly discuss the implications of using a self-determined[1] and deterministic mechanism of exploration to the understanding of variability in animal behavior. In the animal kingdom, there is increasing evidence showing that animals from invertebrates to fish, birds, and mammals are equipped with a surprising degree of variety in response to external stimulation [4, 25, 59, 6]. So far, it is not clear how this behavioral variability is created. Ideas cover the whole range from the quantum effects [30] (pure and inexorable randomness) to thermal fluctuations at the molecular level to the assumption of pure spontaneity [42], rooting the variability in the existence of intrinsic, purely deterministic processes.

This paper shows that a pure spontaneity is enough to produce behavioral variations, and as in animals, their exact source appears “indecipherable” from an observer point of view. If the variation of behavior in animals is produced in a similar way, this would bring new insights into the free will conundrum [10].

2 Methods

We start with the general expressions for the PI and introduce a derived quantity called time-local predictive information (TiPI) more suitable for the intended treatment of nonstationary systems. Based on the specific choice of the time windows we derive estimates of the TiPI for general stochastic dynamical systems and give explicit expressions for the special case of a Gaussian noise. The explicit expressions are used for the derivation of the exploration dynamics obtained by gradient ascending the TiPI. Besides giving the exploration dynamics as a batch rule we also derive, in the sense of a stochastic gradient rule, the one-shot gradient. The resulting combined dynamics (system plus exploration dynamics) is a deterministic dynamical system, where the self-exploration of the system becomes a part of the strategy. These general

[1] Self-determined is understood here has “only based its own internal laws”

results are then applied to the case of the sensorimotor loop and we discuss their Hebbian nature and speculate about the cognitive bootstrapping phenomenon as a potential outcome of our approach.

2.1 Predictive information

The predictive information (PI) of a time discrete process $\{S_t\}_{t=a}^{b}$ with values in $\mathbb{R}^n$ is defined [8] as the mutual information between the past and the future, relative to some instant of time $a \leq t_0 < b$

$$I\left(S_{\text{future}}\,;S_{\text{past}}\right) = \left\langle \ln \frac{p\left(s_{\text{future}}\,,s_{\text{past}}\right)}{p\left(s_{\text{past}}\right)p\left(s_{\text{future}}\right)} \right\rangle = H\left(S_{\text{future}}\right) - H\left(S_{\text{future}}\,|S_{\text{past}}\right) \tag{1}$$

where the averaging is over the joint probability density distribution $p\left(s_{\text{past}}, s_{\text{future}}\right)$ and past $:= \{a,\ldots,t_0\}$ and future $:= \{t_0+1,\ldots,b\}$. In more detail, we use the (differential) entropy $H(S)$ of a random variable S given by

$$H(S) = -\int p(s)\ln p(s)\,\mathrm{d}s$$

where $p(s)$ is the probability density distribution of the random variable S. The conditional entropy $H(S'|S)$ is defined accordingly

$$H\left(S'|S\right) = -\int\int p\left(s'|s\right)\ln p\left(s'|s\right)\,\mathrm{d}s'\,p(s)\,\mathrm{d}s$$

$p(s'|s)$ being the conditional probability density distribution of s' given s. As is well known, in the case of continuous variables, the individual entropy components $H\left(S_{\text{future}}\right)$, $H\left(S_{\text{future}}\,|S_{\text{past}}\right)$ may well be negative whereas the PI is always positive and will exist even in cases where the individual entropies diverge. This is a very favorable property deriving from the explicit scale invariance of the PI [2].

The usefulness of the PI for the development of explorative behaviors of autonomous robots has been discussed earlier[2], see Ay et al. [1], Der et al. [21], Ay et al. [2]. This paper continues these investigations for the case of more general situations than those discussed before. In order to do so, we have to introduce some specifications necessary for the development of a general, versatile and stable algorithm realizing the increase of PI in the sensor process at least approximately.

Let us start with simplifying eq. (1). If $\{S_t\}_{t=a}^{b}$ is a Markov process, see Ay et al. [1], the PI is given by the mutual information (MI) between two successive time steps, i. e. instead of eq. (1) we have

$$I(S_t;S_{t-1}) = \left\langle \ln \frac{p\left(s_t,s_{t-1}\right)}{p\left(s_t\right)p\left(s_{t-1}\right)} \right\rangle = H\left(S_t\right) - H\left(S_t|S_{t-1}\right) \tag{2}$$

the averaging being done over the joint probability density $p\left(s_t,s_{t-1}\right)$. Actually, any realistic sensor process will be purely Markovian only in exceptional cases. However, we can use the simplified expression (2)—let us call it the one-step PI—also for general sensor processes taking it as the **definition** of the objective function driving the autonomous exploration dynamics to be derived.

2.1.1 Nonstationarity and time-local predictive information

Most applications done so far were striving for the evaluation of the PI in a stationary state of the system. With our robotic applications, this is neither necessary nor adequate. The robot is to develop a variety of behavioral modes ideally in a open-ended fashion, which will certainly not lead to a stationary distribution of sensor values. The PI would change on the timescale of the behavior. How can one obtain in this case

[2] In experiments with a coupled chain of robots [21] it was observed that the PI of just a single sensor, one of the wheel counters of an individual robot, already yields essential information on the behavior of the robot chain. The PI turned out to be maximal if the individual robots managed to cooperate so that the chain as a whole could navigate effectively. This is remarkable in that a one-dimensional sensor process can already give essential information on the behavior of a very complex physical object under real world conditions. These results give us some encouragement to study the role of PI and other information measures for specific sensor processes as is done in the present paper.

the probability distributions of $p(s_t)$? The solution we suggest is to introduce a conditioning on an initial state in a moving time window and thus obtain the distributions from our local model as introduced below. More formally, let us consider the following setting. Let t be the current instant of time and τ be the length of a time window τ into the past. We study the process in that window with a fixed starting state $s_{t-\tau}$. What we consider is a fraction of the process starting at time $t-\tau$ and ending at time t so that all distributions in eq. (2) are conditioned on state $s_{t-\tau}$. For instance, instead of $p(s_t)$ in eq. (2), we have to use[3]

$$p(s_t|s_{t-\tau}) = \int \cdots \int p(s_t, \ldots, s_{t-\tau+1}|s_{t-\tau})\, \mathrm{d}s_{t-1} \cdots \mathrm{d}s_{t-\tau+1} \tag{3}$$

and the related expression for $p(s_t, s_{t-1}|s_{t-\tau})$, where $p(s_t, \ldots, s_{t-\tau+1}|s_{t-\tau})$ is the joint probability distribution for the process in the time window, conditioned on $s_{t-\tau}$. As to notation, the conditional probabilities depend explicitly on time so that $p(s_{t'}|s_{t'-1})$ is different from $p(s_{t''}|s_{t''-1})$ in general if $t' \neq t''$, with equality only in the stationary state. As a result we obtain the new quantity, written in a short-hand notation as

$$I^{\tau}(S_t; S_{t-1}) := I(S_t; S_{t-1}|S_{t-\tau} = s_{t-\tau}) \tag{4}$$

which we call *time-local predictive information* (TiPI). Note the difference to the conditional mutual information where an averaging over $s_{t-\tau}$ would take place. Analogously we define the time local entropy as

$$H^{\tau}(S_t) := H(S_t|S_{t-\tau} = s_{t-\tau}) \tag{5}$$

2.2 Estimating the TiPI

To evaluate the TiPI only the kernels have to be known which can be sampled by the agent on the basis of the measured sensor values. However, in order to get explicit update rules driving the increase of the TiPI, these kernels have to be known as a function of the parameters of the system, in particular those of the controller. This can be done by learning the kernels as a function of the parameters. A related approach, followed in this paper, is to learn a model of the times series, i.e. learning a function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ acting as a time series predictor $S_t = \psi(S_{t-1}) + \Xi$ with realization

$$s_t = \psi(s_{t-1}) + \xi_t \tag{6}$$

for any time t, ξ_t being the prediction error, also called the noise in the following. ψ can be realized for instance by a neural network that can be trained with any of the standard supervised learning techniques. A concrete example will be considered below, see eq. (37). The relation to the kernel notation is obtained by observing that

$$p(s'|s) = \int \delta(s' - \psi(s) - \xi)\, p_{\Xi}(\xi)\, \mathrm{d}\xi = p_{\Xi}(s' - \psi(s)) \tag{7}$$

where $\delta(u)$ is the Dirac delta distribution and $p_{\Xi}(\xi)$ is the probability density of the random variable Ξ (prediction error) which may depend on the state s itself (multiplicative noise).

The case of linear systems, where $\psi(s) = Ls + b$ with a constant matrix L, has been treated earlier [2] revealing many interesting properties of the PI. How can we translate the findings of the linear systems to the case of nonlinear systems? As it turns out, the nonlinearities introduce many difficulties into the evaluation of the PI as it becomes clear already in a one-dimensional bistable system as treated in Ay et al. [1]. Higher dimensional systems provide even more of such difficulties so that we propose to consider the TiPI on a new basis. The idea is to study the TiPI of the error propagation dynamics in the stochastic dynamical system instead of the process S_t itself.

[3] In the Markovian case this boils down to

$$p(s_t|s_{t-\tau}) = \int \cdots \int p(s_t|s_{t-1}) \cdots p(s_{t-\tau+1}|s_{t-\tau})\, ds_{t-1} \cdots \mathrm{d}s_{t-\tau+1}$$

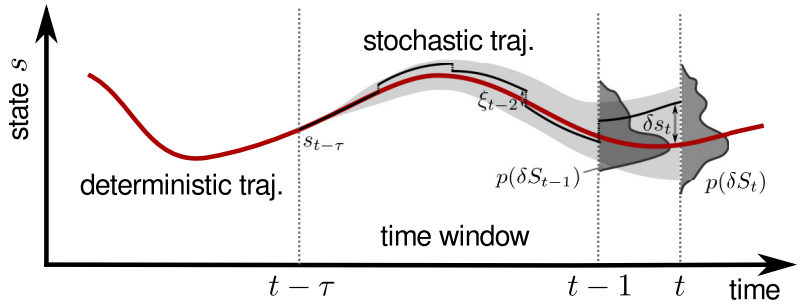

Figure 1: The time window and the error propagation dynamics used for calculating the TiPI, eq. (12). In principle, the process is considered many times with always the same starting value but different realizations of the noise ξ. Note that, when using the one-shot gradients, only one realization is needed, see section 2.3.

2.2.1 Error forward propagation dynamics

Let us assume for the moment that the noise ξ is infinitesimally small so that the following derivations are exact. After formulating the exact results we will discuss how to cope with finite noise situations. As a first step, let us introduce the usual notion of an orbit of the dynamical system defined by the map $\psi : \mathbb{R}^n \to \mathbb{R}^n$. In the time window, see Fig. 1, the orbit is defined by the sequence of states $\hat{s}_{t'} \in \mathbb{R}^n$

$$\hat{s}_{t'} = \psi^{t'-(t-\tau)}(s_{t-\tau}) \tag{8}$$

for any time t' with $t-\tau \leq t' \leq t$, starting state $\hat{s}_{t-\tau} = s_{t-\tau}$, and $\psi^{(0)}(s) = s$. We can consider $\hat{s}_{t'}$ as the predicted state over $t'-(t-\tau)$ time steps. In particular, the prediction over τ steps is $\hat{s}_t = \psi^{\tau}(s_{t-\tau})$.

Instead of the state s itself, let us now consider the dynamics of the differences

$$\delta s_{t'} = s_{t'} - \hat{s}_{t'} \tag{9}$$

between the true state $s_{t'}$, eq. (6), and the state $\hat{s}_{t'}$ obtained by the deterministic dynamics (ψ), see Fig. 1. Assuming ξ is small, $\delta s_{t'}$ obeys the rule[4]

$$\delta s_{t'} = L(s_{t'-1})\,\delta s_{t'-1} + \xi_{t'} + O(\|\xi\|^2) \tag{10}$$

with starting state $\delta s_{t-\tau} = 0$. In the following we will use this approximation which is arbitrary good for infinitesimally small noise. This dynamics corresponds to that of a linear system[5], however with state dependent dynamical operator L.

