Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Online Goal Babbling for rapid bootstrapping of inverse models in high dimensions

Matthias Rolf, Jochen Steil, Michael Gienger

2011

Preprint:

This is an accepted article published in IEEE International Conference on Development and Learning (ICDL) and on Epigenetic Robotics (EpiRob). The final authenticated version is available online at: https://doi.org/[DOI not available]

Online Goal Babbling for rapid bootstrapping of inverse models in high dimensions

Matthias Rolf and Jochen J. Steil
Research Institute for Cognition and Robotics (CoR-Lab)
Bielefeld University

Michael Gienger
Honda Research Institute Europe
Offenbach, Germany

***Abstract*—We present a model of online Goal Babbling for the bootstrapping of sensorimotor coordination. By modeling infants' early goal-directed movements we show that inverse models can be bootstrapped within a few hundred movements even in very high-dimensional sensorimotor spaces. Our model thereby explains how infants might initially acquire reaching skills without the need for exhaustive exploration, and how robots can do so in a feasible way. We show that online learning in a closed loop with exploration allows substantial speed-ups and, in high-dimensional systems, outperforms previously published methods by orders of magnitude.**

***Index Terms*—Goal Babbling, Sensorimotor Exploration**

I. Introduction

Infants are born without being able to perform the most basic motor skills like reaching for an object and must to learn to control their body during the course of development. Successful control of such tasks can be well understood with the notion of *internal models* [1]. Internal models describe relations between motor commands and their consequences. Once internal models are established for a certain task, a forward model predicts the consequence of a motor command, while an inverse model suggests a motor command necessary to achieve a desired outcome. Learning internal models from scratch requires *exploration*. In artificial systems, exploration is traditionally addressed by "motor babbling" [2], [3] in which motor commands are randomly selected and their consequences are observed. This kind of exploration becomes very inefficient with increasing dimensionality of the sensorimotor space. The exploration can be significantly improved by active learning schemes [4], [5], frequently discussed under the notion of intrinsic motivation [6]. Although the risk of generating uninformative examples can be reduced with these methods, they assume that the sensorimotor space can be entirely explored. However, high-dimensional motor systems like human ones *can not be entirely explored in a lifetime*.

Tasks in sensorimotor learning are typically substantially lower-dimensional than the motor systems themselves. Reaching for an object can be done in an infinite number of ways, because human bodies as well as modern humanoid robotic systems have more degrees of freedom than necessary to solve the task. While this redundancy is often considered a problem for sensorimotor learning [7], [8], it also reduces the demand for exploration. If there are multiple ways to achieve the same result there is no inherent need to know all of them. Infants make heavy use of this opportunity in the beginning of their sensorimotor development. They initially master the feedforward control of reaching movements [9], [10], and choose the same set of motor commands for a target every time. Only later on they are able to adapt their movements to various situations and incorporate visual error correction on the fly.

A. Goal-directed Exploration

Learning only one valid solution for reaching can be done with enormous data-efficiency in high dimensions, if appropriate training data is available [11]. If data has to be generated autonomously, the question is how to shape an exploration mechanism that allows to find appropriate solutions without the need to explore everything. It has been shown that infants explore by far not randomly or exhaustively as supposed by "motor babbling". Rather, they attempt goal-directed actions already days after birth [12], [13], which indicates a strong role of "learning by doing". Infants learn to reach by trying to reach. Such goal-directed exploration processes, or "Goal babbling" allow to focus on behaviorally relevant data. The conceptually simple character of Goal Babbling is particularly well suited to explain infants mastery of sensorimotor development. Only one mechanism is needed for exploration and control which alleviates the problem to decide when to stop exploratory behavior and to start acting. Goal Babbling describes an entirely incremental acquisition of sensorimotor coordination which even allows to compensate for ongoing morphological changes like growth [14].

Goal-directed exploration has been part of many learning schemes for inverse models, but only been possible with prior knowledge [15], [16] or non goal-directed pre-training [17], [18]. Only recently, models have been proposed that investigate a consistent, goal-directed bootstrapping of internal models for sensorimotor control. In [19], a partial forward model is learned that is analytically inverted for goal-directed feedback control. A full forward model, as well as a full feedback model can, however, only be learned with exhaustive exploration. We argue that the most direct way to deal with partial exploration of the sensorimotor space is to start with exactly one solution that is learned directly in a feedforward function. In [20], [14] we have introduced a model for learning such functions with Goal Babbling based on batch-gradient learning. We have shown that the approach allows to bootstrap

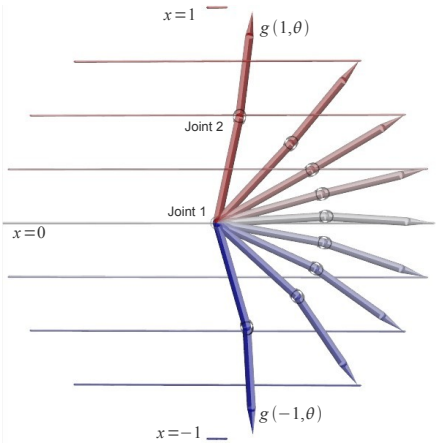


Fig. 1. A robot arm with two joints ($q = (q_{(1)}, q_{(2)}), m = 2$) is controlled to achieve a certain height of the effector ($n = 1$). An inverse estimate $g(x^*)$ suggests one posture for each desired height.

inverse kinematics for humanoid morphologies such as the Honda humanoid and iCub and in the presence of sensory noise or morphological changes. Although the approach does not need substantially more examples with increasing dimensions, the general exploratory cost is comparably high because a number of examples needs to be collected before a gradient update is performed in batch mode. In the present work, we propose an online formulation of this approach and demonstrate that it bootstraps inverse models by orders of magnitude more efficiently.

