Robot Motor Skill Acquisition with Learning in Two Spaces

Jian Fu(✉), Ce Cao, Jinyu Du, and Siyuan Shen

School of Automation, Wuhan University of Technology, Wuhan 430070, China
fujian@whut.edu.cn
http://www.escience.cn/people/fujiane/index.html

Abstract. Motor skill acquisition and refinement is critical for the robot to step in human daily lives, which can endow it with the ability of autonomously performing unfamiliar tasks. However, how does the robot autonomously fulfill the new motion task with preassigned performance based on the demonstration task is still a challenge. We in this paper proposed a novel motor skill acquisition policy to conquer above problem, which is based on improved local weighted regression (iLWR), policy improvement with path integral (PI^2). Besides, the mixture Gaussian regression (GMR) guided self-reconstruction of basis function and the search of weight coefficient in the policy expression are performed alternately in basis function space and weight space to seek the optimal/suboptimal solution. In this way, robot can achieve the gradual acquisition of movement skills from similar tasks which is related to the demonstration to unsimilar task with different criterion. At last, the classical via-points trajectory planning experiment are performed with SCARA manipulator, NAO humanoid robot to verify that the proposed method is effective and feasible.

Keywords: Alternate study in two spaces · GMR-PI^2 · Motor skill acquisition

1 Introduction

It had long been a dream for researchers in the robot communities to endow the robot with motor skill similar to the man. Recently, robot learning from demonstration (LfD) together with reinforcement learning (RL) (Argall et al. 2009; Peters and Schaal 2008; Rombokas et al. 2013; Deisenroth et al. 2013) has attracted significantly increased attention. By means of LfDRL, researchers can derive a robot controller autonomously merely by back-driving or teleoperating it. Furthermore, the controller could be self-improved to refine and expand robot motor capability obtained from demonstration to fulfill the task dissimilar to demonstration task.

J. Fu—The author acknowledges the National Natural Science Foundation of China (61773299, 515754112).

H. Yu et al. (Eds.): ICIRA 2019, LNAI 11745, pp. 379–389, 2019.
https://doi.org/10.1007/978-3-030-27529-7_33

Usually a parametric policy representation is selected firstly for the LfDRL. Generally, a dynamic model with flexible adjustment sounds good, for the reason that it is easy to modulate online and exhibit robustness. DS (Khansari-Zadeh and Billard 2010; 2011) and DMPs (Ude and Gams 2010; Ijspeert et al. 2013) models are current popular dynamic model. DS represents motion scheduling in the form of a nonlinear autonomous dynamic system, which is time-invariance and global/local asymptotic stable. Whereas, DMPs models the movement planning as superimposition of a linear dynamic system and a nonlinear term. The former takes precedence over the latter to guarantee the global convergence at the end of the motion. Then a flexible method to adjust the model is required, including radial basis function networks, regularized kernel least-square, locally weighted regression (Atkeson et al. 1997) and Gaussian process etc.

It is often easier to learn a policy than model a robot with its environment, hence, model-free policy search methods are more popular than model-based policy search methods. Some classic model-free policy search methods were proposed recently, such as Relative Entropy Policy Search (REPS), Covariance Matrix Adaptation-Evolutionary Strategy (CMA-ES). REPS (Parisi et al. 2015; Peters et al. 2010) formulates the policy search problem as an optimization problem in an information theoretic way, meanwhile updates its policy by weighted maximum likelihood estimates. However, CMA-ES is a black-box optimizer. And it uses heuristics to estimate the weight and update the distribution, which often work well in practice but are not founded on a theoretical basis (Gregory et al. 2015).

Aimed to develop a generic method for motor skill learning with good quality for given motion planning task, this paper proposed a DMPs-iLWR policy which learning in two space to search target solution and this paper is organized as follows. Section 2 depicted DMPs-iLWR for policy representation and imitation learning. Section 3 investigates the policy optimization based on GMR-PI^2 from a viewpoint stochastic optimal control. Section 4 describes reconstruction of basis function with auto-encoding and deep iteration process. In Sect. 5, we present in detail the classical benchmark experiment trajectory planning via prior unknown point(s).