In the case of finite noise, we can obtain a related rule by using the mean value theorem of differential calculus stating that under mild restrictions one can find a state $\tilde{s}_{t'} \in [\hat{s}_{t'}, s_{t'}]$ so that

$$\delta s_{t'} = L(\tilde{s}_{t'-1})\,\delta s_{t'-1} + \xi_{t'} \tag{11}$$

yields the exact dynamics of the multi-step prediction error δs_t.

The interesting point now is that $I^{\tau}(S_t : S_{t-1})$ (eq. (4)) is equal to that of the process defined by the error propagation dynamics[6], i. e.

$$I^{\tau}(S_t : S_{t-1}) = I^{\tau}(\delta S_t : \delta S_{t-1}) \tag{12}$$

[4]Proof: Using $\hat{s}_{t'} = \psi(\hat{s}_{t'-1})$ we write

$$\begin{aligned}\delta s_{t'} &= s_{t'} - \hat{s}_{t'} = \psi(s_{t'-1}) + \xi_{t'} - \psi(\hat{s}_{t'-1}) \\ &= \psi(\hat{s}_{t'-1} + \delta s_{t'-1}) - \psi(\hat{s}_{t'-1}) + \xi_{t'} \\ &= L(s_{t'-1})\,\delta s_{t'-1} + \xi_{t'} + O(\|\xi\|^2)\end{aligned}$$

[5]In a linear system, L is independent of the state. In this case $\hat{s}_{t'} = L\hat{s}_{t'-1}$ such that the dynamical evolution of δs and s are the same.

[6]Consider two random vectors S and S' together with the shifted vectors $U = S + a$ and $U' = S' + a'$. Using that the probability distribution functions (pdf) $p_S(s)$ and $p_U(u)$ obey $p_U(u) = p_U(s+a) = p_S(s)$ one obtains $H(S) = H(U)$. Analogously, the joint pdf's obey $p_{UU'}(u,u') = p_{UU}(s+a, s'+a') = p_{SS'}(s,s')$ so that $H(S'|S) = H(U'|U)$.

This result is central for the following arguments—we will make use of the fact that the dynamics eq. (10) is more easily treated to obtain explicit estimates for the TiPI and its gradient.

2.2.2 Explicit expressions

By iterating eq. (10) we obtain an explicit expression for δs_t (using here and in the following $L(t')$ for $L(\tilde{s}_{t'})$)

$$\delta s_t = \sum_{k=0}^{\tau-1} L^{(k)}(t-1)\,\xi_{t-k} \tag{13}$$

with

$$L^{(k)}(t-1) = L(t-1)\cdots L(t-k), \text{ and } L^{(0)} = \mathbf{I} \tag{14}$$

for any t. In general, the entropy of δS_t will be very complicated to be obtained in realistic situations with high dimensional physical systems. Therefore we will base the further considerations on a convenient estimate of the latter. With white Gaussian noise, the process δS_t is Gaussian as well, i. e. $\delta S_t \sim \mathcal{N}(0,\Sigma_t)$ (it is a linear combination of independent Gaussians), so that the entropy is given in terms of the covariance matrix Σ_t of the random vector δS_t as [12]

$$H^\tau(\delta S_t) = \frac{1}{2}\ln|\Sigma_t| + \frac{n}{2}\ln 2\pi e \tag{15}$$

$|A|$ denoting the determinant of a square matrix A and

$$\Sigma_t = \left\langle \delta S_t \delta S_t^\top \right\rangle = \int p(\delta s_t)\,\delta s_t \delta s_t^\top \,\mathrm{d}\delta s_t \tag{16}$$

is the covariance matrix of δS_t and $p(\delta s_t)$ is the probability density distribution of the random variable δS_t. Using eq. (13), explicit expressions for Σ can readily be obtained, see eq. (19) below.

By the same arguments, the conditional entropy is defined, using eq. (7), as

$$H^\tau(\delta S_t|\delta S_{t-1}) = H^\tau(\Xi_t) = \frac{1}{2}\ln|D_t| + \frac{n}{2}\ln 2\pi e \tag{17}$$

with

$$D_t = \left\langle \xi_t \xi_t^\top \right\rangle = -\int p(\xi_t)\,\xi_t \xi_t^\top \,\mathrm{d}\xi_t$$

(where $p(\xi)$ is the probability density function of $\Xi \sim \mathcal{N}(0,D_t)$) so that we obtain the estimate of the TiPI as

$$I^\tau(\delta S_t : \delta S_{t-1}) = \frac{1}{2}\ln|\Sigma_t| - \frac{1}{2}\ln|D_t| \tag{18}$$

which is the entropy of the state δs minus that of the noise.

When looking at eq. (18) one sees that the entropies are expressed in terms of covariance matrices. This is exact in the case of Gaussian distributions. In the general case this may be considered as an approximation to the true TiPI. Alternatively, we can also consider eq. (18) as the definition of a new objective function for any process if we agree to measure variability not in terms of entropies but more directly in terms of the covariance matrices.

2.2.3 White noise

Explicit expressions revealing more details of the theory are obtained for the case of white noise, i. e. $\left\langle \xi_t \xi_{t'}^\top \right\rangle = \mathbf{0}$ if $t \neq t'$, so that using eq. (13) in eq. (16) yields

$$\Sigma = \sum_{k=0}^{\tau-1} L^{(k)} D \left(L^{(k)}\right)^\top \tag{19}$$

In particular, in the case of $\tau = 2$, the shortest nontrivial time window, we find

$$\Sigma = D + LDL^\top.$$

It is also useful to introduce the transformed dynamical operator[7] $\hat{L} = \sqrt{D^{-1}} L \sqrt{D}$ so that the covariance matrix is given by $\Sigma = \sum_{k=0}^{\tau-1} \sqrt{D}\hat{L}^{(k)} \left(\hat{L}^{(k)}\right)^\top \sqrt{D}$ and (using $\left|\sqrt{D}M\sqrt{D}\right| = |MD| = |M|\,|D|$)

$$I^\tau(\delta S_t : \delta S_{t-1}) = \frac{1}{2} \ln \left| \sum_{k=0}^{\tau-1} \hat{L}^{(k)} \left(\hat{L}^{(k)}\right)^\top \right| \tag{20}$$

Interestingly, the $\hat{L}$ operators also exist if the overall noise strength $\lambda = \|\xi\|$ goes to zero, so that I^τ stays finite[8] although the defining entropies, conditioned on the state $s_{t-\tau}$, are equal to zero in the deterministic system.

2.2.4 The linear case

For linear systems explicit expressions for the PI were obtained in Ay et al. [2]. In this case L is not dependent on the state s_t of the system so that $L^{(k)} = L^k$ in eq. (14). Using eq. (19), with $\tau \to \infty$, we reobtain the results[9] of Ay et al. [2]. In particular, for the case of normal matrices and isotropic noise, the explicit expression $\Sigma = \left(\mathbb{I} - LL^\top\right)^{-1}$ was obtained forming the basis for the update rules there.

The nonlinear version, with finite τ, yields more involved update rules (from eq. (20)) given by a sum of terms, see eq. (53). Anyhow, this paper does not rely on large τ since we are interested in treating the general nonstationary case relevant for the self-determined self-exploration of autonomous agents. It should be interesting to further investigate the differences between the two approaches, the leading theme probably being the fact that the most important update information is obtained form the derivative of the Jacobi matrix L and not so much from the states δu and δs it is averaged with, see eq. (53). For the same reason, it seems also not crucial to have large time windows as will become more clear from the applications given below.

2.3 The exploration dynamics

Our aim is the derivation of an algorithm driving the behavior of the agent toward increasing TiPI. Let us assume that the function $\psi : \mathbb{R}^n \to \mathbb{R}^n$ depends on a set of parameters θ so that we may write the dynamics as

$$s_t = \psi(s_{t-1}, \theta) + \xi_{t+1} \tag{21}$$

For instance, if ψ is a neural network as introduced further below, the parameter set θ comprises just the synaptic weights and threshold values of the neurons.

2.3.1 Gradient ascending the TiPI

Based on the TiPI, a rule for the parameter dynamics is given by the gradient step to be executed at each time t

$$\Delta\theta = \varepsilon \frac{\partial I}{\partial \theta} \tag{22}$$

where ε is the update rate. Using eq. (18), we obtain

$$\Delta\theta = \varepsilon \frac{\partial}{\partial \theta} \ln|\Sigma| - \varepsilon \frac{\partial}{\partial \theta} \ln|D| \tag{23}$$

[7] This corresponds to using a so-called whitening transformation on the state dynamics, replacing in eq. (11) the state vector δs by a new vector $\delta x = \sqrt{D^{-1}}\delta s$ so that the covariance matrix of the noise in the δx dynamics is just the unit matrix.

[8] Introducing $\hat{D} = \lambda^{-2} D$ where $\hat{D}$ stays finite with $\lambda \to 0$, we have $\hat{L} = \sqrt{\hat{D}^{-1}} L \sqrt{\hat{D}} = \sqrt{D} L \sqrt{D}$ since λ cancels out.

[9] Note that all eigenvalues of the Jacobi matrix L must be less than one by absolute value so that the limes will exist. This requirement also guarantees that the conditioning on $s_{t-\tau}$ looses its influence for $\tau \to \infty$.

where $\Sigma = \Sigma_t$ was introduced in eq. (16). Considering any (square) matrix M depending on a single parameter θ_k of the set θ we have [10] (see for example Magnus and Neudecker [35])

$$\frac{\partial}{\partial M} \ln |M| = \frac{1}{M^\top}$$

and

$$\frac{\partial}{\partial \theta_k} \ln |M| = \sum_{ij} M^{-1}_{ji} \frac{\partial M_{ij}}{\partial \theta_k} = \mathrm{Tr}\left(\left(M^{-1}\right)^\top \frac{\partial M}{\partial \theta_k} \right)$$

so that, for instance, using $\Sigma = \Sigma^\top = \left\langle \delta s \delta s^\top \right\rangle$

$$\frac{\partial}{\partial \theta_k} \ln |\Sigma| = \mathrm{Tr}\left(\frac{1}{\Sigma} \frac{\partial}{\partial \theta_k} \left\langle \delta s \delta s^\top \right\rangle \right) \tag{24}$$

2.3.2 Characterizing the parameter dynamics

In order to better characterize the parameter dynamics, let us consider for the moment Σ at the r. h. s. of eq. (24) to be some fixed, positive matrix (not depending on the parameters θ_k). Then, we can write

$$\mathrm{Tr}\left(\frac{1}{\Sigma} \frac{\partial}{\partial \theta_k} \left\langle \delta s \delta s^\top \right\rangle \right) = \frac{\partial}{\partial \theta_k} \left\langle \mathrm{Tr}\left(\frac{1}{\Sigma} \delta s \delta s^\top \right) \right\rangle = \frac{\partial}{\partial \theta_k} \left\langle \delta s^\top \frac{1}{\Sigma} \delta s \right\rangle$$

(using the cyclic invariance of the trace in the last step). Using this argument also for the D term in eq. (23), the update rule becomes (self-averaging, see section One-shot gradients for details)

$$\Delta\theta = \varepsilon \frac{\partial}{\partial \theta} \left\| \delta s \right\|^2_\Sigma - \varepsilon \frac{\partial}{\partial \theta} \left\| \xi \right\|^2_D \tag{25}$$

where $\|a\|^2_M = a^\top M^{-1} a$ defines the length of a vector a in the metric given by M (considered fixed in the current gradient step).

As eq. (25) suggests, the effect of following the gradient is to both increase the norm of δs in the Σ metric and decrease the norm of ξ in the D metric. In order to get the latter gradient, we must consider ξ as a function of the parameters of the controller. Assuming that ξ is essentially noise, we stipulate that the latter dependence is weak so that the gradient is small and will be omitted in the following applications. This conjecture can not be considered generally valid but is justifiable in the case of parsimonious control as realized by the low-complexity controller networks in the applications studied below.