B. Online vs. Batch Learning of inverse Models

In machine learning tasks with fixed data sets online learning is typically regarded as a stochastic approximation of batch gradients. If online learning occurs during goal-directed exploration, the situation is different, because the example distribution changes over time depending on the learning dynamics. Example based online learning from goal-directed exploration has been investigated for the tuning of inverse models in various flavors like differential kinematics [17] or operational space control [18]. A central difficulty is that the online learning applies a positive feedback loop on the examples. Perturbations – either self-generated exploratory noise or from external sources – are reinforced by the learning which can cause the inverse model to get instable or drift.

The contribution of this paper is twofold. Firstly, we show that simple and developmentally plausible regularization methods allow to compensate such instabilities and allow for online learning even during an initial bootstrapping. Secondly, we show that online learning is even highly beneficial for bootstrapping: The same feedback loop that amplifies perturbations in the nullspace also amplifies the discovery of new positions in the operational space. Since each step in Goal Babbling is informed by the previous ones this allows substantial speedups in the overall learning process. The next section introduces our model in detail. In section III we investigate the speedups and the scalability experimentally.

II. Online Goal Babbling

In the present work, we investigate the kinematic control of redundant systems. Formally, we consider the relation between joint angles $q \in \mathbf{Q} \subset \mathbb{R}^m$ and effector poses $x \in \mathbf{X} \subset \mathbb{R}^n$, where m is the number of degrees of freedom (DOF) and n is the dimension of the target variable (e.g. $n = 3$ for the spatial position of a hand). The forward kinematics function $f(q) = x$ describes the unique and causal relation between both variables. An inverse model is a function $g(x^*) = q$ that computes joint angles for a given target $x^* \in \mathbb{X}^*$ such that the desired position is actually reached ($f(g(x^*)) = x^*$). This function has parameters θ that are adapted during learning. An example inverse estimate $g(x^*)$ is shown in Fig. 1.

The general idea for the learning of an inverse estimate is to explore sequences of motor commands q_t over timesteps t. These motor commands are applied and the resulting effector positions x_t are observed:

$$x_t \quad = \quad f(q_t). \tag{1}$$

Examples (x_t, q_t) can then be used for supervised adaption of the inverse estimate. In order to generate examples, Goal Babbling starts with an initial inverse estimate $g(x^*, \theta_0)$ that always suggests some comfortable home posture: $g(x^*, \theta_0) = const = q^{home}$. Then continuous paths of target positions x_t^* are iteratively chosen from the set $\mathbb{X}^*$. These targets are tried to reach with the inverse estimate as expressed in the fundamental equation of goal-directed exploration:

$$q_t \quad = \quad g(x_t^*, \theta_t) + E_t(x_t^*). \tag{2}$$

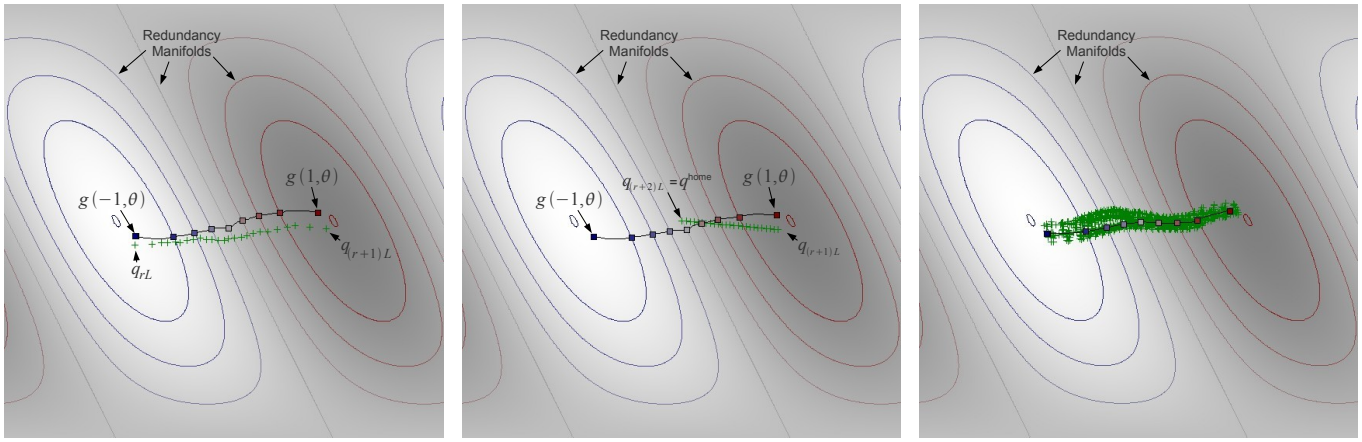
The outcome x_t is observed and the parameters θ_t of the inverse estimate are updated immediately before the next example is generated. The perturbation term $E_t(x^*)$ adds exploratory noise in order to discover new positions or more efficient ways to reach for the targets. This allows to unfold the inverse estimate and finally find correct solutions for all positions in the set of targets $\mathbb{X}^*$.

A. Path Generation

A major aspect of Goal Babbling is how to choose target positions. We do so by generating continuous, piecewise linear target movements. The initial target ($t = 0$) is the effector position belonging to the home posture: $x_0^* = f(q^{home})$. In the first movement, the system tries to move to another target x_L^* which is randomly chosen from $\mathbb{X}^*$. Between the timesteps 0 and L, the target positions are defined by the linear sequence between x_0^* and x_L^*. Afterwards a new target x_{2L}^* is chosen from $\mathbb{X}^*$ and the second movement is attempted from x_L^* to x_{2L}^*. Generally this pattern is denoted as

$$x^*_{r \cdot L + l} = \frac{L - l}{L} \cdot x^*_{r \cdot L} + \frac{l}{L} \cdot x^*_{(r+1) \cdot L}, \tag{3}$$

where r is the number of already executed movements, L is the length of each path and $l = 1 \dots L$ denotes the substeps within one movement. Throughout this paper we use $L = 25$. An example is generated for each of these targets according



(a) A linear target path shown in the joint space. (b) A subsequent "home" path in the joint space. (c) 1000 successive examples.

