Learning Inverse Models in High Dimensions with Goal Babbling and Reward-Weighted Averaging

Matthias Rolf
Department of Adaptive Machine Systems
Osaka University
matthias@ams.eng.osaka-u.ac.jp

Minoru Asada
Department of Adaptive Machine Systems
Osaka University
asada@ams.eng.osaka-u.ac.jp

Abstract

We discuss the difficulty to learn control skills in high-dimensional domains that can not be exhaustively explored. We show how infant-development can serve as a role-model for highly efficient exploration: infants neither explore exhaustively, nor do they learn very versatile skills right in the beginning. Rather, they attempt goal-directed exploration to achieve feedforward control as a first step without requiring full knowledge of the world. This article reviews recent efforts to mimick such pathways by means of "goal babbling", which have led to a series of algorithms that allow for a likewise efficient learning. We show that it permits to learn inverse models from examples even in the presence of non-convex solution sets by utilizing a reward-weighted regression scheme, and that a human-competitive learning speed can be achieved if online learning is applied "in the loop". Results are verified on the "Bionic Handling Assistant", a novel bionic robot that instantiates a wide spread of problems like high dimensions, non-stationary behavior, highly constrained actuators, sensory noise, and very slow response-behavior.

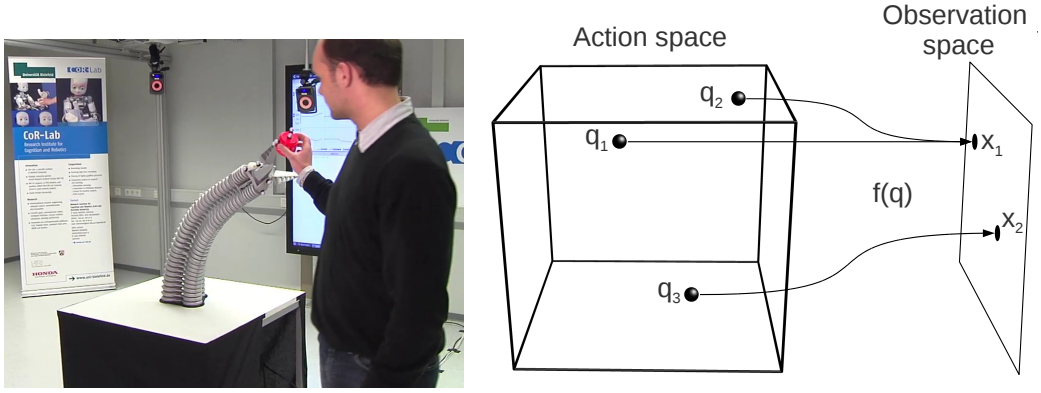
1 Learning Internal Models for Motor Control

New generations of bionic robots combine mechanical flexibility, elastic material, and lightweight actuation like pneumatics. Such robots are often inspired by biological actuators like octopus arms [1], elephant trunks [2] (see figure 1a), or human biomechanics [3], and provide enormous potential for the physical interaction between the robot and the world, and in particular between robots and humans. The downside of their biologically inspired design is that analytic models for their control are hardly available and difficult to design. This qualifies learning as an essential tool for their successful application. Yet, these robots not only challenge analytic control, but also motor learning. They typically comprise many degrees of freedom that can not be exhaustively explored, delayed feedback due to slow pneumatic actuation, and often non-stationary system behavior. We therefore argue for a new paradigm of motor learning that leaves exhaustive exploration, which cannot be achieved for problems of such scale, behind. Rather, we *focus on the achievable* by drawing inspiration from infant development in order to achieve reasonable feedforward-controlled coordination skills that can be learned very efficiently when mimicking infants' exploratory behavior.

As a general framework for motor learning, we can consider an agent that can execute motor commands or actions $q \in \mathbf{Q}$, where $\mathbf{Q}$ is the *action space*. Each action causes an outcome (e.g. an effector position) $x \in \mathbf{X}$ in some *observation space*. The unique causal relation between both variables is formally defined by some forward function f:

$$f : \mathbf{Q} \to \mathbf{X}, \;\; f(q) = x \tag{1}$$

The control-, or *coordination problem* is then to invert this relation in any possible way: achieving some desired outcome, or *goal* $x^* \in \mathbf{X}^*$ out of a set $\mathbf{X}^* \subseteq \mathbf{X}$ requires to estimate an appropriate action $\hat{q}$ that results in the observation of x^* ($f(\hat{q}) = x^*$).