In a nutshell, eq. (25) reveals the central effect of TiPI maximization: increasing the norm of δs is achieved by increasing the amplification of small fluctuations in the sensorimotor dynamics which is equivalent to increasing the instability of the system dynamics, see also the more elaborate discussion in Der et al. [21]. Explicit rules for the parameter dynamics are derived in the Appendix section A.

2.3.3 Learning vs. exploration dynamics

Usually, updating the parameters of a system according to a given objective is called learning. In that sense, the gradient ascent on the TiPI defines a learning dynamics. However, we would like to avoid this notion here, since actually nothing is learnt. Instead by the interplay between the system and the parameter dynamics, the combined system never reaches a final behavior corresponding to the goal of a learning process. We therefore prefer the notion *exploration dynamics* for the dynamics in the parameter space that is driven by the TiPI maximization.

[10] We write $\frac{1}{M}$ for M^{-1} here and in the following.

2.3.4 Special case – Two-step window

Let us consider the case of a very short time window. With $\tau = 1$ there is no learning signal since $\Sigma = \left\langle \xi \xi^\top \right\rangle$ meaning that $I = 0$. So, $\tau = 2$ is the most simple nontrivial case. The general parameter dynamics, see eq. (23) and the explicit expression (53), boils down to

$$\Delta\theta = \varepsilon \delta u_t^\top \frac{\partial L(t-1)}{\partial \theta} \delta s_{t-1} \tag{26}$$

where the auxiliary vector δu is given as

$$\delta u_t = \Sigma_t^{-1} \delta s_t \tag{27}$$

and

$$\delta s_{t-1} = s_{t-1} - \psi(s_{t-2}) \tag{28}$$

$$\delta s_t = s_t - \psi(\psi(s_{t-2})) \tag{29}$$

$$\Sigma_t = \left\langle \delta s_t \delta s_t^\top \right\rangle \tag{30}$$

stipulating the noise is different from zero (though possibly infinitesimal), and ignoring the correlations with possible noisy terms in $L(t-1)$, see eq. (11).

By a whitening transformation, we can give eq. (26) the more symmetric form

$$\Delta\theta = \varepsilon \delta \bar{s}_t^\top \frac{\partial \check{L}(t-1)}{\partial \theta} \delta \bar{s}_{t-1} \tag{31}$$

where $\delta\bar{s} = \sqrt{\Sigma^{-1}}\delta s$ is a white random vector, $\check{L} = \sqrt{\Sigma^{-1}} L \sqrt{\Sigma}$ is the whitened Jacobian matrix. The dependence of Σ on θ is to be ignored when taking the derivative. In the applications described below, already the simple learning rule defined by eqs. (26) and (31) will be seen to create most complex behaviors of the considered physical robots.

2.3.5 One-shot gradients

The formulas given above for the gradient (eqs. (25) and (26)) were obtained by tacitly invoking the self-averaging properties of the gradient, i. e. by simply replacing $\left\langle \delta s \delta s^\top \right\rangle$ with $\delta s \delta s^\top$ in eq. (24). This still needs a little discussion. Actually, the self-averaging is exactly valid only in the limit of sufficiently small ε, with ε eventually being driven to zero in a convenient way. However, our scenario is different. What we are aiming at is the derivation of an intrinsic mechanism for the self-determined and self-directed exploration using the TiPI and related objectives. The essential point is that self-exploration is driven by a deterministic function of the states (sensor values) of the system itself.

Equation (26) obtained from the gradient of the TiPI fulfills these aims very well—any change of the system parameters and hence of the behavior is given in terms of the predecessor states in the short time window. With finite (and often quite large) ε eqs. (26)-(30) are just a rough approximation of the original TiPI but, in view of our goal, the one-shot nature of the gradient is favorable as it supports the explorative nature of the exploration dynamics generating interesting synergy effects.

2.3.6 Synergy of system and exploration dynamics

A further central aspect of our approach is the interplay between the system and the parameter dynamics driven by the TiPI maximization process. In specific cases, the latter may show convergence as in conventional approaches based on stationary states. An example is given by the one-parameter system studied in Ay et al. [1] realizing convergence to the so called effective bifurcation point. However, with a richer parametrization and/or more complex systems, instead of convergence, the combined system (state + parameter dynamics) never comes to a steady state due to the intensive interplay between the two dynamical components if ε is kept finite. An example will be given in the Results section.

Typically, the TiPI landscape permanently changes its shape due to the fact that increasing the TiPI means in general a destabilization of the system dynamics. If the latter is in an attractor, increasing the TiPI destabilizes the attractor until it may disappear altogether with a complete restructuring of the TiPI landscape. This is but one of the possible scenarios where the exploration dynamics engages into an intensive and persistent interplay with the system dynamics. This interplay leads to many synergistic effects between system and exploration dynamics and makes the actual flavor of the method.

2.3.7 Self-directed search

The common approach to solve the exploration–exploitation dilemma in learning problems is to use some randomization of actions in order to get the necessary exploration and then decrease the randomness to exploit the skills acquired so far. This is prone to the curse of dimensionality if the systems are gaining some complexity. Randomness can also be introduced by using a deterministic policy with a random component in the parameters, as quite successfully applied to evolution strategies and reinforcement learning [54, 27].

Our approach is also to use deterministic policies (given by the function K) but aims at making exploration part of the policy. So, instead of relegating exploration to the obscure activities of a random number generator, variation of actions should be generated by the responses of the system itself. This replaces randomness with spontaneity and is hoped (and will be demonstrated) to restrict the search space automatically to the physically relevant dimensions defined by the embodiment of the system.

Formally, we call a search self-directed if there exists a function α so that the change in the parameters

$$\Delta\theta_t = \alpha\left(s_t, \ldots, s_{t-\tau}, \theta_t\right) \tag{32}$$

is given as a deterministic function of the states in a certain time window (of length τ) and the parameter set θ itself. In this paper, α is given by the gradient of the predictive information in the one-shot formulation, see section 2.3.5.

In more general terms, we believe that randomization of actions makes the agent heteronomous, its fate being determined by an obscure (to him) procedure (the pseudo-random number generator) alien to the nature of its dynamics. The agent is autonomous in the 'genuine' sense only if it varies its actions exclusively by its own internal laws [49]. In our approach, according to eq. (32), exploration is driven entirely by the dynamics of the system itself so that exploration is coupled in an intimate way to the pattern of behavior the robot is currently in. The danger might be that in this way the exploration is restricted too much. As our experiments show, this is not so for active motion patterns in high dimensional systems. This fact can be attributed to the destabilization effect incurred by the TiPI maximization, see above and [21]. For stabilizing behaviors, however, the exploration may be too restrictive.

2.4 The sensorimotor loop

Let us now specify the above expressions to the case of a sensorimotor loop, in particular a neurally controlled robotic system. The dynamical systems formulation is obtained now by writing our predictor for the next sensor values as a function of both the sensors and the actions so that

$$s_t = \phi\left(s_{t-1}, a_{t-1}\right) + \xi_t \tag{33}$$

where ϕ represents the so-called forward model and ξ_t is the prediction error as before. As the next step, we consider the controller also as a deterministic function $K : \mathbb{R}^n \to \mathbb{R}^m$ generating actions (motor values) $a_t \in \mathbb{R}^m$ as a function of the sensor values $s_t \in \mathbb{R}^n$ so that

$$a = K\left(s\right) \tag{34}$$

In the applications, K will be realized as a (feed-forward) neural network. Using eq. (34) in eq. (33) we obtain the map ψ modeling our sensor process as

$$\psi\left(s_{t-1}\right) = \phi\left(s_{t-1}, K\left(s_{t-1}\right)\right)$$

In Ay et al. [2] a standard linear control system was studied where $K\left(s\right) = Cs$, $\phi\left(s,a\right) = Ts + Va$, and $\psi\left(s\right) = \left(T + VC\right)s$ This paper will consider a nonlinear generalization of that case in specific robotic applications.

2.4.1 Exploration dynamics for neural control systems

In the present setting, we assume that both the controller K and the forward model ϕ of our robot are realized by neural networks, the controller being given by a single-layer neural network as

$$K(s) = g(Cs+h) \tag{35}$$

the set of parameters θ now given by C and h. In the concrete applications to be given below, we specifically use $g_i(z) = \tanh(z_i)$ (to be understood as a vector function so that $g : \mathbb{R}^n \to \mathbb{R}^n$).

Moreover, the forward model ϕ is given by a layer of linear neurons, so that

$$\phi(s,a) = Va + Ts + b \tag{36}$$

The matrices V, T and the vector b represent the parametrization of the forward model that can be adapted on-line by a supervised gradient procedure as

$$\Delta V = \eta_\phi \xi a^\top,\ \Delta T = \eta_\phi \xi s^\top,\ \Delta b = \eta_\phi \xi \tag{37}$$

with $\xi_t = s_t - \psi(s_{t-1})$. In the applications, the learning rate η_ϕ is large such that the low complexity of the model is compensated by a very fast adaptation process.

The map ψ becomes $\psi(s) = Vg(Cs+h) + Ts + b$ with Jacobian matrix

$$L = VG'(z)C + T \tag{38}$$

where $z = Cs+h$ is the postsynaptic potential and $G'_{ij}(z) = \delta_{ij} g'(z_i)$.

In the applications given below, we are using the short-time window, with the general exploration dynamics given by eq. (26). The explicit exploration dynamics for this neural setting (with $g(z) = \tanh(z)$) are given as (derivation in Appendix section C)

$$\frac{1}{\varepsilon}\Delta C_{ij} = \delta\mu_i \delta s_j - \gamma_i a_i s_j \tag{39}$$

$$\frac{1}{\varepsilon}\Delta h_i = -\gamma_i a_i \tag{40}$$

where all variables are time dependent and are at time t, except δs which is at time $t-1$. The vector $\delta\mu \in \mathbb{R}^m$ is defined as

$$\delta\mu = G'V^\top \delta u_t = G'V^\top \Sigma^{-1} \delta s_t \tag{41}$$

(see eq. (27)), and the channel specific learning rates γ_i are

$$\gamma_i = 2(C\delta s_{t-1})_i \delta\mu_i \tag{42}$$

in the tanh case, see Appendix section C for the general case.

Introducing the diagonal matrix γ by its matrix elements as

$$\gamma_{ij} = \delta_{ij}\gamma_i$$

we may also write in a more compact notation (reintroducing the time indices)

$$\frac{1}{\varepsilon}\Delta C_t = \delta\mu_t \delta s_{t-1}^\top - \gamma_t a_t s_t^\top \tag{43}$$

$$\frac{1}{\varepsilon}\Delta h_t = -\gamma_t a_t \tag{44}$$

The update rules for $\tau > 2$ are given by a sum of such terms, with appropriate redefinitions of the vector $\delta\mu$.

2.4.2 The Hebbian nature of the update rules

In order to interpret these rules in more neural terms, we at first note that the last term in eq. (39) is of an anti-Hebbian structure. In fact, it is given by the product of the output value a_i of neuron i times the input s_j into the j-th synapse of that neuron, the γ_i (which are positive, as a rule) being interpreted as a neuron specific learning rate. Moreover, we may also consider the term $\delta\mu_i\delta s_j$ as a kind of Hebbian since it is again given by a product of values that are present at the ports of the synapse j of neuron i. The factor δs_j can be considered as a signal directly feeding into the input side of the synapse C_{ij}. Moreover, $\delta\mu$ given as $\delta\mu = G'V^\top\delta u$ is obtained by using δu as the vector of output errors in the ψ network and propagating this error back to the layer of the motor neurons by means of the standard backpropagation algorithm.

These results make the generalization to more complicated, multi-layer networks straightforward. However already the simple setting produces an overwhelming behavioral variety, see the case studies below.

2.4.3 Cognitive bootstrapping

The explicit form of the update rules eqs. (43) and (44) allows a direct insight into the nature of the TiPI approach.

