Estimating Biped Gait Using Spline-Based Probability Distribution Function With Q-Learning

Lingyun Hu, Changjiu Zhou, and Zengqi Sun, *Senior Member, IEEE*

***Abstract*—This paper studies the probability distribution functions of the parameters to be learned and optimized in biped gait generation. By formulating the gait pattern generation into a multiobjective optimization problem with consideration of geometric and state constraints, dynamically stable and low energy cost biped gaits are generated and optimized by the proposed method, namely Spline-based Estimation of Distribution Algorithm (EDA) with Q-learning updating rule (EDA_S_Q). Instead of assuming variables as independent ones, the relationship between them is exploited by formulating the corresponding probability models with the Catmull–Rom cubic spline function. Such kind of function is proved to be a suboptimal and adaptive realization of the cubic spline function and is capable of providing high-precision description. Moreover, the probability models are updated autonomously by Q-learning method, which is model-free and adaptive. Thus, EDA_S_Q can deal with complex probability distribution functions without a prior knowledge about the distribution. The biped gait generated by EDA_S_Q has been verified using the simulation model of a humanoid soccer robot Robo-Erectus. It also shows that EDA_S_Q can generate the desired biped gaits autonomously in short learning epochs. An interpretation of the transition probability distribution achieved by EDA_S_Q provides us easy understanding for biped locomotion and better control in humanoid robots.**

***Index Terms*—Biped robot, Estimation of Distribution Algorithm (EDA), gait pattern generation, probability model, Q-learning, spline function.**

I. Introduction

RESEMBLING human walking is essential for biped robot control. It requires the cooperation between all actuated joints in searching for the ideal set of gait characteristics like stability [1], energy efficiency [2], and simplicity. Though several techniques have been proposed to compute the expected biped gaits [3], the problem of generating dynamically stable walking trajectories remains a difficult challenge especially when considering some constraints, e.g., energy cost.

To develop a more efficient biped gait generation and optimization method for the humanoid robot "Robo-Erectus" (RE), one of the foremost leading soccer-playing humanoid robots in the RoboCup Humanoid League developed by Advanced Robotics and Intelligent Control Center at Singapore Polytechnic (www.robo-erectus.org) [4], we proposed the Estimation of Distribution Algorithms (EDAs) [5] based gait optimization approaches in our previous works [6]–[8]. They can generate walking patterns by minimizing the specified objective function under several constraints with the assumption that parameters are independent in EDA, which uses probability distributions derived from the optimization function to generate search points instead of crossover and mutation as done by Genetic Algorithms (GAs). However, those parameters are interrelated and the relationship between them gives an all-around understanding of the joint effect on biped gait stability and energy transformation. To explore and exploit it, the Spline-based EDA with Q-learning-based updating rule (EDA_S_Q) is developed in this paper to study the probability distribution of the parameters to be optimized. Moreover, instead of using the Gaussian distribution as the probability model in traditional EDAs, Catmull–Rom cubic spline [9], [10] is used to achieve suboptimal and self-adaptive probability distribution functions, which are updated automatically by the Q-learning-based [11] method without a predesigned updating rule. As part of the project, the learned trajectories have been tested successfully on the biped robot RE.

The rest of this paper is organized as follows. Section II presents a brief survey of related work with an emphasis on why we choose EDA to solve the biped gait optimization problem. Then, the biped gait optimization framework including the objective function and various constraints is described in Section III. Next, the modified EDA using spline probability distribution function and Q-learning-based updating method, namely EDA_S_Q, for solving the optimization problem is proposed in Section IV. Experimental results achieved by EDA_S_Q are given in Section V, and Section VI is the discussion on the probability model and updating rule in terms of probability learning and optimization for biped gait generation. Finally, some concluding remarks and future work are given in Section VII.