2 DMPS-iLWR Based Robot Motor Skill Learning

Classical DMPs model the robot movement in each degree of freedom (DOF) as independent transformation system, which is synchronized in time dimension by a share phase variable. Specifically, it presents a parametric policy in the representation phase, which comprises the transfer system, canonical system and function approximation. Those systems are described as following

$$\begin{cases} \tau\ddot{x}_t = \underbrace{\alpha_x\left(\beta_x\left(g - x_t\right) - \dot{x}_t\right)}_{spring-damping.system} + \underbrace{\Psi_\theta\left(s_t\right) s_t\left(g - x_0\right)}_{forcing.term} & transf.system \\ \tau\dot{s}_t = -\alpha_s s_t & canon.system \\ \Psi_\theta\left(s_t\right) = \frac{\sum_{i=1}^{K}\psi_i w_i}{\sum_{i=1}^{K}\psi_i} & func.approx \end{cases} \tag{1}$$

where τ is the scaling factor for the duration of motion. $x_t = q_{ref}$ is the reference trajectory generated by transformation system for one DOF, s_t is the phase of the movement generated by canonical system, which decays from 1 to 0 over the same duration with transformation system. ψ_i is the Gaussian kernel function with the variance spaced equally across motion duration.

w_i is the weight associated. The goal g is a point attractor and x_0 is the start state. α_x, β_x, α_s are positive constants, by which the spring damping system is modeled as a 2 order critical damping system. θ is the hyper-parameter for basis functions. As we can see, DMPs for each transformation system is a single-input (time) and single-output (joint) (SISO) system. It constructs a time-dependent control reference rather than traditional state dependent one, which dramatically simplifies the learning process.

We in this paper propose an DMPs-iLWR policy representation in which we employ improved Local Weighted Regression (iLWR) to avoid the poor performance of traditional LWR in the imitation learning and employ the data-driven Gaussian Mixture Model to adjust the feature of trajectories adaptively during the PI^2 learning. It is shown in Eq. (2), which comprises a transformation system, a canonical system, a gating system and a weighted basis function.

$$\begin{cases} \tau\ddot{x}_t = \underbrace{\alpha_x\left(\beta_x\left(g - x_t\right) - \dot{x}_t\right)}_{a_s} + \underbrace{h_t\bar{\mathbf{B}}_{s_t}\bar{\mathbf{w}}}_{a_f} & \text{transf.system} \\ \tau\dot{s}_t = \begin{cases} 1/T & \text{if } t \le T \\ 0 & \text{otherwise} \end{cases} & \text{canon.system} \\ h_t = \frac{1}{1+e^{\alpha_h(t-\tau T)}} & \text{gatin.system} \\ \bar{\mathbf{w}} = \left[\bar{w}_1 \cdots \bar{w}_K \ \bar{w}_{K+1} \cdots \bar{w}_{2K}\right]^T & \text{weight} \\ \bar{\mathbf{B}}_{s_t} = \left[\gamma_1 s_t \ \cdots \gamma_K s_t \ \gamma_1 \cdots \gamma_K\right] & \text{basis.function} \end{cases} \tag{2}$$

where $\bar{\mathbf{B}}$ is the equivalent basis function, and the form of real basis shows as following

$$\gamma^{(i)} = \frac{exp\left(-\frac{1}{2\sigma_i^2}(s_t - c_i)^2\right)}{\sum_{j=1}^{K} exp\left(-\frac{1}{2\sigma_j^2}\left(s_t - c_j\right)^2\right)} \tag{3}$$

The revised canonical system can guarantee that the phase is proportional to time in the transient process, and the gating system is used to guarantee the convergence of forcing term, and the transformation system is the important controllable part, and we adopt the form dot product between basis function $\bar{\mathbf{B}}_{s_t}$ and weight $\bar{\mathbf{w}}$ to approximate forcing term a_s, and the trajectories can be adjusted by controlling the weight $\bar{\mathbf{w}}$.

The traditional LWR presents an restrictive effect of imitation learning. Because the default fitting trajectories must cross the origin of coordinates, which means if the phase variable is close to zero, so must be the forcing term. Therefore we revise the controlling term from pattern ($y = Ax$) to pattern ($y = Ax + B$) to avoid this restriction. And the new forcing term can be seen as Eq. (4).

$$f(s_t) = \frac{\sum_{i=1}^{K} \psi_i(s_t) \begin{bmatrix} A_i & B_i \end{bmatrix}}{\sum_{i=1}^{K} \psi_i(s_t)} \begin{bmatrix} s_t \\ 1 \end{bmatrix} (g - y_0) \tag{4}$$
$$= \sum_{i=1}^{K} \gamma^{(i)} (A_i s_t + B_i)(g - y_0)$$