The first term of eq. (43) acts as a self-amplification and increases the Lyapunov exponents. In the linear case [2] this leads eventually to the divergence of the dynamics such that the PI does not exist any longer. With the nonlinearities, the latter effect is avoided, but the system is driven into the saturation region of the motor neurons. However, the second term, by its anti-Hebbian nature, is seen to counteract this tendency. The net effect of both terms is to drive the motor neurons towards a working regime where the reaction of the motors to the changes in sensor values is maximal. This is understandable, given that maximum entropy in the sensor values requires a high sensorial variety that can be achieved by that strategy. Let us have a more detailed look at the driving term $\delta\mu_t\delta s_{t-1}^\top$ which is responsible for the specific searching behavior. We have

$$\delta\mu_t\delta s_{t-1}^\top = ML\xi_{t-1}\xi_{t-1}^\top + M\xi_t\xi_{t-1}^\top$$

where L is at time $t-1$ and $M = G'V^\top\Sigma^{-1}$ is a matrix linking the s-space to the a-space. When applying this component of ΔC to a state vector s_t, the result is dominated by both $\xi_{t-1}\xi_{t-1}^\top$, the projector onto the prediction error ξ_{t-1}, and the corresponding cross-time projector $\xi_t\xi_{t-1}^\top$. So, in the sense of eq. (32), the change $\Delta\theta$ of the parameters θ is at the very heart defined by the $\xi\xi^\top$ projectors.

The functional role of that term depends on the nature of the prediction error ξ as determined by the ability of the map ψ to represent the systematic part of the sensor dynamics. Representing the process very well means that the map covers all systematic parts of the dynamics so that ξ contains essentially noise. In that case, $\Delta\theta$ is essentially noise so that, driven by $\delta\mu\delta s^\top$, the C matrix is progressively randomized pushing the behavior into regions where the model is bad.

However, in regions of poor prediction, ξ still contains essential parts of the systematic dynamics—the embodiment dominated motions in a physical system[11]. Then, by the projection mechanism, the search will be directed more and more into that physical space—the robot starts discovering a new behavioral pattern. Then, with increasing ability of the model to cover that pattern, the randomization process restarts, introducing new search directions so that the robot may discover in such cycles a large variety of its potential dynamical patterns.

This scenario also proposes a solution to what may be called the cognitive bootstrapping problem emerging automatically in high dimensional systems with a rich and diversified behavior space, see Der and Martius [17] for a more detailed discussion. Given such a complexity, one observes the so called cognitive deprivation effect in the first place: when the system is stabilizing in a specific motion pattern (like walking)—the forward model is developing into an expert for that very small region of the full behavior space, becoming completely dull for other behaviors. However, an autonomous system must have the

[11] The extreme case is given by $\psi = 0$ so that $\xi_t = s_t$ meaning that the search is dictated by the dynamics of the process itself. In the special setting this case corresponds to $V = S = b = 0$ so that the forward model always predicts $s = 0$ independently on the actual physical situation. We may call that the "know nothing" regime of the forward model.

capability of eliciting new behaviors, providing the forward model with fresh information to get it out of the deprivation trap. This constitutes the bootstrapping problem: how can the controller excite behavior patterns that are good "food" for the forward model if the latter does not know how to act outside its cognitive niche?

We argue that the above mechanism does exactly the right thing. In fact, in the state of deprivation, the forward model ϕ and hence the map ψ is, in the niche, essentially correct, so that ξ is essentially noise, and (by the projection mechanism) something new is being tried. In the beginning, this is very random but sufficient to excite some dynamical patterns of the system, guiding the search as described above in the direction of a new physical mode. The ensuing change in the behavior is providing the forward model with fresh signals for relearning, discovering new expert domains. In this way the cognitive deprivation can be overcome, the emerging self-directed search giving the robot new capabilities for discovering its behavioral variety. This scenario works from scratch so that one may start with a "do nothing" and "know nothing" setting, the robot detecting its latent modes in a kind of self-determined search process.

3 Results

We apply our theory to three case studies to illuminate the main features. First a hysteresis systems is considered to exemplify the consequences of nonstationarity and the resulting interplay between the exploration dynamics and the system dynamics in a nutshell. In section 3.2 a physical system of many degrees of freedom is controlled by independent controllers that spontaneously cooperate. Finally in section 3.3 we apply the method to a humanoid robot in various situations.

3.1 Hysteresis systems

Nonstationary processes are the main target of our theory, made accessible by the special windowing and averaging technique presented in this paper for the first time. In order to work out the consequences, let us consider an idealized situation where the above derivations, in particular eqs. (26)-(30), are the exact update rules for increasing the TiPI.

Let us consider a single neuron in an idealized sensorimotor loop, where the sensor values are $s_{t+1} = a_t + \xi_{t+1}$ (the white Gaussian noise ξ is added explicitly). This case corresponds to the dynamical system

$$s_{t+1} = \tanh(Cs_t + h) + \xi_{t+1} \tag{45}$$

where now $s_t \in \mathbb{R}^1$. The system was studied earlier [21] in the special case of $h = 0$ and it was shown that the maximization of the PI self-regulates the system parameter C towards a slightly supercritical value $(2 \ll C > 1)$. There, the system is at the so called effective bifurcation point where it is bistable but still sensitive to the noise.

Let us start with keeping C fixed at some supercritical value and concentrating on the behavior of the bistable system as a function of the threshold value h. The interesting point is that the system shows hysteresis. This can be demonstrated best by rewriting the dynamics in state space as a gradient descent. Let us introduce the postsynaptic potential $z_t = Cs_t + h$ and rewrite eq. (45) in terms of z_t as

$$\Delta z_t = -\frac{\partial}{\partial z_t} U(z_t) + \xi_{t+1} \tag{46}$$

where $\Delta z_t = z_{t+1} - z_t$ and the potential is $U(z) = -C \ln\cosh z - hz + \frac{z^2}{2}$ (using $\frac{\partial}{\partial z} \ln\cosh z = \tanh z$). In that picture, the hysteresis properties of the system are most easily demonstrated by Fig. 2. This phenomenon can be related directly to the destabilization effect of the exploration dynamics. In the potential picture, stability is increasing with the well depth. Hence, the exploration dynamics, aiming at the destabilization of the system, is decreasing the depth of the well more and more until the well disappears altogether, see Fig. 2, and the state switches to the other well where the procedure restarts.

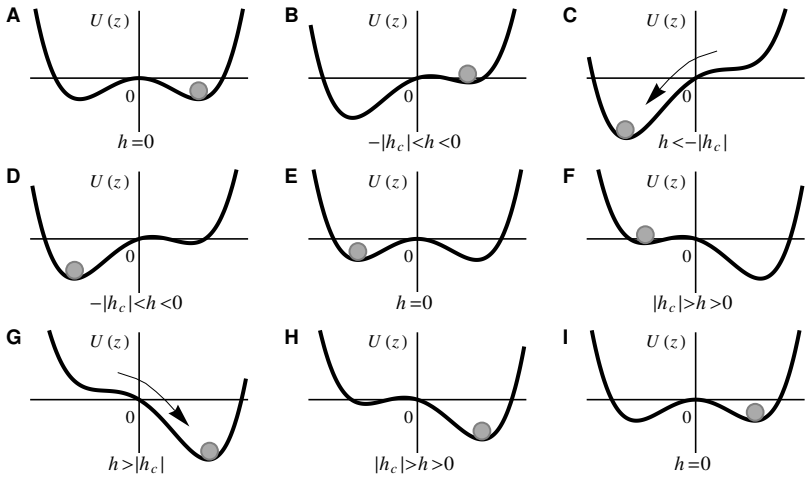

Figure 2: The hysteresis cycle in the gradient picture. The diagrams show the stages of one hysteresis cycle starting from $h = 0$ (**A**) with the state at $z > 0$ as represented by the sphere. Decreasing h creates the asymmetric situation (**B**). If $h = -h_c$ the saddle-node bifurcation happens, i. e. both the maximum at z=0 and the right minimum disappear so that the system shifts to the left minimum of the potential (**C**). Increasing h until $h = 0$ brings us back to the initial situation with the state shifted to the other well see (**D,E**). The diagrams (**F**) and (**G**) depict the switching from the minimum at $z < 0$ to the minimum at $z > 0$ by increasing h. By decreasing h until $h = 0$ the hysteresis cycle is finished, see (**H,I**).

3.1.1 Deterministic self-induced hysteresis oscillation

Now we show that in the one-dimensional case the parameter dynamics is independent of white noise. This implies we can in the state dynamics make the limit of vanishing noise strength and obtain a fully deterministic system. Again we only consider the two-step window ($\tau = 2$). Using $\delta s_t = \xi_t + L\delta s_{t-1} = \xi_t + L\xi_{t-1}$ (eq. (10)) we find that the TiPI, according to eq. (20)

$$I^\tau = \frac{1}{2}\ln\left(1+L^2\right)$$

is independent of the noise. Analogously to eqs. (39)-(42) we obtain the update rules for C and h as the gradient ascent on I^τ and thus the full state-parameter dynamics (with $|\xi| \to 0$) is given by

$$s_t = g\left(Cs_{t-1} + h_{t-1}\right) \tag{47}$$
$$C_t = C_{t-1} + \gamma(1/(2C_{t-1}) - s_{t-1}a_{t-1}) \tag{48}$$
$$h_t = h_{t-1} - \gamma s_t \tag{49}$$

with $\gamma = 2L^2/(1+L^2)$.

Apart from the definition of γ (that just modulates the speed of the parameter dynamics), the extended dynamical system agrees in the one-dimensional case with that derived from the principle of homeokinesis, discussed in detail in Der and Martius [17]. Let us therefore only briefly sketch the most salient features of the dynamics. Keeping C fixed at some supercritical value, as above the most important point is that, instead of converging towards a state of maximum TiPI, the h dynamics drives the neuron through its hysteresis cycle as shown in Fig. 2, which we call a self-induced hysteresis oscillation, see Fig. 3(A).

When we consider the full dynamics (with eq. (48)). Results are given in Fig. 3(B) showing that the feedback strength C in the loop converges indeed toward the regime with the hysteresis oscillation. This demonstrates that the latter is not an artifact present only under the specific parametrization. In fact, we encounter this phenomenon in many applications with complex high-dimensional robotic systems, see the experiments with the ARMBAND below and many examples treated in Der and Martius [17].

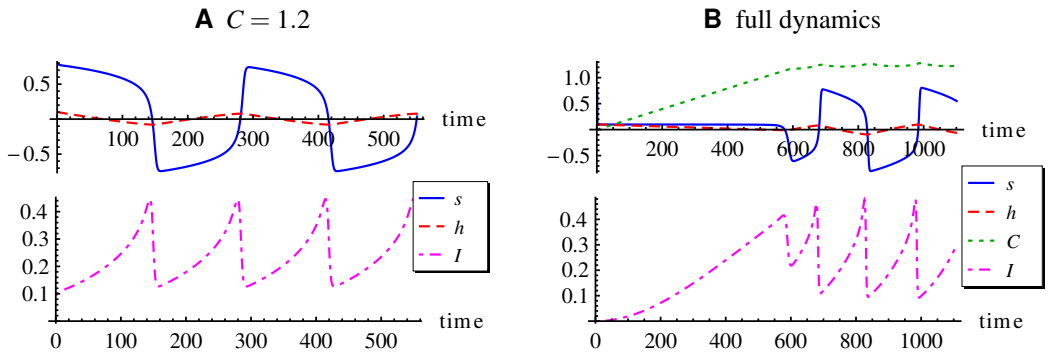

Figure 3: State and parameter dynamics in the one-dimensional system. (**A**) Only h dynamics (fixed $C = 1.2$); the bias h oscillates around zero and causes the state s to jump between the positive and negative fixed points. The TiPI is seen to increase steadily until it eventually drops back when the state is jumping. (**B**) With full dynamics (C,h). C increases until it oscillates around its average at $C \approx 1.2$ where the hysteresis cycle starts. Parameters: $h_0 = 0.1$, $s_0 = 0.8$, $C_0 = 0$ in (**B**), $\varepsilon = 0.002$

Interestingly this behavior is not restricted to simple hysteresis systems but is of more general relevance. For instance, in two-dimensional systems a second order hysteresis was observed, corresponding to a sweep through the frequency space of the self-induced oscillations [17]. It would be interesting to relate this fast synaptic dynamics to the spike-timing-dependent plasticity [38] or other plasticity rules [63] found in the brain.