(a) The Bionic Handling Assistant (b) A forward function connects action and observation space

Figure 1: Left: Goal babbling allows to efficiently learn reaching with the *Bionic Handling Assistant*, which mimics an elephant trunk. Right: Action and observation space are connected by a forward function f that maps actions to their causal outcome.

Learning Forward Models for Control Forward models $\hat{f}$ are predictors that can be used to predict the outcome of some action q without actually performing it. Forward models can *not* solve the coordination problem *directly*. Indirect mechanisms to use them for coordination are, however, widely used, and define a process, such as inverse Jacobian control [4], that dynamically *searches* for an appropriate action q by using the known output and shape of $\hat{f}$ [5]. The learning of forward models for sensorimotor coordination is a heavily investigated and widely used method in motor learning literature, which allows to resemble control mechanisms typically used for the control of robots with analytically known forward functions. The actual learning appears to be a standard *regression* problem: (i) There is a ground truth functional relation f that is to be approximated by the learned forward model $\hat{f}$. (ii) For any input q of the model, the correct output x (or in stochastic domain the output distribution $P(x|q)$) can be queried by executing the forward function. Hence, it is possible to collect a data set $D = \{(q_l, x_l)\}_l$ and learn the forward model, parameterized with some adaptable parameters θ, by reducing the *prediction error* E^P on the data set

$$E^P(D, \theta) = \frac{1}{L} \sum_{l=0}^{L-1} ||\hat{f}(q_l, \theta) - x_l||^2 \approx \int_q ||\hat{f}(q, \theta) - f(q)||^2 P(q) dq \,, \tag{2}$$

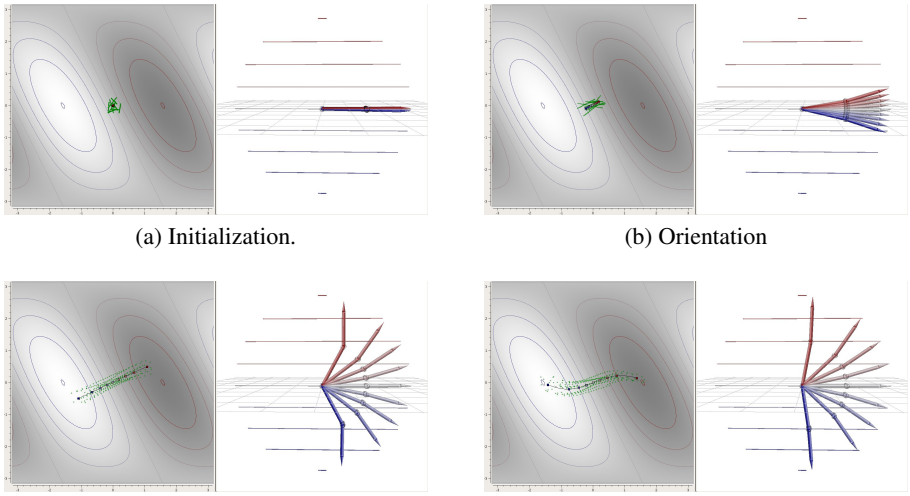
which approximates the expected error on the input distribution $P(q)$. This view of the input distribution exposes a central difference between forward model learning for control and standard regression problems: $P(q)$ usually corresponds to some (at least empirically) known real world distribution that expresses how likely, and thus relevant, certain inputs are to the learner. For the learning of a coordination skill, it is usually not known which actions are relevant to the solution of the coordination problem. Which actions *will* be used in fact depends on the search mechanism that is used on top of the forward model. Since the "true" distribution $P(q)$ during control is unavailable, the standard approach is to assume a uniform distribution, which corresponds to an *exhaustive* sampling in the action space. Hence, sample actions are drawn in a random manner [6, 7], which is often referred to as "motor babbling" [8]. The exhaustive sampling of actions can usually be done in low-dimensional domains. Yet, it does not provide a feasible method when $\mathbf{Q}$ is high-dimensional. Several approaches have been suggested to improve the feasibility of learning forward models for coordination. One is concerned with the incorporation of *prior knowledge*. Ulbrich *et al.* reported an approach that allows to make an exact match $\hat{f} = f$ for robots with only revolute joints. Their approach allows to exactly pinpoint the number of examples needed to 3^m [9], which is still far out of reach for high-dimensional motor systems. Another approach utilizes *active learning*, which has been shown to reduce the absolute number of examples necessary to learn forward models [10]. While it can avoid the generation of entirely uninformative examples, it can not avoid the exhaustive character of the exploration. Active learning still aims at the error reduction over the entire input distribution $P(q)$ of the learner [11].