Associated with the optimization policy of PI^2, we can learn this two parameters (A and B) simultaneously, which can be seen as the slope and the interception of the linear function respectively. The number of real basis are set to K, naturally the equivalent one are twice and the forms of equivalent basis and weight show as following

$$\bar{\mathbf{B}}_{s_t}^{(m)} = \begin{cases} \gamma^{(i)} s_t & m = i, m \leq K \\ \gamma^{(i)} & m = i + K, K < m \leq 2K \end{cases}$$

$$\bar{\mathbf{w}}^{(m)} = \begin{cases} A_i & m = i, m \leq K \\ B_i & m = i + K, K < m \leq 2K \end{cases}$$

Next, LfD can conduct DMPs-iLWR to learn a feasible solution or more in joint space.

3 Policy Improvement Based on iLWR-PI²

Although DMPs-iLWR can effectively replicate and generalize robot demonstration movement, it maybe not a optimal/suboptimal policy for the task. Furthermore, it can not autonomously fulfill the motion different from demonstration one with a high-quality level, such as the task with additional criterion, though vanilla DMPs-iLWR by which $(g - y_0)$ can be adapt to the goal change and scaling law is an orientation-preserving homeomorphism between the original equations using $g - y_0$ and the scaled differential equation using $k\,(g - y_0)$ (Ijspeert et al. 2013). So we combine DMPs-iLWR (policy representation) with path integration (policy improvement) (Theodorou et al. 2010) through stochastic optimal control to meet the requirement. Specifically, we apply Feynman-Kac theorem to derive the state value function based on path integral, and then deduce the optimal control policy. In this way, we solve the Hamilton-Jacobian-Bellman equation indirectly.

As a reinforcement learning, the cost function shows as the Eq. (5), and we add the constraint associated with tasks.

$$S(\tau_i) = \phi_{t_N} + \sum_{j=i}^{N-1} q_{t_j} dt + \frac{1}{2} \sum_{j=i}^{N-1} (\mathbf{w} + \varepsilon)^T \frac{\mathbf{B}_{s_{t_j}} \mathbf{B}_{s_{t_j}}^T}{\mathbf{B}_{s_{t_j}}^T R^{-1} \mathbf{B}_{s_{t_j}}} (\mathbf{w} + \varepsilon_{t_j}) + \frac{\lambda}{2} \sum_{j=i}^{N-1} \ln |\boldsymbol{H}| \tag{5}$$

where $\frac{\lambda}{2}\sum_{j=i}^{N-1}\ln|\boldsymbol{H}|$ is usually removed given that basis function is fixed. The optimal time-variant policy with value function V_{t_i} shows as

$$\begin{aligned}\mathbf{w}_{t_i} &= -R^{-1}\mathbf{B}_{s_t}^T\left(\nabla_{\boldsymbol{z}_{t_i}}V_{t_i}\right)\\ &= \int P(\tau_i)\,u(\tau_i)\,d\tau_i,\end{aligned} \tag{6}$$

where

$$\begin{aligned}P(\tau_i) &= \frac{e^{-\frac{1}{\lambda}S\left(\tau_{t_i}\right)}}{\int e^{-\frac{1}{\lambda}S\left(\tau_{t_i}\right)}d\tau_i}\\ u(\tau_i) &= \frac{R^{-1}\mathbf{B}_{t_i}\mathbf{B}_{t_i}^T}{\mathbf{B}_{t_i}^T R^{-1}\mathbf{B}_{t_i}}\left(\mathbf{w}+\varepsilon_{t_i}\right)\end{aligned} \tag{7}$$

Specifically, $P(\tau_i)$ is the path depended probability distribution and $u(\tau_i)$ is local optimal control derived by value function.

In practical engineering, we usually carry out K roll-outs. And for specific time index i, we gain $P(\tau_{i,k})$ similar to softmax function (taking the frequency as probability), where k is the index of K roll-outs.