3.1.2 About time windows

Before giving the applications to embodied systems, let us have a few remarks on the special nature of the time windowing technique as compared to the common settings. Let us consider again the bistable system with the bias h as the only parameter and with finite noise. Figure 4 depicts a typical situation with $h \neq 0$ so that the wells are of different depth. The figure depicts the qualitative difference between the classical attitude of considering information measures in very large time windows, large enough for the process to reach total equilibrium, as compared to our nonstationarity approach where the TiPI is estimated on the basis of a comparatively short window[12].

While in the former case convergence of the hysteresis parameter h towards the equilibrium condition $h = 0$ is reached, there is no convergence in the nonstationary case. Instead, one obtains a self-induced hysteresis oscillation. This is generic for a large class of phenomena based on the synergy effects between system and exploration dynamics, see section 2.3.6, which open new horizons for the explorative capabilities of the agent. In the context of homeokinesis, this phenomenon has already been investigated in many applications, see Der and Martius [17]. This paper provides a new, information theoretic basis and opens new horizons for applications as the matrix inversions inherent to the homeokinesis approach are avoided.

3.2 Spontaneous cooperation with decentralized control

Let us now give examples illustrating the specific properties of the present approach. We start with an example of strongly decentralized control where the TiPI driven parameter dynamics leads to the emergence of collective modes. Earlier papers have already demonstrated this phenomenon for a chain of passively coupled mobile robots [1, 21, 65]. In the setting of Ay et al. [1], Der et al. [21], each wheel was being controlled by a single neuron with a synapse of strength C defining the feedback strength in each of the sensorimotor loops. There was no bias. As it turned out, the TiPI in the sensorimotor loop is maximal if the synaptic strength C is at its critical value where the system is bistable but still reacts to the external

[12] Note that the time to stay in a well is exponentially increasing with the depth of the well and decreasing exponentially with the strength of the noise [48]. Mean first passage times can readily exceed physical times (on the time scale of the behavior) by orders of magnitudes.

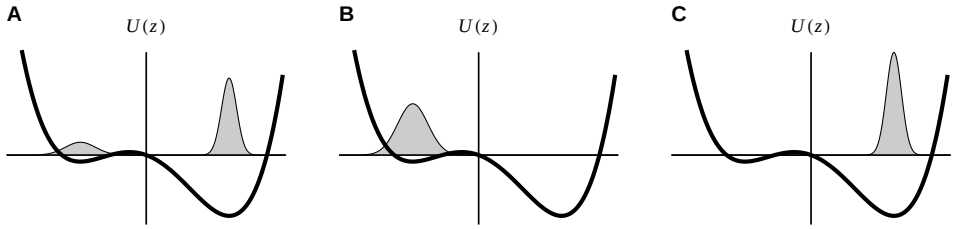

Figure 4: The probability density distributions with different time windows of the stochastic process in an asymmetric double well potential. The mean first passage time T_f of switching between wells is one characteristic time constant of the process [48], T_f increasing exponentially with the barrier height. If observing the process in a window of length $T \gg T_f$, the distribution of **(A)** will be observed. In that situation, the TiPI is maximal if the wells are of equal depth ($h = 0$). However, with windows of length $T \ll T_f$, the system state will be predominantly in one of the wells generating the distributions shown in **(B)**, **(C)**. Gradient ascending the TiPI will decrease the well depth as long as the probability mass is still concentrated in that well. This is what drives the hysteresis cycle depicted in Fig. 2.

perturbations, i. e. the system is at its so-called effective bifurcation point [17]. As compared to the present setting, these results correspond to using a time window of infinite length, stipulating the presence of a stationary state.

The situation is entirely different when using the short time window and large update rates allowing for the synergy effects. In experiments with the robot chain, we observe better cooperativity with the hysteresis oscillations and better exploration capabilities. The reason can be seen in the fact that the self-regulated bias oscillations help the chain to better get out of impasse situations. We do not give details here, since we will study in the following an example that demonstrates the synergy effects even more convincingly.

3.2.1 The ARMBAND

The ARMBAND considered here is a complicated physical object with 18 degrees of freedom, see Fig. 5. Each joint is controlled by an individual controller, a single neuron driven by TiPI maximization, as with the robot chain treated in Der et al. [21]. The controller receives the measured joint angle or slider position and the output of the controller defines the target joint angle or target slider position to be realized by the motor. The motors are implemented as simulated servomotors in order to be as close to reality as possible. Moreover, the forces are limited so that, due to by the interaction with obstacles or the entanglement of the system's different degrees of freedom, the true joint angle may differ substantially from the target angle. These deviations drive the interplay between system and exploration dynamics.

In the experiments, we use the controller given by eq. (35) and the update rules for the parameter dynamics as given by eqs. (48) and (49). The adaptive forward model is given by eq. (36) with $T = 0$ and the appropriate learning rules eq. (37). In order to demonstrate the constitutive role of the synergy effect, we started by studying the system with fixed C and $h = 0$. In contrast to the chain of mobile robots, with fixed parameters there is no parameter regime where the ARMBAND shows substantial locomotion. This result suggests that, as compared to the chain of mobile robots, the specific embodiment of the ARMBAND is more demanding for the emergence of the collective effect.

In order to assess the effects appropriately, note that potential locomotion depends on the forces the motors are able to realize. For instance, if the robot is strongly actuated, the command $a = 0$ for each of the motors drives each joint to its center position so that the shape of the robot is nearly circular, locomotion readily taking place under the influence of very weak external influences. In order to avoid such trivial effects, we use an underactuated setting so that gravitational or environmental forces are deforming the robot substantially, see Fig. 5.

The situation changes drastically if the h dynamics is included. As demonstrated by Fig. 6, substantial locomotion sets in only if ε is large enough so that the exploration dynamics is sufficiently fast for the synergy effect to unfold. Also, as the experiments show, the effect is stable for a very wide range of ε and under varying external conditions. It is also notable, that the ARMBAND robot shows a definite reaction to

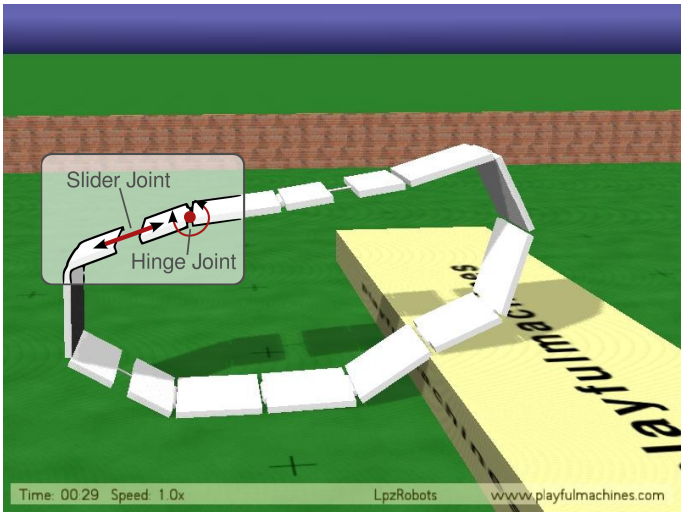

Figure 5: The ARMBAND. The robot has 12 hinge and 6 slider joints, each actuated by a servo motor and equipped with a proprioceptive sensor measuring the joint angle or slider length. The robot is strongly underactuated so that it can not take on a wheel like form where locomotion were trivial.

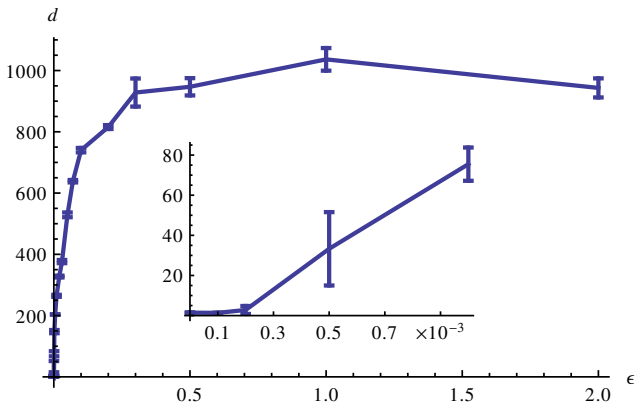

Figure 6: Role of the fast synaptic dynamics: depending on the speed of the synaptic dynamics defined by ε, the locomotion properties are changing drastically. Depicted is the distance traveled by the robot in 10 min simulated time on an empty plane. The inset gives a close up view for low ε, demonstrating that the locomotion starts only if ε exceeds a certain threshold value. Shown is the mean and standard deviation of 10 runs each.

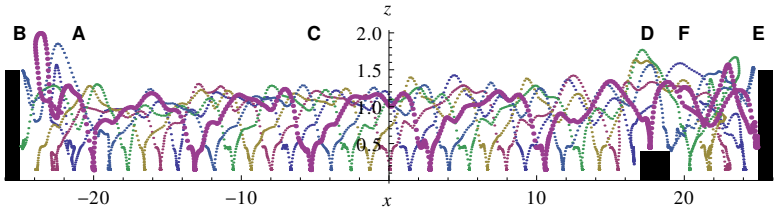

Figure 7: Regular locomotion pattern and interaction with the environment. Plotted are the center positions of the 6 rigid segments in space for an interval of 40 sec. One line is highlighted for visibility. The trajectory starts while the robot is moving to the left (**A**) and is hitting the wall (**B**) (black box) and locomotes to the right (**C**) showing a very regular pattern. Then it overcomes an obstacle (**D**) and hits the wall (**E**) and moves back (**F**). The behavior is cyclic. Parameter: $\varepsilon = 0.5$.

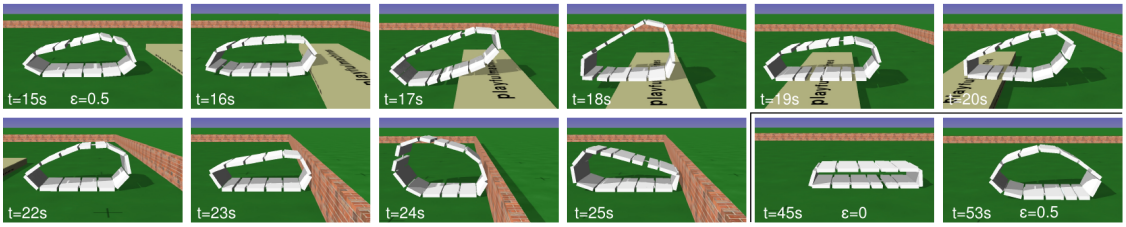

Figure 8: ARMBAND robot surmounting an obstacle and inverting speed at a wall. Screen shots from the simulation for Fig. 7. The order is row-wise from left to right. The last two pictures show the situation after switching off the parameter dynamics $\varepsilon = 0$ for a few seconds (the robots stops) and enabling it again (starts moving).

external influences. For instance, obstacles in its path are either surmounted or cause the robot to invert its velocity, see Fig. 7. The latter effect is observed in particular in the underactuated regime defined above, so that the reflection is not the result of the elastic collision but it is actively controlled by the involvement of the exploration dynamics. The role of the latter is also demonstrated by the fact that locomotion stops as soon as the rate parameter ε is put to zero, see Fig. 8 and the corresponding video S1 on [41].

The ARMBAND has also been investigated recently using artificial evolution for the controller [47], demonstrating convincingly the usefulness of the evolution strategy for obtaining recurrent neural networks that make the ARMBAND roll into a given direction. There are several differences to our approach, both conceptually and in the results. While in the evolution strategy the fitness function was designed for the specific task and many generations were necessary to get the performance, in our approach the rolling modes are emerging right away by themselves. Moreover, the modes are sensitive to the environment, for instance by inverting velocity upon collisions with a wall, they are flexible (changing to a jumping behavior on several occasions) and resilient under widely differing physical conditions. Interestingly, these behaviors are achieved with an extremely simple neural controller, the functionality of a recurrent network being substituted by the fast synaptic dynamics.