Fig. 2. Online Goal Babbling in the joint space of example Fig. 1. The x-axis encodes the angle of the first joint, the y-axis the angle of the second joint. Colored redundancy manifolds show the joint angles that apply the same effector height. The inverse estimate is shown as markered manifold through the two-dimensional joint space. Explored examples are shown as small green crosses. Our exploration scheme involves two kinds of movements: (a) the inverse estimate is used for trying to move from a target x^*_{rL} (here -1) to some other target $x^*_{(r+1)L}$ (here $+1$). Due to the perturbation term E_t the explored postures are not exactly aligned with the solutions suggested by the inverse estimate. (b) the effector moves from the last actuated joint angles back to the home posture. Fig. (c) shows how the varying perturbation terms cover the local surrounding of the inverse estimate.

to equations 1 and 2. An exemplary movement generated in this way is shown in Fig. 2a.

In non-linear redundant domains it is generally possible to generate inconsistent examples with same effector pose but different joint angles. Learning from such examples leads to invalid solutions [16]. We have previously shown in [20] that the structure of goal-directed exploration allows to resolve such inconsistencies by means of a simple weighting scheme:

$$w_t^{dir} = \frac{1}{2}\left(1 + cos\sphericalangle(x_t^* - x_{t-1}^*,\ x_t - x_{t-1})\right) , \quad (4)$$

$$w_t^{eff} = ||x_t - x_{t-1}|| \cdot ||q_t - q_{t-1}||^{-1} , \quad (5)$$

$$w_t = w_t^{dir} \cdot w_t^{eff} . \quad (6)$$

w_t^{dir} measures whether the actually observed movement and the intended movement have the same direction. w_t^{eff} measures the the kinematic efficiency of the movement and assigns high weight to examples that achieve a maximum of effector movement with a minimum of joint movement. For learning, each example (x_t, q_t) is weighted by w_t. In addition to resolving inconsistencies, the weighting also regularizes the inverse estimate. Efficient movements will dominate the learning in the long term and cause the inverse estimate to select smooth and comfortable solutions.

A special kind of movement is used to regularize the inverse estimate and prevent drifts into irrelevant regions of the sensorimotor space. With a probability p^{home} (0.1 throughout this paper), the next movement after a target x^*_{rL} has been applied is not another goal-directed movement. Instead, the system returns to its home posture. Similar to infants practicing their motor skills, the system returns to a stable point after a while and starts to practice again. This kind of movement leads to a repetitive presentation of examples close to the home posture and forces the inverse estimate to reproduce these postures for goal-directed movements. It acts as a developmentally plausible stabilizer that helps to stay in known areas of the sensorimotor space [19], [20]. We model this movement as a linear path in joint space in order to get smooth and continuous behavior for online learning: the system moves from the last actuated joint angles $q_{r \cdot L}$ to its home posture q^{home}, whereas Eqn. 2 is replaced by the following expression:

$$q_{r \cdot L + l} = \frac{L - l}{L} \cdot q_{r \cdot L} + \frac{l}{L} \cdot q^{home} . \quad (7)$$

For every generated joint configuration, the resulting effector pose is observed (Eqn. 1) and learning is applied online in the same way as for goal-directed movements. These examples are only weighted with w_t^{eff}, because targets x_t^* for the evaluation w_t^{dir} do not exist during this homeward movement. After the home posture has been reached, a goal-directed movement is attempted from the initial target $x^*_{(r+1) \cdot L} = f(q^{home})$. An example of this movement type is shown in Fig. 2b.

B. Structured Continuous Variation

In order to find kinematic solutions for all target positions, it is necessary to consider exploratory noise, or rather perturbations of the motor system. Such perturbations arise naturally in physical systems and lead to the exploration of new postures that would not be suggested by the inverse estimate. Physical perturbations typically lead to smooth variations of the intended movements. At any point in time we model this effect by adding a small, randomly chosen linear function to the inverse estimate.

$$E_t(x^*) = A_t \cdot x^* + b_t, \quad A_t \in \mathbb{R}^{m \times n},\ b_t \in \mathbb{R}^m \quad (8)$$

Initially, all entries e_0^i of the matrix A_0 and the vector b_0 are chosen i.i.d. from a normal distribution with zero mean and variance σ^2. In order to explore the local surrounding of

the inverse estimate, we vary these parameters slowly with a normalized Gaussian random walk. A small value δ^i_{t+1} is chosen from a normal distribution $N(0, \sigma^2_\Delta)$ with $\sigma^2_\Delta \ll \sigma^2$, and added to the previous value e^i_t. The variance of the resulting value is the sum of the individual variances $\sigma^2+\sigma^2_\Delta$. We normalize with the factor $\sqrt{\sigma^2/(\sigma^2+\sigma^2_\Delta)}$ to keep the overall deviation stable at σ:

$$\begin{aligned} e^i_0 &\sim N(0,\sigma^2), \qquad \delta^i_{t+1} \sim N(0,\sigma^2_\Delta)\ , \\ e^i_{t+1} &= \sqrt{\frac{\sigma^2}{\sigma^2+\sigma^2_\Delta}} \cdot (e^i_t + \delta^i_{t+1}) \sim N(0,\sigma^2)\ . \end{aligned}$$

Hence, $E_t(x^*)$ is a slowly changing linear function. It is smooth at any time, which is important for the evaluation of the weighting scheme (Eqn. 4 and 5). It is furthermore zero centered and limited to a fixed variance which leads to a local exploration around the inverse estimate (see Fig. 2c).