II. Related Work

For the lumped-mass model of biped mechanisms, a commonly used method to generate a dynamically stable trajectory is prescribing a time evolution of the Zero Moment Point (ZMP) [12] trajectory such that it lies within the supporting polygon constructed by the biped feet [1], [13]. Besides conceptual simplicity, the advantages of this kind of method include that

Manuscript received January 31, 2006; revised August 23, 2007. This work was supported in part by the Singapore Tote Fund, by the Singapore Polytechnic R&D Fund, and by the National Key Project for Basic Research of China under Grant 2002CB312205. This research was conducted when L. Hu was a Research Associate in Advanced Robotics and Intelligent Control Center (ARICC) at Singapore Polytechnic.

L. Hu and C. Zhou are with the School of Electrical and Electronic Engineering, Singapore Polytechnic, Singapore 139651 (e-mail: ZhouCJ@sp.edu.sg).

Z. Sun is with the Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China (e-mail: szq-dcs@tsinghua.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIE.2007.908526

the closed form solution can be easily derived for on-line implementation. By introducing some computational intelligent algorithms, like neural networks [14], fuzzy systems [15], and GAs [16], the point mass model has been applied widely in biped gait generation and optimization.

However, the number of degrees of freedom (DOFs) in modern humanoid robots is large. It increases the difficulty of finding and representing an optimal biped trajectory for the point mass model. Moreover, parameters to be optimized become sensitive to each other. For such cases, a probability model seems to be one of the preferable solutions because it can describe the distribution with few variables. While reviewing the existing literature, there exist many papers on biped gait synthesis [13], [17]. However, there are very few papers that report the use of a probability model to accelerate the search in high dimensional coupling space for biped gait generation and optimization. In our first attempt to use probability modeling and evolutionary computation, we developed an EDA-based framework for biped gait generation and optimization [6]. Three key poses are selected in one gait cycle to produce the complete gait trajectory by third-order spline functions. Joint coordinates at the transition phases are the parameters to be optimized. Based on the framework, EDAs with different probability distribution functions and updating rules have been tested and compared for different requirements [7], [8], [18], [19].

Among them, the EDA with Catmull–Rom cubic spline-based probability function shows quicker convergence speed compared to traditional EDAs. Such a kind of probability distribution function can approximate an arbitrary continuous function with arbitrarily high quality [20], and it can provide precise description for complex multimodal probability distribution functions. Such cases as shown in the simulation results are most likely to appear in the 2-D probability distribution functions used to describe the relationship between parameters. Traditional probability functions like the Gaussian function are not suitable to describe this relationship. Therefore, the Catmull–Rom cubic spline function is applied in EDA_S_Q as the probability distribution function, which has been proved in this paper to be a suboptimal and adaptive realization of the cubic spline function with the kernels $|x^3|$ (see Appendix I). We also provide guidelines on how to improve the approximation precision in Appendix II.

For the specially-designed probability model, the selection of a suitable updating method needs experience and may convert to new optimization problems. To build and update the probability model from learning instead of exclusively being determined in advance, the idea of reinforcement learning (RL) [4] is incorporated with EDA in this paper. RL can learn the unknown desired distribution by providing an agent with suitable evaluation of its performance. By formulating the relationship between parameters with conditional probability distribution functions, probability models can be updated autonomously by the RL agents, which provide us a new viewpoint to explore the inherent dynamics in biped locomotion.

In conclusion, EDA_S_Q is able to find out the preferable solutions in high dimensional coupling space quickly and approximate the probability model autonomously with the help of the spline-based probability function and Q-learning-based updating rule. As indicated by the comparison experiment with traditional EDAs in Section VI, the faster convergence of EDA_S_Q brought by the spline-based probability function may be helpful for online application. Also, the conditional probability functions updated by Q-learning-based rule can give some tips in biped locomotion and robot control.