Learning Differential Inverse Models for Control Forward models are often used for establishing a feedback controller that suggests actuator velocities $\dot{q}$ for a desired effector movement $\dot{x}^*$, based on the forward model's inverse Jacobian. This relation can also be learned directly in a differential *inverse* model g, which maps $\dot{x}^* \rightarrow \dot{q}$ [12]. The necessary actuator velocity depends on the current state, e.g. the posture q, such that the inverse model needs to be learned with respect to that dependency: $g(\dot{x}^*, q) = \dot{q}$. Similarly, learning the inverse dynamics $\ddot{x}^* \rightarrow \ddot{q}$ (or mapping on a motor torque τ directly) has been proposed [13, 14]. Therefore, the current actuator velocity $\dot{q}$ needs to be considered as dependency, which gives an inverse model $g(\ddot{x}^*, q, \dot{q}) = \ddot{q}$. If applicable, differential inverse models permit highly versatile feedback control. However, it is not always applicable for two reasons: Firstly, they critically rely on immediate feedback [15], which is not always available. Secondly, and more importantly, they inherit and *worsen* the scalability problems of forward model learning. Fully learning such models requires full knowledge about the action space, and even all combinations of actions $\dot{q}$/$\ddot{q}$ and there dependencies with with posture and velocity. Consequently, the approaches for learning such models start with an random exploration [12, 13], after which the model can be refined while performing goal-directed actions.

Learning Direct Inverse Models for Control Direct inverse models $g\colon x^* \rightarrow q$ represent a different class of control solutions. They directly suggest a motor command q for a goal x^*, which can be applied in a feedforward manner. While such feedforward control does not allow for such highly versatile control as feedback models, it has two potential advantages for difficult learning problems. Firstly, feedforward control is entirely insensitive to delayed, noisy, or missing feedback. Secondly, direct inverse models *do not require full knowledge* about the sensorimotor space if the control problem has redundancy: only a single solution q for any goal x^* must be known, even if infinitely many other solutions exist. Such models can be learned from very few examples if appropriate data exists [16]. Learning direct inverse models has been attempted in two different ways. *Error-based* methods measure the current performance error $E^X = (x^* - f(g(x^*)))^2$ between a goal x^* and the outcome $f(g(x^*))$ of trying to reach it. They then attempt to adapt the inverse model by means of gradient descent on E^X. The complication of this approach is that computing the gradient of E^X requires *prior knowledge*, because differentiating E^X requires to know the Jacobian of f. This knowledge alone could solve the control problem by means of feedback control. In *feedback-error learning* [17] it is simply assumed that a mechanism to derive the gradient, and thus a feedback controller, is already given. Learning with *distal teacher* [18] avoids a pre-existing controller, but requires to first exhaustively learn a forward model $\hat{f}$. Differently from error-based methods, *example-based* methods collect exploratory data (x, q), and attempt to adapt the inverse model by fitting it to the data, i.e. minimizing $E^Q = (q - g(x))^2$. This method can be successful for simple and low-dimensional control problems by fitting randomly collected data. Predicting the success, however, is far from trivial because the error functional E^Q used for learning is different from the actual performance metric E^X. For problems without redundancy it is easy to see that learning with E^Q is sound. We have recently proven [19] for linear control problems, even with redundancy, that the gradients of E^Q and E^X satisfy a non-negative relation, so that minimizing E^Q leads to minimizing E^X. For non-linear redundant problems, however, example based learning of inverse models is not generally applicable: such problems possess non-convex sets of solutions q for the same outcome x [18]. Example-based learning can then average them, which leads to invalid estimates. For differential models this can typically be neglected: they are local by construction which implies approximately convex solution sets. For direct inverse models, learning at least from arbitrary data is prohibitive due to this effect.

2 Lessons Learned from Babies

The standard models for the learning of control discussed in the last section demand either an exhaustive exploration or prior knowledge. Exhaustive exploration is prohibitive in high dimensions, which gets even worse when the robot to be controlled is non-stationary. How is it possible to learn and maintain a control skill in a non-stationary domain that can *not even once* be explored exhaustively? This tough challenge is not unique to robots. To the opposite, humans face the same problem to learn control a very high-dimensional, ever-changing motor system right after birth. It turns out that infants follow a pathway very different from computational models that require an exhaustive exploration. It has long been argued in computational learning literature [20, 8] that infants' early exploratory behavior is essentially random, and thus exhaustive. But this claim does not withstand evidence from three decades of developmental research, which has shown conclusive evidence



(a) Initialization. (b) Orientation

(c) Expansion (d) Tuning

Figure 2: The inverse estimate is initialized around zero in joint space. During goal babbling it unfolds successively and reaches an accurate solution [27].