3.3 High dimensional case – the HUMANOID

Let us now study the properties of the exploration dynamics in a general (not decentralized) control task. We consider a humanoid robot with 17 degrees of freedom. Each joint is driven by a simulated servo motor, the motor values $a \in \mathbb{R}^{17}$ sent by the controller are the target angles of the joints and sensor values $s \in \mathbb{R}^{17}$ are the true, observed angles. This is the only knowledge the robot has about its physical state. The physics of the robot is simulated realistically in the LPZROBOTS simulator [40].

The aim of this experiment is to investigate in how far the robot develops behaviors with high variability so that it explores its sensorimotor contingencies. Given that there is no externally defined goal for the behavior development, will the robot develop a high behavioral variety depending on its physics and the

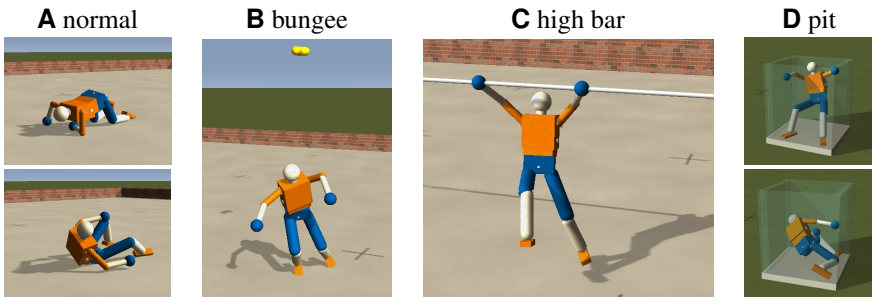

Figure 9: The HUMANOID robot in four different scenarios. (**A**) Normal environment with flat ground. (**B**) The robot is hanging at a bungee like spring. (**C**) The robot is attached to a high bar. (**D**) Robot is fallen into a narrow pit.

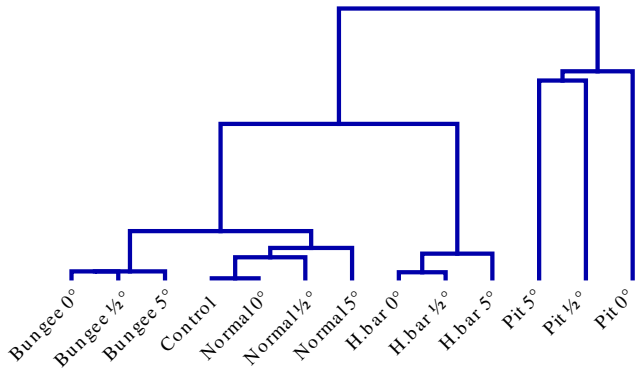

Figure 10: Parameter similarity for the behavior in different environments (Fig. 9). Plotted is the results of a hierarchical clustering based on the difference between the parameters in each of the simulations (averaged over time). For each of the four environments there are three initial poses: 0° (straight upright), 0.5° and 5° slanted to the front. The parameters for runs in the same environment are clustered together. This supports the observation that the embodiment plays an essential role in the generation of behavior. More importantly the physical conditions are reflected in the parameters and are thus internalized. We used the squared norm of the difference of the absolute values of the matrix elements. The absolute values were used because a common structure in the parameters are rotation matrices and there the same qualitative behavior is obtained with inverted signs. Parameters: $\varepsilon = 0.001$ ($\eta = 0.005$)

environment it is dynamically embedded into?

That this happens indeed is demonstrated by the videos S2, S3, S4, S5 on [41]. However, we want a more objective quantity to assess the relation between body and behavior. We provide two different measures for that purpose. One idea is to use the parameter constellation of the controller itself for characterizing the behavior—different behaviors should reflect in characteristic parameter configurations of the controller. In order to study this idea, we place the robot in different scenarios, see Fig. 9, always starting with the same initial parameter configuration (using the result of a preparatory learning phase in the bungee setting), letting the robot move independently for 40 min physical time. Without any additional noise, the dynamics is deterministic so that variations are introduced by starting the robot in different poses, i. e. in a straight upright position and in slightly tilted poses (0.5° and 5° slanted to the front). We then compared the parameter values of the controller matrix C at each second for all simulations and calculated a hierarchical clustering reflecting the differences between the C matrices. Figure 10 shows the resulting dendrogram.

Obviously, there is a distinct grouping of the C matrices according to the environment the robot is in and the behaviors developing in the respective situation. Distances between the groups are different, the most pronounced group corresponding to the behavior in the pit situation. This seems plausible since the constraints are most distinctive here, driving the robot to behaviors that are markedly different from the

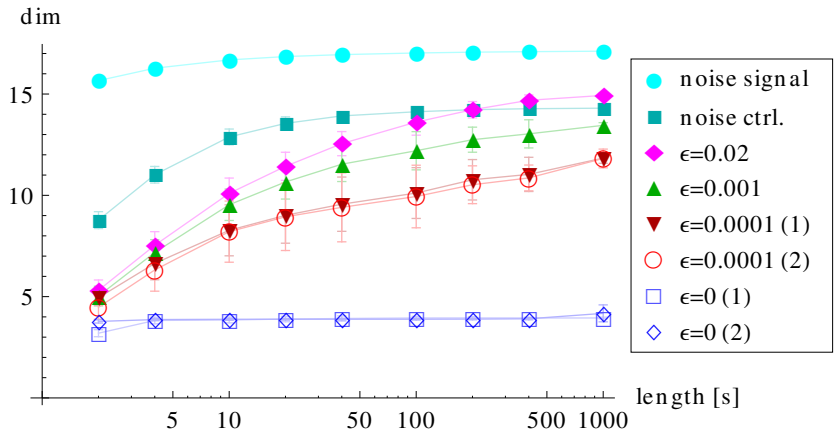

Figure 11: Dimensionality of behavior on different time scales. HUMANOID robot in bungee setup running 40 min with different control settings. The sensor data is partitioned into chunks of a fixed length, the graph depicting the effective dimension over the length of the chunks for different settings. In order to test the method we start with a uniformly distributed noise signal for motor commands ("noise signal"). As expected the observed dimension is maximal. The sensor values produced by that random controller show a lower dimension ("noise ctrl.") as is expected due to the low pass filtering property of the mechanical system. All other cases are with the TiPI maximization controller with different update rates ϵ. In particular, the comparison with the $\epsilon = 0$ case demonstrates that the exploration dynamics produces more complex behaviors than any fixed controller.

situation with the bungee setting, say, where all joints (extremities, hip, back) can move much more freely. There is a second pronounced group – the robot clinging to the high bar – whereas the distances between the C matrices controlling the robot lying on the ground and hanging at the bungee rope is less pronounced. However, by visual inspection the emerging behaviors in the two latter situations appear quite different – a finding that is not so clear in the matrix distance method.

In order to get an additional measure we start from the idea that the TiPI maximization method produces a series of behaviors that are qualified by a high dynamical complexity generated in a controlled way. The latter point means that the dimensionality of the time series of the sensor values is much less than that of the mechanical system – if the behavior of the robot is well controlled (think of a walking pattern) a few master observables will be sufficient to describe the dynamics of the mechanical system. We have tried different methods from dynamical system theory for finding the effective dimension of that time series without much success. The reason was found to be in the strongly nonstationary nature of the compound dynamics (system plus exploration dynamics) making low dimensional behaviors to emerge and disappear in a rapid sequence. So, in the long run the full space of the dynamical system is visited so that globally a seemingly high dimensional behavior is observed.

In order to cope with this nonstationary characteristic, we developed a different method, splitting the whole time series into chunks and using an elementary principal component analysis (PCA) in order to define the effective dimension in each chunk: on each chunk a PCA is performed and the number of principal components required to capture 95% of the data's variance is plotted (mean and standard deviation for all chunks of the same length). In order to avoid discretization artifacts we linearly interpolate the required number of components to obtain a real number.

The results presented in Fig. 11 corroborate the above hypothesis on the dimensionality of the behaviors. In particular, we observe the increase of the effective dimension if the chunk length is increasing, mixing different low dimensional behaviors. The latter point is made even more obvious in Fig. 12 depicting the overlap between the behaviors in chunks at different times. This overlap is large if the behaviors are essentially the same and small if the behavior has changed in the time span between the chunks. As the figure demonstrates, the overlap is indeed large for short time spans, but behaviors can reemerge after

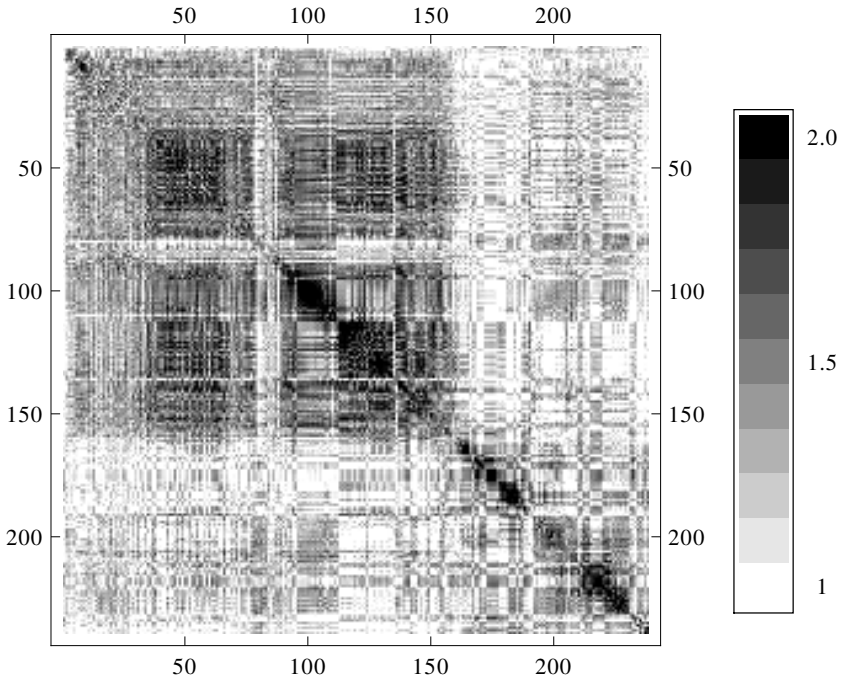

Figure 12: Behavioral changes with time. Pairwise distances of chunks with length 10 s. Distance is defined as the length of the vector of maximal projections of the first 6 principal components.

some time. Altogether, the results demonstrate that our TiPI maximization method effectively explores the behavior space of high-dimensional robotic systems by exciting their low-dimensional modes, avoiding in this way the curse of dimensionality.

4 Discussion

Can a robot develop its skills completely on its own, driven by the sole objective to gain more and more information about its body and its interaction with the world? This question raises immediately further issues such as (i) what is the relevant information for the robot and (ii) how can one find a convenient update rule that realizes the gradient ascent on this information measure. We have studied the predictive information of the stream of sensor values as a tentative answer to the first question and, based on that, could give exact answers to the second question for simple cases. Earlier work was restricted to linear systems [2]. In order to be applicable to actual robotic systems we extend it to the case of nonlinear controllers and to nonstationary processes leading to a new measure called TiPI (time-local predictive information). Using several approximations we have been still able to obtained analytical results. In this way we derived an explicit exploration dynamics for the controller parameters based on an information maximization principle, namely by maximizing the TiPI using gradient ascent. For neural networks the gradient yields a fast synaptic dynamics which is essentially local in nature. Interestingly the TiPI landscape (on which the gradient is calculated) continuously changes its shape due to the general destabilization of the system dynamics inherent in maximizing the TiPI. For instance if the system dynamics is in an attractor, increasing the TiPI destabilizes the attractor until it may disappear altogether with a complete restructuring of the TiPI landscape. This is another reason why nonstationary processes have to be handled and why no convergence of the parameter dynamics is desired.

We studied a one-dimensional hysteresis system in order to work out the consequences of the nonstationary. The parameter dynamics leads to a slightly supercritical regime and additionally a self-induced hysteresis oscillation emerges. This is a useful new property as shown in the experiment with the ARMBAND robot, a high-dimensional robot with a complicated dynamics. Despite the highly decentralized

control—each joint is controlled individually—the robot develops coherent and global pattern of behavior. This is enabled by the continuous adaptation and spontaneous mutual cooperation of the individual controllers (hysteresis elements). We find the effect to be very robust against the speed of the exploration dynamics. Interestingly in the one-dimensional case the update formulas are independent of white noise and we can obtain an exploration dynamics in a fully deterministic system.