In this way, we achieve weighted average of adjustment $\delta\mathbf{w}_{t_i}=\mathbf{w}_{t_i}-\mathbf{w}$ across N time index as equivalent time-invariant policy, which could be expressed as

$$\delta\mathbf{w}=\frac{\sum_{i=0}^{N-1}(N-i)\,\mathbf{B}_{t_i}\delta\mathbf{w}_{t_i}}{\sum_{i=0}^{N-1}(N-i)\,\mathbf{B}_{t_i}} \tag{8}$$

4 Basis Function Auto-Encoding and Alternate Learning in Two Space

Embodiment feature indicates the intelligence what an agent is capable of (from low-level sensory-motor activities to high-level cognitive activities) is closely related to the morphology what the agent is composed of and the way which agent interact with the environment (Pfeifer and Bongard 2006). Therefore, basis function in the policy representation plays an decisive role in determining the feasible coverage of the robot's kinematic intelligent capability. When the robot faces new task which is different from demonstration (for example with additional performance criterion), the new task is less correlative to the experience obtained from demonstration task. In other words, there exists the relative large gap between the experimental basis function from demonstration task and appropriate one for new task. As for new task the minimal return in finite horizon from experimental basis function is ordinarily large than that with appropriate one. We specially draw the Fig. 1 to facilitate the comprehending. Basis function B1 is obtained from demonstration task and blue elliptical region is the projection from space spanned by basis function B1 to feature/measure space. And the location of optimal approximation which associate the optimal weight with basis B1 indicates the distance L3 is relative large. We assume basis function B3 is appropriate basis for the new task and red elliptical region is associated projection. As seen, there are no overlap between two projection regions. That

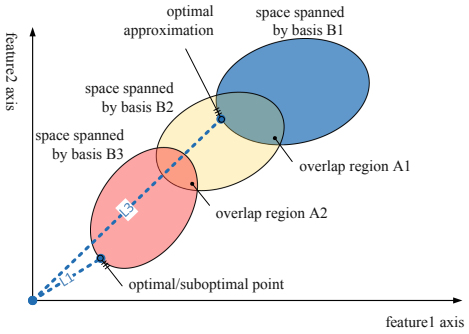


Fig. 1. Diagrammatic sketch of idea for RBAE (Color figure online)

means no correlation between two tasks which is one of the difficulties for the robot skill learning and autonomous skill acquisition.

Since it is unrealistic for the robot to preset the appropriate basis function such as B3 prior to the new task arising in the unstructured scenarios, it is necessary to endow the robot with the capability of gradual learning and skill acquisition. In other words, we can design a sequence of projection zone overlapped each other to join the demonstration task and new task, as yellow elliptical region works shown in Fig. 1. By means of the correlation (overlap) between anteroposterior projection zone in which robot evolving along with, for example form blue zone via yellow zone to red zone shown in Fig. 1, autonomous specific skill improvement could be achieved.

We put forward the alternate learning in two spaces based on above analysis to conduct motion skill acquisition from demonstration to new task. Specifically, reinforcement learning in weight space will gradually drive the candidate elites from blue zone B1 into the overlap region A1 by means of the distance (reward) in feature space. By means of the presentation learning on the data generated from candidate elites, the algorithms can automatically generate the new basis and zone B3 in which the better performance capability is available. In this way, robot can eventually approach the optimal/suboptimal point for the specific new skill by repeating the procedure. So we propose the improved LWR and PI^2 learning in two space to overcome the above challenge. Its main flow is ① →②→③→④→⑤ →⑥→④→⑤→⑥$\cdots$ shown in Fig. 2. Policy representation (iLWR) for motor skill is firstly constructed based on the principle of maximum entropy. In other words, alike Gaussian functions ψ_i with identical variances are evenly assigned along the phase duration, which can provide the policy the most flexible for learning unknown motion given that the number of basis function is fixed. (①). Then LfD is conducted to seek the appropriate weight to replicate the demonstration(②). Next, for new task (with different performance criterion), we apply RL(iLWR-PI^2) to search the suitable weight until the cost doesn't decrease apparently any more (③④). Usually, the best trajectories so far imply the feature of new task. So K-means++ is applied to cluster

classification on the data generated by those candidate elites. Then EM-GMM is adopted to estimates the appropriate parameters μ_k and Σ_k for respective Gaussian distribution component. Next, these parameters will be assign to ψ_i and $r^{(m)}$ to construct the new basis function $\bar{\mathbf{B}}_{s_t}$. In a sense, more appropriate basis functions are constructed adaptively based on data-driven according to the character of targeted task. Also, LfD is performed to seek weights to replicate the best trajectories so far with a posterior maximization.(⑤). Based on them, we again apply DL(iLWR-PI^2) to search the best approximation in the new space(⑥). This procedure repeat again and again (④→⑤→⑥) until a satisfied trajectory is obtained.