for very coordinated behavior even in newborns. Examples include orienting towards sounds [21], tracking of visual targets [22], and apparent reflexes that have been re-discovered as goal-directed actions [23]. In the case of reaching, it has been shown that newborns attempt goal-directed movements already few days after birth [24]. "Before infants master reaching, they spend hours and hours trying to get the hand to an object in spite of the fact that they will fail, at least to begin with" [25]. From a machine learning point of view, these findings motivate to devise methods that intertwine exploration and learning in a goal-directed manner right away. Findings of early goal-directed actions are complemented by studies investigating the structure of infants' reaching attempts. When infants perform the first *successful* reaching movements around the age of four months, these movements are controlled entirely *feedforward* [26]. This strongly indicates the use of direct inverse model as discussed in the last section, which selects one solution and applies it without corrections. It seems that infants follow a very efficient pathway by focusing on the *achievable*: instead of trying (and failing) to explore everything before starting goal-directed behavior, they gather few but appropriate solutions and *use* them directly. Only later on these movements are gradually optimized and become more adaptive as it is needed. While this pathway is very intuitive, it is orthogonal to the random exploration approach which first attempts to gather full knowledge about the sensorimotor space.

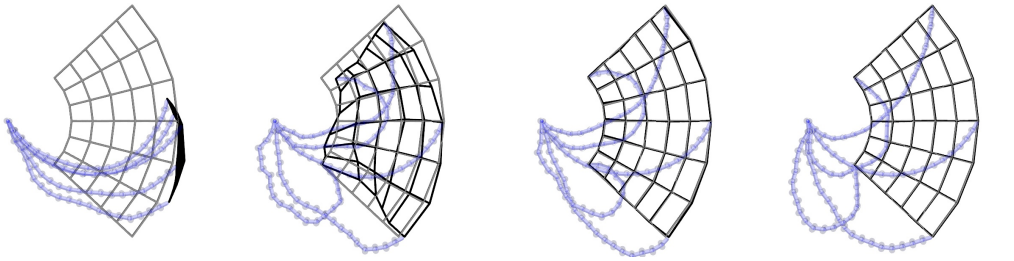
3 Learning Inverse Models with Goal Babbling

The general idea that connects early goal-directed movements and initial feedforward control is to take redundancy as an *opportunity* to reduce the demand for exploration. If there are multiple ways to achieve some behavioral goal, there is no inherent need to know all of them. One can attempt a *partial exploration* of the action space that is just enough to achieve any goal. In order to capitalize on these insights, we have previously introduced the concept of goal babbling:

> **Definition [27]: Goal babbling** is the bootstrapping of a coordination skill by repetitively trying to accomplish multiple goals related to that skill.

Goal babbling aims at the *bootstrapping* of skills. In contrast, goal-directed exploration has been used in several approaches only for the fine-tuning of well initialized models [12, 28], or requiring prior knowledge [17, 18]. Furthermore, goal babbling applies to domains with multiple goals, in contrast to typical scenarios in reinforcement learning, in which only a single desired behavior is considered [29], and also algorithms in control domains which only consider a single goal [30].

Goal babbling does not refer to a particular algorithm, but to a concept that can be implemented in various ways. We investigate the learning of *direct inverse models* by means of goal babbling.



(a) $r = 10$ movements (b) $r = 100$ movements (c) $r = 1000$ movements (d) $r = 10000$ movements

Figure 3: Example of online goal babbling for a 20 DOF planar arm. The inverse estimate unfolds with high speed even in high dimensions. The selected postures get smoother and more comfortable over time, as indicated by the number r of point-to-point movements [31].

This approach resembles infants' developmental pathway, which serves as an example of efficiency, by acquiring at first one valid solution that can be used for feedforward control. For the learning of such models, error-based methods are clearly disqualified by their inherent need for prior knowledge, which leaves the choice to learn by fitting self-generated examples. Thereby motor commands q are chosen by querying the inverse model with successive goals x^* and adding exploratory noise E:

$$q_t \quad = \quad g(x_t^*, \theta_t) + E_t(x_t^*). \tag{3}$$

Thereby goals can be chosen along continuous random paths [31]. However, approaches to choose goals based on predicted learning progress also exist [32]. Learning steps are then performed by observing the result $x_t = f(q_t)$ and fitting the inverse model to (x_t, q_t). For linear problems we have recently proven [19] that this scheme does not only lead to a valid inverse model, but even to an optimal least-squares solution when redundancy is present.