III. Problem Definition

The definition of the biped gait generation and optimization problem consists of three elements. 1) Parameters to be optimized. As mentioned in Section II, biped trajectories are parameterized by a third-order spline function of the joint angles at transition poses, which are the correct parameters to be optimized. 2) Desired features of the expected gaits. The commonly used criteria include dynamical stability [12], energy efficiency [2], and so on. In this paper, ZMP displacement and required joint torques are applied to form the objective function pro rata. 3) Description of physical feasibility for biped walking. Geometric constraints for joint motion range and actuator torques are first considered to satisfy the practical requirement. Additionally, state constraints to guarantee the stability and state feasibility like maximum joint rotational velocities are also considered in final problem definition. They are expressed by a set of inequality and equality constraints.

Consequently, among the set of all trajectories satisfying the above constraints, the one with the least sum of ZMP displacement and actuator torques is desired to be found using the proposed method.

For the biped gaits optimization and learning problem, we consider the following objective function constructed using the energy cost and stability criteria

$$\begin{aligned}
\text{Minimize } \Upsilon(\cdot) &= \beta_f f(\cdot) + \beta_g g(\cdot) \\
&= \sum_{i=1}^{N_s} \left(\beta_f \mathbf{N}\left(f_i(\cdot)\right) + \beta_g \mathbf{N}\left(g_i(\cdot)\right)\right) \\
\text{Subject to GC1 } & A_p \le p(\phi) \le B_p \\
\text{GC2 } & A_q \le q \le B_q \\
\text{FC1 } & \sum_{i=1}^{N_1}(m_i \ddot{p}_i) = f_{\mathrm{R}} + f_{\mathrm{L}} + \sum_{i=1}^{N_1}(m_i g) \\
\text{FC2 } & f_{\mathrm{d}} = \min\left(\mathrm{FI}(f_{\mathrm{R}}, f_{\mathrm{L}})\right) \\
\text{VC } & A_{\dot{q}} \le \dot{q} \le B_{\dot{q}} \\
\text{ZC } & A_{\mathrm{zmp}} \le p_{\mathrm{zmp}} \le B_{\mathrm{zmp}}
\end{aligned} \tag{1}$$

where $\Upsilon(\cdot)$ represents the optimization goal to achieve the dynamically stable and energy-efficient biped gaits. $f_i = \|p_{\mathrm{zmp}}^i - p_{\mathrm{zmp}}^d\|_2$ calculates the Euclidean distance between the actual ZMP p_{zmp}^i and the desired ZMP p_{zmp}^d at the ith sample index. $\|\cdot\|_2$ is the second-order norm. $g_i = \sum_{j=1}^{N_q} \tau_{ij}$ summarizes the torque of all actuated joints at the ith sample index. $\mathbf{N}$ is the normalization operator to make the two targets comparable. N_{s} is the number of sampling points in one gait cycle. β_f and β_g are weights satisfying $\beta_f + \beta_g = 1$ to notify the desired character of generated gaits.

The ZMP is defined as the point on the ground at which the net moment of the inertial forces and the gravity forces have no component along the horizontal axes [12]. For kinematic chain structure, the ZMP coordinate in x- and y-direction can be calculated by the following:

$$x_{\mathrm{zmp}} = \frac{\sum_{i=1}^{N_{\mathrm{q}}} \left(m_i(\ddot{z}_i + g)x_i - m_i\ddot{x}_i z_i - (I_i\ddot{\theta}_i)_y \right)}{\sum_{i=1}^{N_{\mathrm{q}}} (m_i(\ddot{z}_i + g))} \tag{2}$$

$$y_{\mathrm{zmp}} = \frac{\sum_{i=1}^{N_{\mathrm{q}}} \left(m_i(\ddot{z}_i + g)y_i - m_i\ddot{y}_i z_i + (I_i\ddot{\theta}_i)_x \right)}{\sum_{i=1}^{N_{\mathrm{q}}} (m_i(\ddot{z}_i + g))} \tag{3}$$

where m_i is the mass of link i and g is the gravity acceleration. The coordinate of link i is described by (x_i, y_i, z_i). Correspondingly, accelerations of link i in x-, y- and z-direction are represented by $\ddot{x}_i$, $\ddot{y}_i$, and $\ddot{z}_i$, respectively. $(I_i)_x$ and $(I_i)_y$ are the inertial components. $(\ddot{\theta}_i)_x$ and $(\ddot{\theta}_i)_y$ are the absolute angular acceleration component around x- and y-axis at the center of gravity of link i.