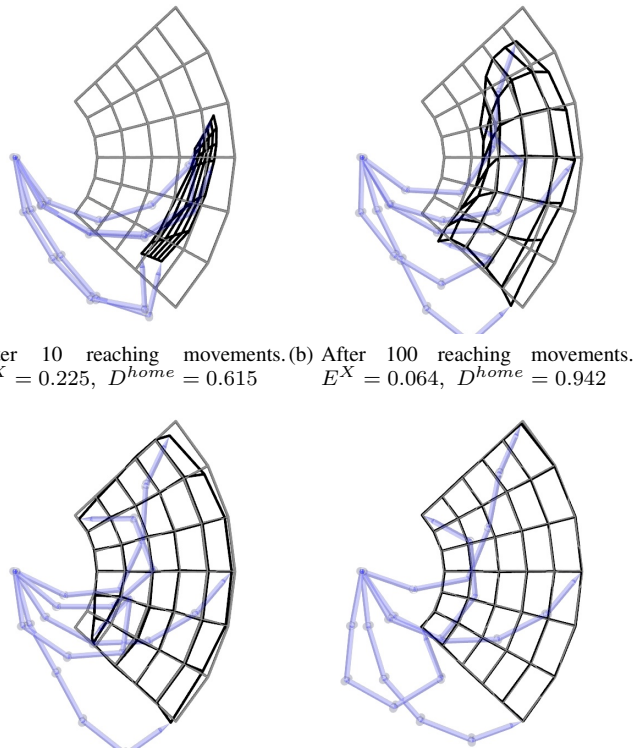
C. Incremental Regression Model

For learning, a regression mechanism is needed in order to represent and adapt the inverse estimate $g(x^*)$. The goal-directed exploration itself does not require particular knowledge about the functioning of this regressor, such that in principal any regression algorithm can be used. For a safe and incremental online learning we have chosen a *local-linear map* [21] for our experiments. The inverse estimate consists of different linear functions $g^{(k)}(x)$, which are centered around prototype vectors $p^{(k)}$ and active only in its close vicinity which is defined by a radius d. The function $g(x^*)$ is a linear combination of these local linear functions, weighted by a Gaussian responsibility function $b(x)$:

$$\begin{aligned} g(x^*) &= \frac{1}{n(x^*)} \sum_{k=1}^{K} b\left(\frac{x^*-p^{(k)}}{d}\right) \cdot g^{(k)}\left(\frac{x^*-p^{(k)}}{d}\right) \\ b(x) &= \exp\left(-||x||^2\right) \\ n(x) &= \sum_{k=1}^{K} b\left(\frac{x-p^{(k)}}{d}\right) \\ g^{(k)}(x) &= W^{(k)} \cdot x + o^{(k)} \end{aligned}$$

The normalization $n(x)$ scales the sum of influences of the components to unity, which is known as *soft-max*.

The inverse estimate is initialized with a single local function with center $p^{(1)} = f(q^{home})$, that outputs the constant value q^{home} ($W^{(1)} = 0_{m\times n}$ and $o^{(1)} = q^{home}$). New local functions and their prototype vectors are added dynamically. Whenever the learner receives an input x, that has a distance of at least d to all existing prototypes, a new prototype $p^{K+1} = x$ is created. In order to avoid abrupt changes in the inverse estimate, the function $g^{K+1}(x)$ is initialized such that its insertion does not change the local behavior of $g(x^*)$ at the position x. We set the offset vector o^{K+1} to the value of the inverse estimate before the insertion of the new local function: $o^{K+1} = g(x)$. The weight matrix is initialized with the Jacobian matrix $J(x) = \frac{\partial g(x)}{\partial x}$ of inverse estimate: $W^{K+1} = J(x)$.



(a) After 10 reaching movements. $E^X = 0.225,\ D^{home} = 0.615$
(b) After 100 reaching movements. $E^X = 0.064,\ D^{home} = 0.942$
(c) After 1000 reaching movements. $E^X = 0.011,\ D^{home} = 0.995$
(d) After 10000 reaching movements. $E^X = 0.002,\ D^{home} = 0.737$

Fig. 3. Example of the bootstrapping dynamics for a five DOF arm with learning rate $\eta = 0.1$. The inverse estimate rapidly finds valid solutions as the actual position (black grid) become congruent with the targets (gray grid). Blue posture show how the redundancy is resolved.

In each timestep, the inverse estimate is fitted to the currently generated example (x_t, q_t) by reducing the weighted square error:

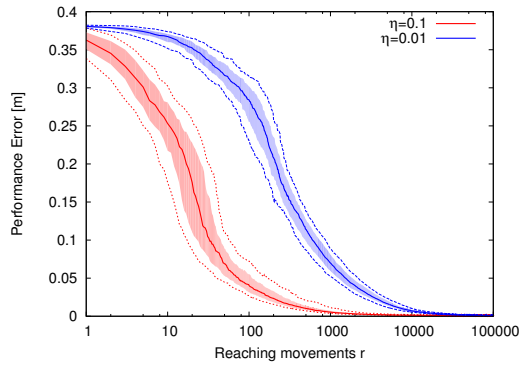
$$E^Q_w = w_t \cdot ||q_t - g(x_t)||^2$$

The parameters $\theta = \{W^{(k)}, o^{(k)}\}_k$ of $g(x^*)$ are updated using online gradient descent on E^Q_w with a learning rate η:

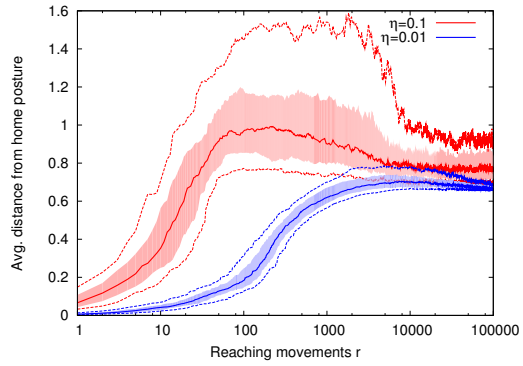
$$W^{(k)}_{t+1} = W^{(k)}_t - \eta \frac{\partial E^Q_w}{\partial W^{(k)}}\ , \qquad o^{(k)}_{t+1} = o^{(k)}_t - \eta \frac{\partial E^Q_w}{\partial o^{(k)}}$$

III. Experiments

We use simulated planar robot arms as an illustrative test case to investigate online Goal Babbling. The aim is to control the effector-position in the 2D plane ($n = 2$). An example with five degrees of freedom is shown in Fig. 3. The target positions x^* are arranged in the gray grid structure. The black grid shows the actually reached positions ($x = f(g(x^*))$). Initially, the inverse estimate is fixed at the home position, but expands rapidly towards the target positions. After a number of movements, target and actual grids are in congruence. An accurate inverse estimate has been bootstrapped. Blue postures show configurations generated by the inverse estimate for several different target positions and thus how the redundancy is resolved.