Reward-weighted Regression For non-linear problems the additional problem of non-convex solution sets [18] needs to be considered. Previous studies have only shown how to deal with non-convexity locally, for instance by reformulating the problem into a differential one [12]. However, it turns out that goal babbling provides an elegant solution [27] for this long-standing problem. During goal babbling, it is possible to utilize the *goals as reference structure* in order to resolve inconsistent solutions. When sampling *continuous paths* of goal-directed movements, inconsistent examples can only appear if either (i) the observed movement x_t leads in the opposite direction of x_t^*, or (ii) there is no observed movement x_t at all despite a movement of the motors. This finding motivated to simply exclude such examples by means of a weighting scheme. The first case can be simply be detected by the angle of intended and observed movement:

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

The second case is characterized by a minimum of movement efficiency, which can also be easily detected. The weights for both measures are then multiplied in order to exclude any of the two cases:

$$w_t^{eff} = \frac{||x_t - x_{t-1}||}{||q_t - q_{t-1}||}\ , \qquad w_t = w_t^{dir} \cdot w_t^{eff}\ . \tag{5}$$

Such a weight is assigned to each example, and the *weighted error* $E_w^Q = w_t \cdot (q_t - g(x_t))^2$ is minimized. A very similar scheme of reward weighted regression for efficiency has been used in [28], although coming from a completely different direction. A very positive effect besides resolving the non-convexity issue is that the efficiency weighting causes the inverse model to select very efficient solutions that smoothly relate to each other, because solutions with high efficiency dominate the averaging across exploratory noise. Even in high-dimensions the procedure can therefore find very elegant solutions for resolving the redundancy [27].

Partial Exploration requires Stabilization Non-convexity is not the only problem to deal with. A more general problem is that goal-directed exploration processes tend to *drift* between different

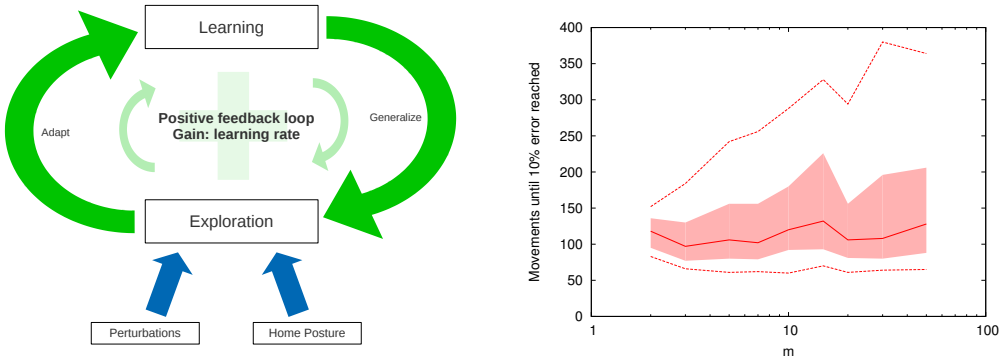
redundancy resolutions, similarly to non-cyclic controllers like pure inverse Jacobian control. This is a problem when attempting an only partial exploration because the regime of known solutions can be left which leads to an entire degeneration of the skill. Performing goal babbling therefore requires a stabilization mechanism to prevent such drifts. A very effective way has been proposed independently in [27] and [33]: Instead of permanently performing goal-directed movements, the learner returns to a "rest" or "home" position after some time, which corresponds to executing some action q^{home}, which is also repeatedly incorporated into the learning. The inverse model will generally tend to reproduce the connection between q^{home} and $x^{home} = f(q^{home})$ if it is used for learning: $g(x^{home}) \approx q^{home}$. This stable point prevents the inverse estimate to drift away. Learning can start around the home posture and proceed to other targets. Similar approaches can be applied to forward-model-based learning and control, e.g. by applying a Nullspace-optimization towards the home posture during goal babbling [34]. Together with the efficiency weighting, this starting and return point also allows for a decent control of the overall redundancy resolution, which can be exploited for learning different solution branches if necessary [35].

Example An example of the overall procedure is shown for a toy-problem in Fig. 2. A 2 DOF robot arm is used to control the height (color-coded from blue to red) of the effector. Goal babbling starts in the home posture, and directly spreads out along the steepest gradient. It rapidly expands and finally reaches a phase of non-linear tuning. It is well visible that training data (green dots) is not generated exhaustively, but *partially* covers the action space (left side of each (a)-(d)) along a low-dimensional manifold, which can succeed even if $\mathbf{Q}$ is very high-dimensional.