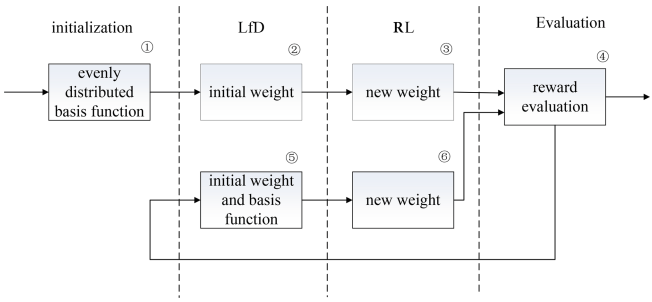


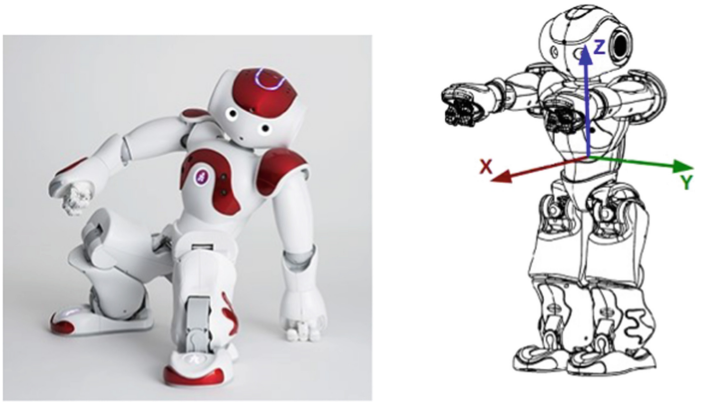
Fig. 2. Algorithm flow for RBAE

5 Simulation Experiment

In this part, we employ NAO robot shown in Fig. 3(a) to verify the effectiveness of the proposed method. This robot is the first humanoid robot of the Aldebaran Robotics company. And it is an interactive companion robot. There are twenty-five joints on the NAO robot. But this experiment only involves five revolute joints with the right arm of NAO: RShulderRoll, RShouderPitch, RElbowRoll, RElbowYaw and RWristYaw.

In this experiment, we fix RShouderPitch and RWristYaw at 0 degree. Experiences are depicted as following. Firstly, man drag the right hand of NAO from starting point to the end point. And at the same time we read the data of three joint angles every 0.05 s. There are 100 sets of data used as the original trajectory. Then we randomly put a small landmark in the domain, which the arm can reach (excluding the start point and the end point). So the arm is supposed to move passing through this via-point (for example strike) from previous start point to end point with the same duration. In addition, the cost which is described in Eq. (9) are met.

In the experiment, we set the start point with (112.62, 62.56, −15.53) in the operation space, which is corresponding to (0.0705, −0.8054, 1.1137) in the joint space. And the coordinates of terminal point is (118.03, 30.59, 214.26)



(a) the experiment platform (b) the position of base frame

Fig. 3. The position of base frame and joints

corresponding to (−0.15, 0.70, 1.46) in the joint space. We choose randomly a point (157.46, 122.69, 78.30) as via point. In the joint space, this via point is with corresponding several groups of joint angle. We choose irregularly and randomly the data (0.03, 0.15, 1.04) in multi group data as the value of via point in the joint space. A clear contrast between our proposed iLWR-PI^2 and classical LWR-PI^2 will be illustrated later.

We now proceeded to test the performance of iLWR-PI^2 and LWR-PI^2. Detail results are listed in Table 1. The cost function is in the form of

$$S = 0.5\sum_{j=1}^{3}\sum_{i=1}^{N-1}\left[10^3\left(\ddot{x}^{(j)}(i)\right)^2 + \left(a_f^{(j)}(i)\right)^2\right] + \sum_{j=1}^{3}10^{10}\left[\left(x^{(j)}(m) - x_v^{(j)}\right)^2\right] + \sum_{j=1}^{3}10^3\left[\left(\dot{x}^{(j)}(N)\right)^2 + \left(x^{(j)}(N) - x_g^{(j)}\right)^2\right] \tag{9}$$

Where i indicates the time index from 1 to N, j indicates the joint index from 1 to 3. Besides, $x_g^{(j)}$ denotes the expected position of point j when the task ends. When the time index equals m, $x^{(j)}(m)$ is the position of joint j which is corresponding to the expected via-point $x_v^{(j)}$ of the joint j. Apparently, $\ddot{x}^{(j)}$ is an arbitary state-dependent cost value, $a_f^{(j)}(i)$ is the acceleration (forcing term) relevant to joint j at the time index i. $\dot{x}^{(j)}(N)$ is the velocity of joint when time index is N.