(a)



(b)

Fig. 4. Statistics of the bootstrapping dynamics for two different learning rates. (a) The performance error E^X decreases rapidly over the number of movements. A ten times higher learning rate results in a speed up of approx. 20. (b) The distance from the home posture initially increases as the inverse estimate unfolds. High learning rates η initially select less comfortable solutions which are then gradually optimized.

We are interested in three different experimental measures to assess the characteristics of the bootstrapping:

1) *Accuracy* of the bootstrapped inverse models.
2) *Comfort* of the selected solution.
3) *Speed* of the bootstrapping process.

The accuracy is measured as the average distance between the target positions $x_i^* \in \mathbb{X}^*$ and the actually reached positions:

$$E_r^X = \frac{1}{N} \sum_{i=1}^{N} ||x_i^* - f(g(x_i^*, \theta_{rL}))|| \tag{9}$$

As a measure of comfort we compute how far the suggested postures are away from the home posture:

$$D_r^{home} = \frac{1}{N} \sum_{i=1}^{N} \left|\left|q^{home} - g(x_i^*, \theta_{rL})\right|\right| \tag{10}$$

This measure can not be zero for a bootstrapped model, because the home posture has to be left in order to reach for different targets. Nevertheless it allows to compare how comfortably different inverse estimates resolve the redundancy.

We assess the speed of bootstrapping by measuring the number of movements until a certain percentage of independent trials has reached some accuracy level:

$$S(Q, e^X) = \underset{r}{argmin}(\ Q \leq p(E_r^X \leq e^X)\ ) \tag{11}$$

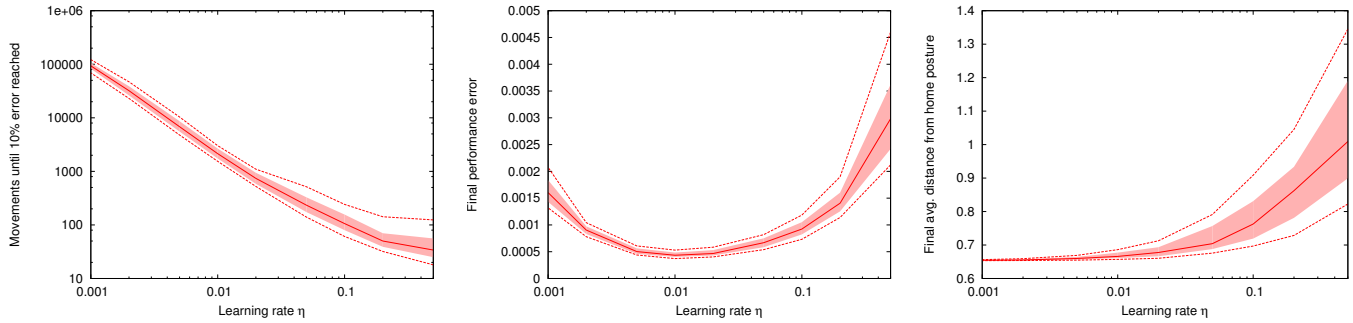
For instance, $S(0.9, 0.1)$ counts the number of reaching movements, until 90% of the trials have reached a performance error below or equal to 0.1. The statistics presented in this section are all computed over 100 independent trials.

A. Effects of the Learning Rate

The most important variable for online learning from goal-directed exploration is the learning rate η. In supervised learning from fixed data sets, online learning is used as stochastic approximation of batch methods. In goal-directed exploration, however, the data set is not fixed but continuously constructed by the learner. This interweaved relation of data generation and learning leads to non-trivial effects with respect to the choice of the learning rate.

We used parameters $\sigma = 0.05$, $\sigma_\Delta = 0.005$ and $d = 0.1$ for our experiments. The home posture is defined by setting the first joint to $-\frac{\pi}{3}$ and the remaining joints to $\frac{\pi}{6}$, which corresponds to a slightly bent shape with the effector at zero height. Fig. 4a shows the development of the performance error E_r^X over the number of movements r for the 5 DOF planar arm with a total length of $1m$. Bold lines show the median error, thin lines the 10% and 90% quantiles and the filled areas correspond to the range between the 25% and 75% quantiles. For both $\eta = 0.1$ and $\eta = 0.01$ the error decreases reliably and we obtain an accurate inverse model. Obviously the bootstrapping is faster for the higher learning rate, but the speedup does not scale with the factor 10 between the learning rates. For $\eta = 0.1$ the performance error has reached a median level of 0.04 after 100 movements. For $\eta = 0.01$ it takes 2000 movements to reach the same error level. Hence, the bootstrapping is 20 times faster for the high learning rate, although the rate itself is only 10 times higher. The reason for this non-trivial speedup is the incremental character of the goal-directed data-generation. Because the creation of each example is already informed by learning from the previous examples, learning does not only improve the inverse estimate, but will also result in a more informative next example. In an online scenario, this causes a positive feedback loop. Higher learning rates imply a higher "gain" in this loop and accelerate the bootstrapping over the sheer values of the learning rates.