4 Online Learning in the Loop

Online learning during goal babbling has turned out to be highly beneficial – and to expose effects very different from online learning on fixed data sets during which online gradient descent is a stochastic approximation of batch gradient descent. An interesting effect can be observed when manipulating the learning rate during the gradient descent on E_w^Q. It turned out that the speed of learning scales in a very non-linear manner with this learning rate: increasing the learning rate by a factor of 10 can lead to an effective speedup of a factor 20 or 50 [31]. During goal babbling, exploration and learning inform *each other*, instead of only learning being informed by exploration. Improving with a higher learning rate then results in a more informative example in the next exploration step, which in turn accelerates learning. This behavior can be understood by means of a *positive feedback loop* (see Fig. 4a), in which the learning rate acts as a gain. In fact, the underlying learning dynamics of goal babbling have been shown to resemble those of *explosive combustions* [19] which also comprise a very non-linear phase of high-speed expansion. As a result of this feedback loop, and the only partial exploration described by goal babbling, exploration becomes both very scalable to high dimensions and very fast in terms of absolute time needed to succeed. An example is shown for a planar arm with 20 degrees of freedom in Fig. 3. After only 100 continuous point-to-point movements, goal babbling can already roughly reach throughout the entire workspace. Later-on refinement yields a very smooth redundancy resolution, and an accuracy in a $1mm$ range. Fig. 4b shows a systematic comparison across different dimensions for this setup. The cost for the bootstrapping of a solution is measured in terms of the movements necessary to reduce the performance error to 10% of its initial value. The median cost is approximately constant in the entire range from two to 50 DOF. Plus, it is very low around only 100 movements, which is a speed thoroughly competitive to human learning [36]. Recent work [37] shows that this procedure also allows for a very efficient *identification* of the reachable workspace. Sampling goals along random directions allows to fully cover a workspace with unknown shape and size without any representation of it, similarly to vacuum cleaning robots that move along random directions and get repulsed by obstacles. While the learning of inverse models themselves can not be evaluated against a random exploration baseline (since inverse models *cannot* be learned from random data), this setup allows to compare the workspace's coverage of random compared to goal-directed exploration. Results show that goal babbling permits a good coverage after 10^6 examples along continuous paths in a challenging 50 DOF domain in which a pure random strategy would require at least 10^{14} examples.

A practically highly relevant use case for these algorithms is to master the control of the *Bionic Handling Assistant* (BHA, see Fig. 1a). This robot comprises nine main actuators, which are bellows inflated by a pneumatic actuation. Pneumatics alone are not sufficient for successful control, since



(a) Positive feedback loop of exploration and learning.

(b) Scalability and speed

Figure 4: Exploration and learning mutually inform each other in goal babbling. This constitutes a positive feedback-loop during bootstrapping which substantially accelerates learning. Goal babbling scales to very high-dimensional problems, as shown by the only marginal increase of exploratory cost for reaching with between $m=2$ and $m=50$ degrees of freedom [31].

friction and the visco-elasticity of the bellows' polyamide material can cause different postures when applying the same pressure to the actuators. In our setup the robot is entirely controlled by means of the actuator-lengths, which can be measured by cable-potentiometers, and be controlled by adjusting the pressure with a learned equilibrium-point controller [38]. The central difficulties for this robot are its dimensionality, its very slow steady-state dynamics (it takes up to 20 seconds to fully converge into a posture), and the very narrow and ever-changing ranges of the single actuators. Each actuator's movement is very restricted, which requires a fully cooperative movement of *all* actuators to move through the workspace. Since the bellows' material is visco-elastic, the maximum and minimum length it can take changes over time. This implies that previously possible motor commands can become impossible since the length is out of range. Hence, learning needs to continuously re-discover solutions how to reach for goals. These challenges can be mastered very efficiently with online goal babbling [39]. Left/right and forward/backward movements can be accurately controlled after less than 30 minutes of exploration. In contrast to revolute joint robots the BHA can also *stretch* by inflating all of its actuators in a top/down direction. Such movements are more difficult to discover since they require a highly cooperative behavior of all actuators that is difficult to coordinate within the narrow and changing limits. Nevertheless results show that a full 3D control of the BHA's effector is possible within few hours of real-world exploration. Just after learning, the control achieves accuracies around $2cm$, which is already reasonable considering the robot's widely opened, elastic gripper that can simply surround objects to grasp them. The feedforward control thereby allows to perform very quick movements despite the robot's long response time. If needed, however, the accuracy can be further improved to 6-8mm by applying an additional cartesian feedback control on top [39], which comes with the cost that movements need to be performed very slowly in order not to get unstable. The accuracy is thereby close to the robot's repetition accuracy of $5mm$, which serves as an absolute baseline of how accurately the robot can be controlled at all.