According to the result, the proposed iLWR-PI^2 methods execute better than original LWR-PI^2. As for Table 1, we run fifteen times under the same condition. Shown in Table 1 the cost generated by iLWR-PI^2 are remarkably smaller than that of LWR-PI^2, which the overall weighted evaluation include energy consuming, via point and terminal status etc. It indicates that iLWR-PI^2 outperform LWR-PI^2.

Table 1. The final costs when algorithm stop

Times	$Cost_1$ (LWR-PI^2)	$Cost_2$ (iLWR-PI^2)
1st	1.39E+08	2.24E+07
2nd	1.58E+08	2.21E+07
3rd	1.32E+08	2.88E+07
4th	2.00E+08	2.25E+07
5th	2.06E+08	1.74E+07
6th	1.68E+08	1.96E+07
7th	2.03E+08	1.87E+07
8th	2.40E+08	2.03E+07
9th	1.24E+08	2.68E+07
10th	1.79E+08	1.72E+07
Mean	1.75E+08	2.16E+07

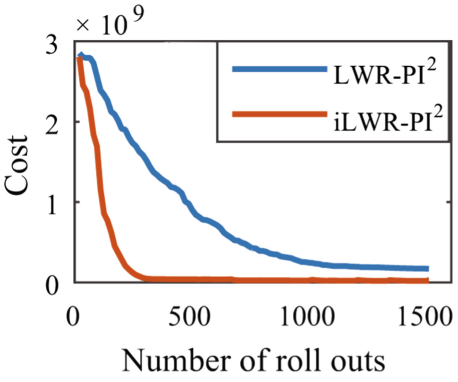
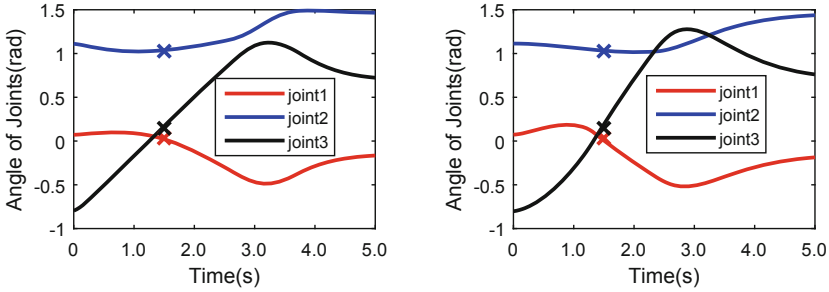


Fig. 4. The red curve shows the cost caused by iLWR-PI^2, The blue curve shows the cost caused by LWR-PI^2(Color figure online)

(a) the origin track (b) the track of iLWR-PI^2 (c) the track of LWR-PI^2

Fig. 5. The track on the platform



(a) the curve of iLWR-PI^2 through via-point

(b) the curve of LWR-PI^2 through via-point

Fig. 6. The curve of iLWR-PI^2 and LWR-PI^2 through via-point in the joint space

Besides, we compare the learning rate between two methods. As shown in Fig. 4, quicker coverage speed of iLWR-PI^2 is manifest.

And Fig. 6 shows the results of iLWR-PI^2 and LWR-PI^2 through via-point in the joint space. They both have excellent results when they pass the via-point. But clearly, the curve of iLWR-PI^2 shown in the Fig. 6(a) is more smooth than LWR-PI^2. The origin track is shown in Fig. 5(a). The actual track of iLWR-PI^2 and LWR-PI^2 on the robot is displayed in Fig. 5(b) and 5(c). Compared that with Fig. 5(b), the track of iLWR-PI^2 and LWR-PI^2 can go through the via-point. So they all have certain learning abilities. Besides, we can get that the trajectories of iLWR-PI^2 shown in the left figure of Fig. 5 is smoother than LWR-PI^2. And the impact to mark of trajectory generated by iLWR-PI^2 is stronger, which indicates that the precision of iLWR-PI^2 is higher.