The distance from the home posture D_r^{home} for the same trials is shown in Fig. 4b and displays another qualitative effect of the learning rate. Low learning rates let the distance from the home posture increase gradually as the inverse estimate unfolds. It finally reaches a stable level which corresponds to a comfortable solution. High learning rates, in contrast, cause a rapid increase with high variance. The bootstrapping initially sticks to the very first solution that is observed due to the random perturbation term. This can result in a less comfortable redundancy resolution. After several thousand movements, the distance decreases again as comfortable solutions receive higher weights w^{eff} and dominate the learning in the long term.

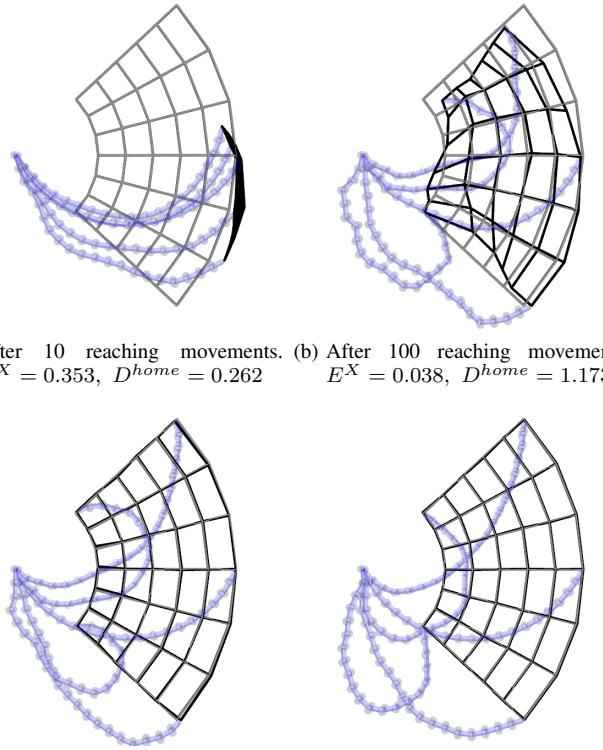


(a) (b) (c)

Fig. 5. Bootstrapping results for various learning rates between 0.001 and 0.5. (a) The number of movements needed to reach 10% of the initial error decreases rapidly for higher learning rates. (b) The performance error after 10^7 movements is very low for all learning rates. Very low learning rates are not fully converged. (c) The final distance from the home posture increases gradually for higher learning rates.

An example trial for $\eta = 0.1$ is shown in Fig. 3. Already after 10 movements the inverse estimate has expanded from the home posture and is roughly aligned with the correct movement directions, and rapidly expands further. After 1000 movements, the inverse estimate starts to consolidate the redundancy resolution and the selected postures become closer to the home posture.

Results for a high range of learning rates $[0.001; 0.5]$ are summarized in Fig. 5. The bootstrapping speed is continuously increased for higher learning rates. Fig. 5a shows the number of movements, until the performance error is reduced 10% of its initial value ($S(Q, 0.1 \cdot E_0^X)$, quantiles Q shown are 10, 25, 50, 75, 90). For the highest rate $\eta = 0.5$, 50% of the trials have reached this level already after 34 movements ($S(0.5, 0.1 \cdot E_0^X) = 34$). After a total number of 10^7 movements the trials for all learning rates have reached a performance error in the millimeter range (Fig. 5b). For very low learning rates the inverse estimates are not fully converged after that time, as indicated by the slightly increased error. For high learning rates both final performance error and the home posture distance (Fig. 5c) increase gradually. The positive feedback loop between exploration and learning, which causes the speedup effects, also starts to destabilize the learning at high rates. Nevertheless the results are in a very good range due to the regularization by the home posture and the weighting scheme.



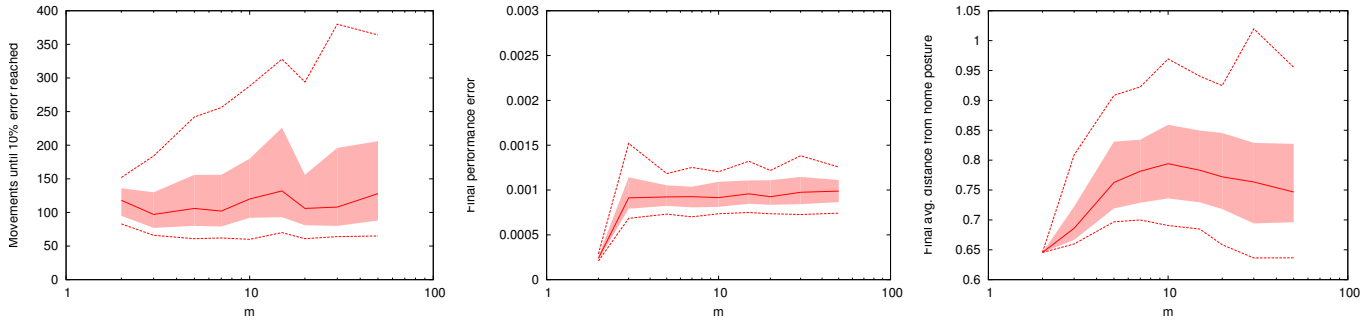
(a) After 10 reaching movements. $E^X = 0.353$, $D^{home} = 0.262$
(b) After 100 reaching movements. $E^X = 0.038$, $D^{home} = 1.173$
(c) After 1000 reaching movements. $E^X = 0.004$, $D^{home} = 0.940$
(d) After 10000 reaching movements. $E^X = 0.002$, $D^{home} = 0.922$

Fig. 6. Example of the bootstrapping dynamics for 20 degrees of freedom. The inverse estimate unfolds with high speed also in high dimensions. The selected postures get smoother and more comfortable over time.