GC1 and GC2 are geometric constraints to guarantee the feasibility of generated gaits. $p = [p_{t,i}]^{\mathrm{T}}$, $A_p = [A_{p_{t,i}}]^{\mathrm{T}}$, $B_p = [B_{p_{t,i}}]^{\mathrm{T}}$. $p_{t,i} = [x_{t,i}, y_{t,i}, z_{t,i}]$ denotes center position of link i at time t, $i = 1, 2, \ldots, N_{\mathrm{l}}$. $A_{p_{t,i}}$ and $B_{p_{t,i}}$ are lower and upper boundaries of $p_{t,i}$. $q = [q_{t,i}]^{\mathrm{T}}$ stands for the ith joint angle at time t, $i = 1, 2, \ldots, N_{\mathrm{q}}$. $A_q = [A_{q_{t,i}}]^{\mathrm{T}}$, $B_q = [B_{q_{t,i}}]^{\mathrm{T}}$, $A_{q_{t,i}}$ and $B_{q_{t,i}}$ are lower and upper boundaries of $q_{t,i}$. N_{l} and N_{q} are the number of links and actuated torques in biped robot.

Besides geometric constraints, state constraints including force, velocity, and ZMP constraints are also considered in this paper.

FC1 and FC2 are force constraints. f_{R} and f_{L} are the ground reaction force at the right and left foot, respectively. $\ddot{p}_i$ is the acceleration of link i. Since only the sum $f_{\mathrm{R}} + f_{\mathrm{L}}$ is known during the double support phase, the force constraint FC2 is assumed as the internal force f_{d}, which must be minimized in the closed loop structure. FI is the function to calculate the internal force.

VC is velocity constraint. $A_{\dot{q}} = [A_{\dot{q}_{t,i}}]^{\mathrm{T}}$ and $B_{\dot{q}} = [B_{\dot{q}_{t,i}}]^{\mathrm{T}}$ are lower and upper boundaries of $\dot{q}_{t,i}$.

ZC stands for ZMP constraint. According to the dynamic stability criterion defined by ZMP [12], the position of ZMPs should be within the stable region, which changes from the area of the standing feet to the convex polygon formulated by the two feet when gaits change from single support to double support phase. Where $p_{\mathrm{zmp}} = [p_{\mathrm{zmp},i}]^{\mathrm{T}}$, $p_{\mathrm{zmp},i} = (x_{\mathrm{zmp},i}, y_{\mathrm{zmp},i}, 0)$ is position of the ith ZMP. $A_{\mathrm{zmp}} = [A_{\mathrm{zmp},i}]^{\mathrm{T}}$ and $B_{\mathrm{zmp}} = [B_{\mathrm{zmp},i}]^{\mathrm{T}}$ are lower and upper boundaries of the stable region, respectively.

For details of the dynamical and gait pattern models used to calculate the parameters in gait generation, please refer to [7], [8], and [21].

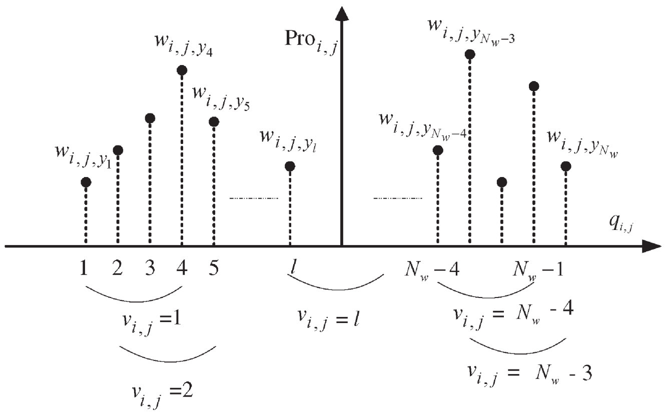

Fig. 1. Structure of spline function for $\mathrm{Pro}_{i,j}$.