5 Discussion

High-dimensional motor systems can not be fully explored even once, not to mention a re-exploration necessary for non-stationary systems. Infant developmental studies show ways to deal with such challenges, by starting right away with non-exhaustive, *goal-directed* exploration, and learning *simple*, e.g. feedforward, skills in the beginning. This strategy allows to perform an effective and efficient *partial exploration* just to the extent that goals can be achieved. Goal babbling allows to mimic this tremendous efficiency. This strategy is highly beneficial for technological problems with the scale of the Bionic Handling Assistant. We have introduced a series of algorithms for the learning of direct inverse models by means of goal babbling and reward weighted regression. These algorithms allow for a learning of very elegant solutions even in high dimensions, and with human-competitive learning speed. Other implementations of goal babbling have recently been pro-

posed, and confirm the success of goal babbling, as well as its superiority over random exploration in terms of bootstrapping efficiency [32, 34, 40, 41]. These results also demonstrate the general validity of the goal babbling concept. Goal-directed exploration itself is not a new idea. But considering it as a first-class concept not only for a fine-tuning of skills has revealed interesting phenomena like the existence of a positive feedback loop. Likewise, performing only partial exploration only goes *together* with admitting that a full mastery of all possible solutions and full-scale feedback control in the entire action space is not possible for large scale problems. This requires a stabilization that keeps exploration and control in known regions of the action space, but with the benefit that even large problems can be solved in a very pragmatic manner.

Acknowledgments This study has been supported by the JSPS Grant-in-Aid for Specially promoted Research (No. 24000012).

References

[1] C. Laschi, B. Mazzolai, V. Mattoli, M. Cianchetti, and P. Dario. Design of a biomimetic robotic octopus arm. *Bioinspiration & Biomimetics*, 4(1), 2009.

[2] Kenneth J Korane. Robot imitates nature. *Machine Design*, 82(18):68–70, 2010.

[3] Koh Hosoda, Shunsuke Sekimoto, Yoichi Nishigori, Shinya Takamuku, and Shuhei Ikemoto. Anthropomorphic muscular-skeletal robotic upper limb for understanding embodied intelligence. *Advanced Robotics*, 26(7):729–744, 2012.

[4] A. Liegeois. Automatic supervisory control of configuration and behavior of multibody mechanisms. *IEEE Transactions on Systems, Man and Cybernetics*, 7(12):861–871, 1977.

[5] Ken Waldron and Jim Schmiedeler. Chapter 1: Kinematics. In Bruno Siciliano and Oussama Khatib, editors, *Handbook of Robotics*, pages 9–33. Springer New York, 2007.

[6] G. Sun and B. Scassellati. Reaching through learned forward models. In *IEEE-RAS/RSJ International Conference on Humanoid Robots (Humanoids)*, 2004.

[7] Camille Salaün, Vincent Padois, and Olivier Sigaud. Learning forward models for the operational space control of redundant robots. In Olivier Sigaud and Jan Peters, editors, *From Motor Learning to Interaction Learning in Robots*, pages 169–192. Springer, 2010.

[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*, 5(4):408–435, 1993.

[9] S. Ulbrich, V. Ruiz de Angulo, T. Asfour, C. Torras, and R. Dillmann. Kinematic bzier maps. *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, 42(4):1215–1230, 2012.

[10] R. Martinez-Cantin, M. Lopes, and L. Montesano. Body schema acquisition through active learning. In *IEEE Int. Conf. Robotics and Automation (ICRA)*, 2010.

[11] D. A. Cohn, Z. Ghahramani, and M. I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4(1):129–145, 1996.

[12] Aaron D'Souza, Sethu Vijayakumar, and Stefan Schaal. Learning inverse kinematics. In *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2001.

[13] Jan Peters and Stefan Schaal. Learning to control in operational space. *The International Journal of Robotics Research*, 27(2):197–212, 2008.

[14] Duy Nguyen-Tuong, M. Seeger, and J. Peters. Computed torque control with nonparametric regression models. In *American Control Conference*, 2008.

[15] Michael I. Jordan. Computational aspects of motor control and motor learning. In *Handbook of Perception and Action: Motor Skills*. Academic Press, 1996.

[16] Matthias Rolf, Jochen J. Steil, and Michael Gienger. Efficient exploration and learning of whole body kinematics. In *IEEE Int. Conf. Development and Learning (ICDL)*, 2009.