The new theoretical basis also allows for controlling complex high-dimensional robotic systems. This is demonstrated by a series of experiments with the HUMANOID robot, now jointly controlled by a single high-dimensional controller. Given that there is no externally defined goal for the behavior development, will the robot develop a high behavioral variety depending on its physics and the environment it is dynamically embedded into? Our results support a positive answer to this question. We quantify the dimensionality and temporal structure of the behavior and find a succession of low-dimensional modes that increasingly explore the behavior space. Furthermore we show that environmental factors influence the internal as well as behavioral development. Without additional noise, the deterministic dynamics leads to an individual development which depends decisively on the particular experiences made during the lifetime.

The exploration dynamics can be viewed as a self-directed search process, where the directions to explore are created from the dynamics of the system itself. Without a random component the changes of the parameters are deterministically given as a function of the sensor values and internal parameters in a certain time window. For an embodied system this means in particular that constraints, responses and current knowledge of the dynamical interaction with the environment can directly be used to advance further exploration. Randomness is replaced with spontaneity which we demonstrate to restrict the search space automatically to the physically relevant dimensions. Its effectiveness is shown in the HUMANOID experiments and we argue that this is a promising way to avoid the curse of dimensionality.

What is the relation of the parameter dynamics described here to other work on maximizing information quantities in neural systems? Maximizing the mutual information between input and output of a neuron, known as InfoMax, yields a very similar parameter dynamics [5]. Interestingly, when applied to a feed-forward network an independent component analysis can be performed. Also similar rules have been obtained in Triesch [61] where the entropy of the output of a neuron was maximized under the condition of a fixed average output firing-rate Triesch [61]. The resulting dynamics is called intrinsic plasticity as it acts on the membrane instead of on the synaptic level and it was shown to result in the emergence of complex dynamical phenomena [11, 31, 62, 32]. In Markovic and Gros [36, 37] a related dynamics is obtained at the synaptic level of a feedback circuit realized by an autaptic (self) connection. In a recurrent network of such neurons it was shown that any finite update rate (ε in our case) destroys all attractors, leading to intermittently bursting behavior and self-organized chaos.

Our work differs in two aspects. On the one hand, we use the information theoretical principle at the behavioral level of the whole system by maximizing the TiPI on the full sensorimotor loop, whereas they use it at the neuronal level. Nevertheless we manage to root the information paradigm back to the level of the synaptic dynamics of the involved neurons. On the other hand, as a direct consequence of that approach, there is no need to specify the average output activity of the neurons. Instead the latter is self-regulating by the closed loop setting. Independent of the specific realization, the general message is that these self-regulating neurons realize a specific working regime where they are both active and sensitive to influences of their environment. If embedded into a feedback setting many interesting phenomena are produced. Instead of studying them in internal (inside the "brain") recurrences, we embed such neurons into a feedback loop with complex physical systems where the self-active and highly responsive nature of these neurons produces similar phenomena at the behavioral level.

In the current form, our approach is limited to the control of robots where the sensorimotor dynamics can be, in its essence, modeled by a simple feed-forward neural network. The parameter dynamics can also be calculated for more complex controllers, such as recurrent networks, which remains for future work. In this study only proprioceptive sensors measuring joint angles have been used. However, our newest experiences have shown that also other sensors e. g. current sensors, acceleration sensor or velocity sensors can be successfully integrated.

To conclude, information theory is a powerful tool to express principles to drive autonomous systems because it is domain invariant and allows for an intuitive interpretation. We present for the first time, to our knowledge, a method linking information theoretic quantities on the behavioral level (sensor values) to explicit dynamical rules on the internal level (synaptic weights) in a systematic way. This opens new

horizons for the applicability of information theory to the sensorimotor loop and autonomous systems.

Acknowledgments

The project was supported by the DFG (SPP 1527).

References

[1] N. Ay, N. Bertschinger, R. Der, F. Güttler, and E. Olbrich. Predictive information and explorative behavior of autonomous robots. *The European Physical Journal B - Condensed Matter and Complex Systems*, 63(3):329–339, 2008. doi: 10.1140/epjb/e2008-00175-0.

[2] N. Ay, H. Bernigau, R. Der, and M. Prokopenko. Information driven self-organization: The dynamical systems approach to autonomous robot behavior. *Theory Biosci.*, 131(3):161–179, 2012. URL `http://www.mdpi.com/1999-4893/2/1/398`.

[3] A. G. Barto. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of 3rd Int. Conference Development Learn.*, pages 112–119, San Diego, CA, USA, 2004.

[4] M. Bekoff and J. A. Byers, editors. *Animal Play: Evolutionary, Comparative and Ecological Perspectives*. Cambridge University Press, 1998.

[5] A. J. Bell and T. J. Sejnowski. An information-maximisation approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[6] D. E. Berlyne. Curiosity and exploration. *Science*, 153(3731):25–33, 1966.

[7] N. Bertschinger, E. Olbrich, N. Ay, and J. Jost. Autonomy: An information theoretic perspective. *Biosystems*, 91(2):331–345, 2008.

[8] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity and learning. *Neural Computation*, 13:2409, 2001.

[9] M. A. Boden. Autonomy: What is it? *Biosystems*, 91(2):305–308, 2008.

[10] B. Brembs. Towards a scientific concept of free will as a biological trait: spontaneous actions and decision-making in invertebrates. *Proc. R. Soc. B*, 278:930–939, 2011.

[11] N. Butko and J. Triesch. Exploring the role of intrinsic plasticity for the learning of sensory representations. In *ESANN*, pages 467–472, 2006.

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2006.

[13] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Lett.*, 63:105–108, 1989.

[14] R. Der. Self-organized acquisition of situated behaviors. *Theory in Biosci.*, 120:179–187, 2001.

[15] R. Der and R. Liebscher. True autonomy from self-organized adaptivity. In *Proc. of EPSRC/BBSRC Intl. Workshop on Biologically Inspired Robotics*, HP Labs Bristol, 2002.

[16] R. Der and G. Martius. From motor babbling to purposive actions: Emerging self-exploration in a dynamical systems approach to early robot development. In S. Nolfi, G. Baldassarre, R. Calabretta, J. C. T. Hallam, D. Marocco, J.-A. Meyer, O. Miglino, and D. Parisi, editors, *From Animals to Animats 9 (SAB 2006)*, volume 4095 of *LNCS*, pages 406–421. Springer, 2006. ISBN 3-540-38608-4.

[17] R. Der and G. Martius. *The Playful Machine - Theoretical Foundation and Practical Realization of Self-Organizing Robots*. Springer, 2012.

[18] R. Der, F. Hesse, and G. Martius. Learning to feel the physics of a body. In *Proc. Intl. Conf. on Computational Intelligence for Modelling, Control and Automation (CIMCA 06)*, pages 252–257, Washington, DC, USA, 2005. IEEE Computer Society. ISBN 0-7695-2504-0-02.

[19] R. Der, F. Hesse, and G. Martius. Rocking stamper and jumping snake from a dynamical system approach to artificial life. *Adaptive Behavior*, 14(2):105–115, 2006. doi: 10.1177/105971230601400202.

[20] R. Der, G. Martius, and F. Hesse. Let it roll – emerging sensorimotor coordination in a spherical robot. In L. M. Rocha, L. S. Yaeger, M. A. Bedau, D. Floreano, R. L. Goldstone, and A. Vespignani, editors, *Proc, Artificial Life X*, pages 192–198. Intl. Society for Artificial Life, MIT Press, August 2006.

[21] R. Der, F. Güttler, and N. Ay. Predictive information and emergent cooperativity in a chain of mobile robots. In S. Bullock, J. Noble, R. Watson, and M. A. Bedau, editors, *Proc. Artificial Life XI*, pages 166–172. MIT Press, Cambridge, MA, 2008.

[22] M. O. Duff. *Optimal learning: computational procedures for bayes-adaptive markov decision processes*. PhD thesis, University of Massachusetts Amherst, 2002. AAI3039353.

[23] K. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2:56–78, 1995.

[24] M. Garofalo, T. Nieus, P. Massobrio, and S. Martinoia. Evaluation of the performance of information theory-based methods and cross-correlation to estimate the functional connectivity in cortical networks. *PLoS ONE*, 4(8):e6482, 08 2009. doi: 10.1371/journal.pone.0006482.

[25] S. Glickman and R. Sroges. Curiosity in zoo animals. *Behaviour*, pages 151–188, 1966.

[26] P. Grassberger. Toward a quantitative theory of self-generated complexity. *Int. J. Theor. Phys.*, 25(9): 907–938, 1986.

[27] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[28] T. Jung, D. Polani, and P. Stone. Empowerment for continuous agent-environment systems. *Adaptive Behavior - Animals, Animats, Software Agents, Robots, Adaptive Systems*, 19(1):16–39, Feb. 2011. ISSN 1059-7123. doi: 10.1177/1059712310392389.

[29] F. Kaplan and P.-Y. Oudeyer. Maximizing learning progress: An internal reward system for development. *Embodied Artificial Intelligence*, pages 629–629, 2004.

[30] C. Koch. Free Will, Physics, Biology, and the Brain. In N. Murphy, G. F. R. Ellis, and T. O'Connor, editors, *Downward Causation and the Neurobiology of Free Will*, pages 31–52. Springer, 2009. doi: 10.1007/978-3-642-03205-9_2.

[31] A. Lazar, G. Pipa, and J. Triesch. The combination of STDP and intrinsic plasticity yields complex dynamics in recurrent spiking networks. In *ESANN*, pages 647–652, 2006.

[32] A. Lazar, G. Pipa, and J. Triesch. Emerging bayesian priors in a self-organizing recurrent network. In *ICANN (2)*, pages 127–134, 2011.

[33] M. Lungarella and O. Sporns. Mapping information flow in sensorimotor networks. *PLoS Comput Biol*, 2(10):e144, 10 2006. doi: 10.1371/journal.pcbi.0020144.

[34] M. Lungarella, T. Pegors, D. Bulwinkle, and O. Sporns. Methods for quantifying the informational structure of sensory and motor data. *Neuroinformatics*, 3(3):243–262, 2005.

[35] J. Magnus and H. Neudecker. *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons, New York, NY, USA, 1988.

[36] D. Markovic and C. Gros. Self-Organized chaos through polyhomeostatic optimization. *Physical Review Letters*, 105(6):068702+, 2010. doi: 10.1103/PhysRevLett.105.068702.

[37] D. Markovic and C. Gros. Intrinsic adaptation in autonomous recurrent neural networks. *Neural Computation*, 24(2):523–540, 2012.

[38] H. Markram, J. Lübke, M. Frotscher, and B. Sakmann. Regulation of synaptic efficacy by coincidence of postsynaptic aps and epsps. *Science*, 275(5297):213–215, 1997. doi: 10.1126/science.275.5297.213.

[39] G. Martius, J. M. Herrmann, and R. Der. Guided self-organisation for autonomous robot development. In F. Almeida e Costa, L. Rocha, E. Costa, I. Harvey, and A. Coutinho, editors, *Proc. Advances in Artificial Life, 9th European Conf. (ECAL 2007)*, volume 4648 of *LNCS*, pages 766–775. Springer, 2007. ISBN 978-3-540-74912-7.

[40] G. Martius, F. Hesse, F. Güttler, and R. Der. LPZROBOTS: A free and powerful robot simulator. `http://robot.informatik.uni-leipzig.de/software`, 2010.

[41] G. Martius, R. Der, and N. Ay. Supplementary material:. `http://playfulmachines.com/TiPI`, 2013.

[42] A. Maye, C.-h. Hsieh, G. Sugihara, and B. Brembs. Order in spontaneous behavior. *PLoS ONE*, 2(5): e443, 05 2007. doi: 10.1371/journal.pone.0000443.

[43] P.-Y. Oudeyer, F. Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *Evolutionary Computation, IEEE Transactions on*, 11(2):265–286, April 2007.