B. Scalability up to 50 DOF

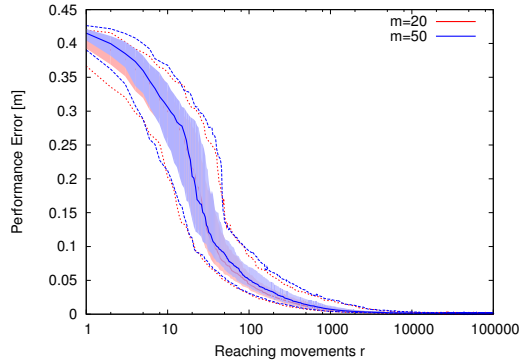
The outstanding property of Goal Babbling is its scalability to many dimensions, which we investigate in the following experiment. For a direct comparison to the previous experiment with five degrees of freedom we investigate the same setup, but split the arm in m segments of equal length, each actuated by one joint. Hence, an arm with 20 degrees of freedom comprises 20 segments of length $0.05m$. The home posture is chosen as $-\frac{\pi}{3}$ for the first joint and $\frac{2\pi}{3(m-1)}$ for the remaining joints. The target positions are identical to those in the first experiment as indicated in Fig. 6. For a maximum of comparability, we need to scale down the perturbation term: if the variability per joint is constant, it has a higher effect of the end effector for high dimensional systems. This leads to a faster discovery of effector positions but also more instability. We can approximate the deviation at the effector σ^X for an entirely stretched arm as a function of the joint variability σ and the number of DOF m:

$$\sigma^X = \sigma \cdot \sqrt{\frac{m+1}{2}} \ . \tag{12}$$

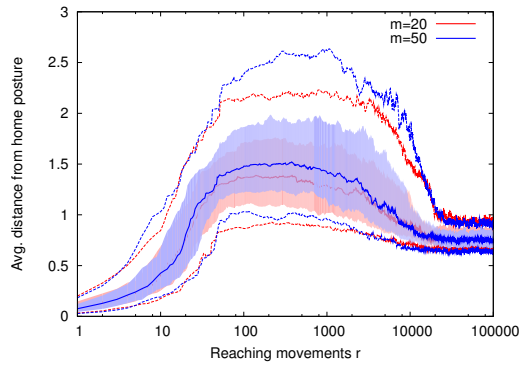


(a) (b) (c)

Fig. 8. Bootstrapping results for various numbers of joints. (a) The number of movements needed to reach 10% of the initial error increases only very gradually. (b) The performance error after 10^6 movements is very low in all cases. (c) The final distance from the home posture.



(a)



(b)

Fig. 7. Statistics of the bootstrapping dynamics for 20 and 50 degrees of freedom. Both performance error (a) and home posture distance (b) show a very similar behavior for 20 and 50 DOF. Goal Babbling scales without substantial extra cost in high dimensions.

For this experiment we scale σ such that σ^X is constant at $0.05 \cdot \sqrt{3}$ which is the variability in the first experiment. The update parameter is set to $\sigma^\Delta = 0.1 \cdot \sigma$. The learning rate is set to $\eta = 0.1$.

An example trial for $m = 20$ is shown in Fig. 6. The behavior over time, and in particular the speed of bootstrapping, is very similar to the previous five DOF example. The performance error is reduced very rapidly during the first 100 movements. After 1000 movements the inverse estimate is already very accurate, but does not yet use optimally comfortable joint configurations. These are further optimized in the following movements as the configurations get smoother and the avg. distance to the home posture decreases. Fig. 7 shows a comparison between $m = 20$ and $m = 50$ over time. The temporal characteristics of the performance error are virtually identical in the two cases and also compared to the $m = 5$ experiment (see Fig. 4). Also the home distance values show the same behavior, with slightly increased values for $m = 50$ in the intermediate movements.

Results for values of m between 2 and 50 are summarized in Fig. 8. The most important result is that the average bootstrapping speed is virtually constant across the entire range of m. Even for 50 degrees of freedom, 50% of the trials have reached the 10% error level after after 128 movements ($S(0.5, 0.1 \cdot E_0^X) = 128$). However, the distribution becomes increasingly heavy-tailed as the values for the 90% quantile $S(0.9, 0.1)$ grow very slowly (e.g. $S(0.9, 0.1) = 364$ for $m = 50$). After a total number of 10^6 movements the performance error is approximately constant and very low at $1mm$. Only for $m = 2$ it is even lower with almost zero variance. Here the problem does not contain local redundancy, but only two separated choices "elbow up" and "elbow down" that can not be flipped by local perturbations. Higher values of m allow to modify the redundancy resolution continuously, which causes minor local averaging errors.

IV. Discussion

High dimensional sensorimotor spaces can not be entirely explored in a lifetime. Goal-directed exploration processes provide a developmentally plausible and technically powerful way to circumvent this problem by focusing on behaviorally relevant regions in the sensorimotor space. Our experiments show that inverse models can be entirely bootstrapped with high speed in very high dimensional domains. Thereby the problematic amplification of perturbations is regularized by our weighting scheme and the consistent use of a home posture. The results show that online learning during Goal Babbling is not only possible, but highly beneficial. The search

for solutions is boosted in the feedback loop of exploration and learning along continuous paths. This allows a very rapid discovery of valid solutions within a few hundred movements, which is a time span also required by humans to learn new sensorimotor mappings [22]. In contrast, our previous batch-gradient model [20] required several thousand epochs, each involving a few hundred movements, to reach a comparable accuracy. Also the model in [23] was reported to take several ten-thousand to hundred-thousand movements on a 15 DOF planar arm to reach a coarse level of coordination. A precise numeric comparison between these models is certainly delicate, because they have entirely different parametrizations. The actual results, however, show that our present model outperforms these previous approaches by several orders of magnitude.