[17] M. Kawato. Feedback-error-learning neural network for supervised motor learning. In R. Eckmiller, editor, *Advanced Neural Computers*, pages 365–372. Elsevier, 1990.

[18] M. I. Jordan and D. E. Rumelhart. Forward models: supervised learning with a distal teacher. *Cognitive Science*, 16(3):307–354, 1992.

[19] Matthias Rolf and Jochen J. Steil. Explorative learning of inverse models: a theoretical perspective. *Neurocomputing*, 2013. In Press.

[20] M. Kuperstein. Neural model of adaptive hand-eye coordination for single postures. *Science*, 239(4845):1308–1311, 1988.

[21] Rachel K. Clifton, B. A. Morrongiello, W. Kulig, and J.M. Dowd. Developmental changes in auditory localization in infancy. In *Development of Perception, Vol. 1, Psychobiological Perspectives*, pages 141–160. New York: Academic Press, 1981.

[22] H. Bloch and I. Carchon. On the onset of eye-head coordination in infants. *Behavioral Brain Research*, 49(1):85–90, 1992.

[23] AL van der Meer, FR van der Weel, and DN Lee. The functional significance of arm movements in neonates. *Science*, 267(5198):693–695, 1995.

[24] Claes von Hofsten. Eye-hand coordination in the newborn. *Developmental Psychology*, 18(3):450–461, 1982.

[25] Claes von Hofsten. An action perspective on motor development. *Trends in Cognitive Science*, 8(6):266–272, 2004.

[26] Rachel K. Clifton, Darwin W. Muir, Daniel H. Ashmead, and Marsha G. Clarkson. Is visually guided reaching in early infancy a myth? *Child Development*, 64(4):1099–1110, 1993.

[27] Matthias Rolf, Jochen J. Steil, and Michael Gienger. Goal babbling permits direct learning of inverse kinematics. *IEEE Trans. Autonomous Mental Development*, 2(3), 2010.

[28] Jan Peters and Stefan Schaal. Reinforcement learning by reward-weighted regression for operational space control. In *Int. Conf. Machine Learning (ICML)*, 2007.

[29] E. Theodorou, J. Buchli, and S. Schaal. Reinforcement learning of motor skills in high dimensions: A path integral approach. In *IEEE Int. Conf. Robotics and Automation*, 2010.

[30] Stefan Schaal and Christopher G. Atkeson. Assessing the quality of learned local models. In *Advances in Neural Information Processing Systems (NIPS)*, 1994.

[31] Matthias Rolf, Jochen J. Steil, and Michael Gienger. Online goal babbling for rapid bootstrapping of inverse models in high dimensions. In *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2011.

[32] A. Baranes and P-Y. Oudeyer. Active learning of inverse models with intrinsically motivated goal exploration in robots. *Robotics and Autonomous Systems*, 61(1):49–73, 2013.

[33] Adrien Baranes and Pierre-Yves Oudeyer. Maturationally-constrained competence-based intrinsically motivated learning. In *IEEE Int. Conf. Development and Learning (ICDL)*, 2010.

[34] Lorenzo Jamone, Lorenzo Natale, Kenji Hashimoto, Giulio Sandini, and Atsuo Takanishi. Learning task space control through goal directed exploration. In *IEEE Int. Conf. Robotics and Biomimetics (ROBIO)*, 2011.

[35] Rene Felix Reinhart and Matthias Rolf. Learning versatile sensorimotor coordination with goal babbling and neural associative dynamics. In *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2013.

[36] Uta Sailer, J. Randall Flanagan, and Roland S. Johansson. Eye–hand coordination during learning of a novel visuomotor task. *Journal of Neuroscience*, 25(39):8833–8842, 2005.

[37] Matthias Rolf. Goal babbling with unknown ranges: A direction-sampling approach. In *IEEE Int. Joint Conf. Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2013.

[38] K. Neumann, M. Rolf, and J. J. Steil. Reliable integration of continuous constraints into extreme learning machines. *Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 21(supp02):35–50, 2013.

[39] Matthias Rolf and Jochen J. Steil. Efficient exploratory learning of inverse kinematics on a bionic elephant trunk. *IEEE Trans. Neural Networks and Learning Systems*, 2013. In Press.

[40] Patrick O. Stalph and Martin V. Butz. Learning local linear jacobians for flexible and adaptive robot arm control. *Genetic Programming and Evolvable Machines*, 13(2):137–157, 2012.

[41] Christoph Hartmann, Joschka Boedecker, Oliver Obst, Shuhei Ikemoto, and Minoru Asada. Real-time inverse dynamics learning for musculoskeletal robots based on echo state gaussian process regression. In *RSS*, 2012.