[44] R. Pfeifer and J. C. Bongard. *How the Body Shapes the Way We Think: A New View of Intelligence*. MIT Press, Cambridge, MA, November 2006. ISBN 0262162393.

[45] R. Pfeifer, M. Lungarella, and F. Iida. Self-organization, embodiment, and biologically inspired robotics. *Science*, 318:1088–1093, 2007.

[46] M. Prokopenko, V. Gerasimov, and I. Tanev. Evolving spatiotemporal coordination in a modular robotic system. In S. Nolfi, G. Baldassarre, R. Calabretta, J. Hallam, D. Marocco, J.-A. Meyer, and D. Parisi, editors, *From Animals to Animats 9*, volume 4095 of *LNCS*, pages 558–569. Springer, 2006.

[47] C. W. Rempis. *Evolving complex neuro-controllers with interactively constrained neuro-evolution*. PhD thesis, University of Osnabrück, 2012.

[48] H. Risken. *The Fokker-Planck Equation, 2nd edition*. Springer, 1989.

[49] M. Rohde and J. Stewart. Ascriptional and 'genuine' autonomy. *Biosystems*, 91(2):424–433, 2008.

[50] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats (SAB 1991)*, pages 222–227, Cambridge, MA, USA, 1990. MIT Press.

[51] J. Schmidhuber. Curious model-building control systems. In *In Proc. Intl. Joint Conf. on Neural Networks, Singapore*, pages 1458–1463. IEEE, 1991.

[52] J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. *Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76, 2009.

[53] N. M. Schmidt, M. Hoffmann, K. Nakajima, and R. Pfeifer. Bootstrapping perception using information theory: case study in a quadruped robot running on different grounds. *Advances in Complex Systems, submitted*, 2012.

[54] F. Sehnke, C. Osendorfer, T. Rückstieß, A. Graves, J. Peters, and J. Schmidhuber. Parameter-exploring policy gradients. *Neural Networks*, 23(4):551–559, 2010.

[55] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Trans. on Auton. Ment. Dev.*, 2(2):70–82, June 2010. ISSN 1943-0604. doi: 10.1109/TAMD.2010.2051031.

[56] O. Sporns and G. Tononi. Classes of network connectivity and dynamics. *Complexity*, 7:2002, 2002.

[57] L. Steels. The autotelic principle. *Embodied Artificial Intelligence*, pages 629–629, 2004.

[58] J. Storck, S. Hochreiter, and J. Schmidhuber. Reinforcement driven information acquisition in non-deterministic environments. In *Proceedings of the International Conference on Artificial Neural Networks*, pages 159–164, 1995.

[59] M. Stöwe, T. Bugnyar, M. Loretto, C. Schlögl, F. Range, and K. Kotrschal. Novel object exploration in ravens (Corvus corax): Effects of social relationships. *Behavioural Processes*, 73:68–75, 2006.

[60] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, March 1998.

[61] J. Triesch. A gradient rule for the plasticity of a neuron's intrinsic excitability. In *Proceedings of the 15th international conference on Artificial Neural Networks: biological Inspirations - Volume Part I*, ICANN'05, pages 65–70, Berlin, Heidelberg, 2005. Springer-Verlag. ISBN 3-540-28752-3, 978-3-540-28752-0. doi: 10.1007/11550822_11.

[62] J. Triesch. Synergies between intrinsic and synaptic plasticity mechanisms. *Neural Computation*, 19 (4):885–909, 2007.

[63] G. G. Turrigiano, K. R. Leslie, N. S. Desai, L. C. Rutherford, and S. B. Nelson. Activity-dependent scaling of quantal amplitude in neocortical neurons. *Nature*, 391:892–896, 1998.

[64] P. L. Williams and R. D. Beer. Information dynamics of evolved agents. In S. Doncieux, B. Girard, A. Guillot, J. Hallam, J.-A. Meyer, and J.-B. Mouret, editors, *SAB*, volume 6226 of *Lecture Notes in Computer Science*, pages 38–49. Springer, 2010. ISBN 978-3-642-15192-7.

[65] K. Zahedi, N. Ay, and R. Der. Higher coordination with less control – A result of information maximization in the sensorimotor loop. *Adaptive Behavior*, 18(3-4):338–355, 2010. doi: 10.1177/1059712310375314.

[66] K. Zahedi, G. Martius, and N. Ay. Predictive information in reinforcement learning of embodied agents. In *Int. Workshop on Guided Self-Organization 5*, 2012. Abstract.

A Explicit gradient step

In order to derive the update rule, we start from eq. (24) considering

$$\frac{\partial I}{\partial \theta} = \left\langle \delta s_t^\top \Sigma^{-1} \frac{\partial}{\partial \theta} \delta s_t \right\rangle \tag{50}$$

By eq. (11) we obtain (ignoring the dependence of ξ on the parameter)

$$\frac{\partial}{\partial \theta} \delta s_{t'} = \frac{\partial L(t'-1)}{\partial \theta} \delta s_{t'-1} + L\left(t'-1\right) \frac{\partial}{\partial \theta} \delta s_{t'-1}$$

so that

$$\frac{\partial}{\partial\theta}\delta s_t = \sum_{l=1}^{\tau-1} L^{(l-1)}(t-1)\frac{\partial L(t-l)}{\partial\theta}\delta s_{t-l}$$

where $L^{(k)}(t-1)$ is obtained from eq. (14). Using $a^\top W b = \left(W^\top a\right)^\top b$, we write

$$\frac{\partial I}{\partial\theta} = \sum_{l=1}^{\tau-1}\left\langle \delta u_{t-l+1}^\top \frac{\partial L(t-l)}{\partial\theta}\delta s_{t-l}\right\rangle \tag{51}$$

where (Σ is symmetric)

$$\delta u_{t-l+1} = \left(L^{(l-1)}(t-1)\right)^\top \Sigma_t^{-1}\delta s_t \tag{52}$$

Stipulating the self-averaging property of the gradient, we realize the learning rule as

$$\Delta\theta = \varepsilon\sum_{l=1}^{\tau-1}\delta u_{t-l+1}^\top \frac{\partial L(t-l)}{\partial\theta}\delta s_{t-l} \tag{53}$$

B Learning the inverse covariance matrix

Note that the covariance matrix given in eq. (28) can be easily obtained by the on-line update rule

$$\Delta\Sigma_t = \eta\left(\delta s_t\delta s_t^\top - \Sigma_t\right) \tag{54}$$

or

$$\Sigma_{t+1} = (1-\eta)\Sigma_t + \eta\delta s_t\delta s_t^\top \tag{55}$$

realizing a sampling over a restricted period of time. The update rate η defines the time horizon $t_H \propto \eta^{-1}$ for the averaging. The only remaining nontrivial operation in that setting is the inversion of the covariance matrix Σ. However, this can also be reduced to elementary operations by using the Sherman-Morrison formula as given by

$$\left(A+uv^\top\right)^{-1} = A^{-1} - \frac{1}{1+v^\top A^{-1}u}A^{-1}uv^\top A^{-1}$$

Putting $A = (1-\eta)\Sigma$ and $uv^\top = \eta\delta s\delta s^\top$ we get

$$\left((1-\eta)\Sigma_t + \eta\delta s_t\delta s_t^\top\right)^{-1} = \frac{1}{1-\eta}\Sigma_t^{-1} - \frac{\eta}{(1-\eta)^2\left(1+\frac{\eta}{1-\eta}\delta s_t^\top\Sigma_t^{-1}\delta s_t\right)}\Sigma_t^{-1}\delta s_t\delta s_t^\top\Sigma_t^{-1}$$

and thus

$$\Sigma_{t+1}^{-1} = \frac{1}{1-\eta}\Sigma_t^{-1} - \frac{\beta}{1-\eta}\Sigma_t^{-1}\delta s_t\delta s_t^\top\Sigma_t^{-1}$$

where $\beta\in\mathbb{R}$ is given by

$$\beta = \frac{\eta}{\left(1-\eta+\eta\delta s_t^\top\Sigma_t^{-1}\delta s_t\right)}$$

Note that $\delta s_t^\top\Sigma_t^{-1}\delta s_t$ featuring in the denominator of β is a scalar so that with Σ_t^{-1} given there is no matrix inversion to be done.

If Σ_t is an $n\times n$ matrix, the cost of getting Σ_{t+1} is $O\left(n^2\right)$. This is very favorable if the dimension of the sensor space is large. Using the above formula, the only true inversion (of order $O\left(n^3\right)$) has to be done just once, when starting the process (with a convenient initialization of Σ).

C Neural networks—derivation of the update rule

Let us consider the case that $L = VG'(z)C + T$ with $z = Cs + h$ so that for any two vectors a, b we have

$$\begin{aligned} a^\top \frac{\partial L}{\partial C_{ij}} b &= a^\top V G' \frac{\partial C}{\partial C_{ij}} b + a^\top V \frac{\partial G'}{\partial C_{ij}} Cb \\ &= \left(G'V^\top a\right)_i b_j + a^\top V \frac{\partial G'}{\partial C_{ij}} Cb \end{aligned}$$

The second term is treated by introducing the column $\delta^i \in \mathbb{R}^m$ with

$$\left(\delta^i\right)_l = \delta_{il}$$

and using that quite generally

$$\frac{\partial G'(z)}{\partial C_{ij}} = \sum_r G''(z)\, \delta^r \delta^{r^\top} \frac{\partial z_r}{\partial C_{ij}} \tag{56}$$

(note that $\delta^r \delta^{r^\top}$ is a matrix and the derivative of z_r is a scalar). Considering only the explicit dependence (the full case is discussed with eq. (64) below), we have $\partial z_r / \partial C_{ij} = \delta_{ir} s_j$ or $\partial z / \partial C_{ij} = \delta^i s_j$ so that

$$a^\top V \frac{\partial G'}{\partial C_{ij}} Cb = a^\top V G'' \delta^i \delta^{i^\top} Cb s_j \tag{57}$$

Eventually, we obtain

$$\delta u^\top \frac{\partial L(t-l)}{\partial C_{ij}} \delta s = \delta\mu_i \delta s_j + \kappa_i s_j \tag{58}$$

where G' is at time $t-l$ and

$$\delta\mu_i = \left(G'V^\top \delta u\right)_i, \text{ and } \kappa_i = \left(G''V^\top \delta u\right)_i (C\delta s)_i \tag{59}$$

Analogously we obtain with the h gradient

$$\delta u^\top \frac{\partial L(t-l)}{\partial h_i} \delta s = \kappa_i \tag{60}$$

In the case of $g(z) = \tanh(z)$ we find, using $G''(z) = -2G'(z)G(z)$, that

$$\kappa_i = -\gamma_i a_i, \text{ with } \gamma_i = 2\left(C\delta s\right)_i \delta\mu_i \tag{61}$$

since $g_i(z) = a_i$. The update rule reads in that case

$$\Delta C_{ij} = \varepsilon \delta\mu_i \delta s_j - \gamma_i a_i s_j \tag{62}$$

In the case of arbitrary neuron activation functions g we obtain equivalent formulae by writing the diagonal matrix G'' as

$$G'' = -WGG'$$

where the diagonal matrix W is given by $W = 2\mathbb{I}$ in the tanh case. In that way we can again factor the g and g' factors out of the κ term as with eq. (61). Then, eq. (62) is recovered with the definition

$$\gamma_i = -\frac{g_i''}{g_i' g_i} \left(C\delta s\right)_i \delta\mu_i \tag{63}$$

of the γ_i factors.

In the derivation of eq. (57) we ignored the dependence of the state z in $G'(z)$ on the parameters θ. This dependence can be considered explicitly if the state is at a fixed point. In that case, a more detailed discussion in Der and Martius [17] shows that the effect of the derivative can be condensed into the so-called "sense" parameter α multiplying γ, i. e. by replacing γ in eqs. (61) and (63) as

$$\gamma_i \leftarrow \alpha\gamma_i \tag{64}$$

where α is an empirical constant, typically $\alpha \geq 1$, by which the sensitivity of the sensorimotor dynamics to external perturbations can be regulated. This works also in more general cases like a limit cycle dynamics, see Der and Martius [17]. The update rules for h are obtained analogously.