6 Conclusion

Recently motor skill acquisition has been received strong attention to, and it is also the highlight for robot learning. However associative dilemma, broad learning and targeted improvement for robot, has still been a challenge. In this paper, we propose a novel GMR-PI^2 motor skill learning based on RBAE to overcome the dilemma throughout the all phases of LfDRL. GMR-PI^2 comprise two parts: DMPs-GMR for LfD, GMR-PI^2 for RL. Besides, affiliated RBAE can extract features discovered so far and perform target-oriented exploration with the basis function generated from previous RBAE. After this process iteratively to a certain depth, robot can obtain the capability to fulfilling unfamiliar task with an optimal/suboptimal criterion.

The most important and prominent part of our work is that we propose general RBAE framework and associated algorithms. Specially, we applied the auto-encoding method into GMR-PI^2, and this promotes the optimization more accurately. There are a few interesting future research directions along this topic. Firstly, how to seek the optimal hyper-parameters such as the noise and number of updating the parameters for PI^2 need to study, since it is somewhat

time-consuming to set them suitably. Secondly, we only make uses of the latest information (best trajectories so far) in the algorithms, it is intuitive to allocate and integrate the old/new information which may make the robot more flexible. Thirdly, there are many set of basis functions generated by RBAE which are the features in different proficiency scale. How to utilized them in parallel is a promising research direction.

References

Argall, B.D., Chernova, S., Veloso, M., Browning, B.: A survey of robot learning from demonstration. Robot. Auton. Syst. **57**(5), 469–483 (2009)

Atkeson, C.G., Moore, A.W., Schaal, S.: Locally weighted learning. Artif. Intell. Rev. **11**(1), 11–73 (1997)

Deisenroth, M., Neumann, G., Peters, J.: A survey on policy search for robotics. J. Intell. Rob. Syst. **15**(1), 1–2 (2013)

Gregory, M.D., Martin, S.V., Werner, D.H.: Improved electromagnetics optimization: the covariance matrix adaptation evolutionary strategy. IEEE Antennas Propag. Mag. **57**(3), 48–59 (2015). https://doi.org/10.1109/MAP.2015.2437277

Ijspeert, A.J., Nakanishi, J., Hoffmann, H., Pastor, P., Schaal, S.: Dynamical movement primitives: learning attractor models for motor behaviors. Neural Comput. **25**(2), 328–373 (2013)

Khansari-Zadeh, S.M., Billard, A.: BM: an iterative algorithm to learn stable non-linear dynamical systems with Gaussian mixture models. In: 2010 IEEE International Conference on Robotics and Automation, Anchorage, USA, pp. 2381–2388 (2010)

Khansari-Zadeh, S.M., Billard, A.: Learning stable nonlinear dynamical systems with Gaussian mixture models. IEEE Trans. Robot. **27**(5), 943–957 (2011)

Parisi, S., Abdulsamad, H., Paraschos, A., Daniel, C., Peters, J.: Reinforcement learning vs human programming in tetherball robot games. In: 2015 IEEE International conference on Intelligent Robots and Systems, Hamburg, Germany, pp. 6428–6434 (2015)

Peters, J., Schaal, S.: Reinforcement learning of motor skills with policy gradients. Neural Netw. Off. J. Int. Neural Netw. Soc. **21**(4), 682 (2008)

Peters, J., Mülling, K., Altun, Y.: Relative entropy policy search. In: 24th AAAI, Atlanta, Westin, USA, pp. 1607–1612 (2010)

Pfeifer, R., Bongard, J.: How the Body Shapes the Way We Think: A New View of Intelligence. MIT Press, Cambridge (2006)

Rombokas, E., Malhotra, M., Theodorou, E.A., Todorov, E., Matsuoka, Y.: Reinforcement learning and synergistic control of the ACT hand. IEEE Trans. Mechatron. **18**(2), 569–577 (2013). https://doi.org/10.1109/TMECH.2012.2219880

Theodorou, E., Buchli, J., Schaal, S.: A generalized path integral control approach to reinforcement learning. J. Mach. Learn. Res. **11**, 3137–3181 (2010)

Ude, A., Asfour, G.T.A.: Task-specific generalization of discrete and periodic dynamic movement primitives. IEEE Trans. Robot. **26**(5), 800–815 (2010)