The learning of just one solution might appear to be a deficit and a lack of flexibility. In contrast to an exhaustive exploration it is, however, systematically possible in high dimensions. The number of required movements increases only very gradually with the number of dimensions in which the exploration is performed. Our experiments show an almost constant behavior while an exhaustive motor babbling approach would require an exponential increase of the exploration to cover the sensorimotor space. Infants need to master a very high dimensional motor system and develop on a similar pathway: starting with early goal-directed movements [12] they initially master feedforward control before reaching movements become adaptive and feedback control can be applied [9], [10]. The mastery of feedforward control *before* adaptive feedback mechanisms can be applied seems very natural. Yet, traditional models of sensorimotor learning like feedback-error learning [15] and learning with distal teacher [16] fail to explain this ordering of skill acquisition. Both models require full knowledge about the sensorimotor space which can then be used to extract a feedforward controller. In contrast, our model of Goal Babbling shows how a coherent solution can be bootstrapped without prior knowledge by means of a simple "trying to reach" method that is developmentally plausible and technically feasible. For high learning rates we find that initially one solution is picked that might be uncomfortable. Later on it becomes smooth and more comfortable, which is an efficient developmental path also observed in infants [24].

Although the learning of a single valid solution does not permit instantaneous adaptation to novel situations, it allows to adapt to ongoing changes such as body growth [14]. In fact, the role of feedforward control does not diminish in adult sensorimotor control, which is well known from prism-glass experiments [25]. An important objective for future research is to continue the developmental pathway with learning architectures that allow for more adaptivity. Here the rapid bootstrapping needs to be combined with the expressiveness and flexibility of other models such as [26]. This describes an efficient goal-directed pathway on which at first valid solutions are discovered as fast as possible and which are later on augmented as it is necessary to achieve desired results.

ACKNOWLEDGEMENTS

Matthias Rolf gratefully acknowledges the financial support from Honda Research Institute Europe for the Project "Neural Learning of Flexible Full Body Motion".

REFERENCES

[1] D. Wolpert, R. C. Miall, and M. Kawato, "Internal models in the cerebellum," *Trends in Cog. Sci.*, vol. 2, no. 9, 1998.
[2] P. Gaudiano and D. Bullock, "Vector associative maps unsupervised real-time error-based learning and control of movement trajectoires," *Neural networks*, vol. 4, no. 2, pp. 147–183, 1991.
[3] Y. Demiris and A. Dearden, "From motor babbling to hierarchical learning by imitation: A robot developmental pathway," in *EpiRob*, 2005.
[4] R. Martinez-Cantin, M. Lopes, and L. Montesano, "Body schema acquisition through active learning," in *ICRA*, 2010.
[5] A. Baranes and P.-Y. Oudeyer, "Robust intrinsically motivated exploration and active learning," in *ICDL*, 2009.
[6] E. L. Deci and R. M. Ryan, *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum Press, 1985.
[7] N. Bernstein, *The Co-ordination and Regulation of Movements*. Pergamon Press, 1967.
[8] D. Bullock, S. Grossberg, and F. H. Guenther, "A self-organizing neural model of motor equivalent reaching and tool use by a multijoint arm," *Cognitive Neuroscience*, vol. 5, no. 4, 1993.
[9] E. W. Bushnell, "The decline of visually guided reaching during infancy," *Infant Behavior and Development*, vol. 8, no. 2, 1985.
[10] L. Out, A. J. van Soest, G. P. Savelsbergh, and B. Hopkins, "The effect of posture on early reaching movements," *Journal of Motor Behavior*, vol. 30, no. 3, 1998.
[11] M. Rolf, J. J. Steil, and M. Gienger, "Efficient exploration and learning of whole body kinematics," in *ICDL*, 2009.
[12] C. von Hofsten, "Eye-hand coordination in the newborn," *Developmental Psychology*, vol. 18, no. 3, 1982.
[13] L. Ronnquist and C. von Hofsten, "Neonatal finger and arm movements as determined by a social and an object context," *Early Development and Parenting*, vol. 3, no. 2, pp. 81–94, 1994.
[14] M. Rolf, J. J. Steil, and M. Gienger, "Mastering growth while bootstrapping sensorimotor coordination," in *EpiRob*, 2010.
[15] M. Kawato, "Feedback-error-learning neural network for supervised motor learning," in *Advanced Neural Computers*. Elsevier, 1990.
[16] M. Jordan and D. Rumelhart, "Forward models: supervised learning with distal teacher," *Cognitive Science*, vol. 16, pp. 307–354, 1992.
[17] A. D'Souza, S. Vijayakumar, and S. Schaal, "Learning inverse kinematics," in *IROS*, 2001.
[18] J. Peters and S. Schaal, "Reinforcement learning by reward-weighted regression for operational space control," in *ICML*, 2007.
[19] A. Baranes and P.-Y. Oudeyer, "Maturationally-constrained competence-based intrinsically motivated learning," in *ICDL*, 2010.
[20] M. Rolf, J. J. Steil, and M. Gienger, "Goal babbling permits direct learning of inverse kinematics," *IEEE Trans. Auto. Mental Development*, vol. 2, no. 3, 2010.
[21] H. Ritter, "Learning with the self-organizing map," in *Artificial Neural Networks*, T. Kohonen, Ed. Elsevier Science, 1991.
[22] U. Sailer, J. R. Flanagan, and R. S. Johansson, "Eye–hand coordination during learning of a novel visuomotor task," *Journal of Neuroscience*, vol. 25, no. 39, pp. 8833–8842, 2005.
[23] A. Baranes and P.-Y. Oudeyer, "Intrinsically motivated goal exploration for active motor learning in robots: A case study," in *IROS*, 2010.
[24] N. E. Berthier and R. Keen, "Development of reaching in infancy," *Experimental Brain Research*, vol. 169, no. 4, 2005.
[25] J. Baily, "Adaptation to prisms: Do proprioceptive changes mediate adapted behaviour with ballistic arm movements?" *The Quarterly Journal of Experimental Psychology*, 1972.
[26] M. V. Butz, O. Herbort, and J. Hoffmann, "Exploiting redundancy for flexible behavior: unsupervised learning in a modular sensorimotor control architecture," *Psychological Review*, vol. 114, no. 4, 2007.