IV. EDA_S_Q for Biped Gait Optimization

For the simplified nonlinear parameter optimization problem with inequality constraints, if all DOFs would be optimized, trajectories with a smaller objective function value could be obtained. Different from previous heuristic methods, the proposed algorithm adopts a new evolutionary algorithm called EDAs to search for the joint angle permutation. For unknown probability distribution functions, EDAs build the probability model by using the selected set of solutions and making use of this estimation to generate new solutions. As the input coordinates q_i $(i = 1, 2, \ldots, N_{\mathrm{l}})$ of the N_{l} links in joint space are considered to be interrelated, the interrelationship between parameters needs to be formulated autonomously and accurately. This is realized by setting a spline-based probability function and Q-learning-based updating rule in EDA_S_Q.

A. Spline Function Based Probability Model

Different from traditional probability models, the proposed algorithm employs piecewise polynomial spline interpolation schemes, which are a continuous first derivative and have local adaptation, to describe the probability model. The structure of probability distribution function $\mathrm{Pro}_{i,j}$ is illustrated in Fig. 1.

Supposing $q_{i,j}$ is the angular position of joint j at the ith key pose in one gait cycle, $i = 1, 2, \ldots, N_{\mathrm{kp}}$, $j = 1, 2, \ldots, N_{\mathrm{q}}$, then the output of this probability distribution function $\mathrm{Pro}_{i,j}$ can be calculated by

$$v_{i,j} = \left\lfloor \frac{q_{i,j}}{\Delta w} + \frac{N_w}{2} \right\rfloor \tag{4}$$

$$u_{i,j} = \frac{q_{i,j}}{\Delta w} + \frac{N_w}{2} - v_{i,j} \tag{5}$$

$$\mathrm{Pro}_{i,j} = F_{i,j,y_v}(\cdot) = \sum_{m=0}^{3} w_{i,j,x_{v_{i,j}+m}} C_m(u_{i,j}). \tag{6}$$

Equations (4) and (5) implement the computation for local parameters. $\lfloor\ \rfloor$ is the floor operator and it is designed to guarantee that $u_{i,j}$ is always nonnegative. $\Delta w = w_{x_{v+1}} - w_{x_v}$. It is proved that function smoothness can be tuned by Δw, which is insensitive to generalization error in a suitable range [22]. Hence, it is not necessary to learn the optimal Δw using a complex procedure. The proposed method sets it by manual tuning. N_w is the number of points and

$$\begin{pmatrix} C_0(u) \\ C_1(u) \\ C_2(u) \\ C_3(u) \end{pmatrix}^{\mathrm{T}} = \frac{1}{2}\begin{pmatrix} u^3 \\ u^2 \\ u \\ 1 \end{pmatrix}^{\mathrm{T}} \times \begin{pmatrix} -1 & 3 & -3 & 1 \\ 2 & -5 & 4 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix}.$$

As shown in Appendix II, an accurate prediction of the Catmull–Rom spline approximation error can be described by function $w_{\text{int}}\Delta x^L|\tilde{F}^L(v)|_{L_2}$ with the sampling step Δw and approximation order L.

Remark 1: The higher the approximation order L, the better the approximation quality. However, the high order will inevitably add to the computational load. The third order is a practical choice.

Remark 2: For basis functions of identical approximation order, the smaller the approximation constant w_{int}, the better the approximation quality.

Since the Catmull–Rom spline is a suboptimal of the cubic spline function as demonstrated by Appendix I, it can also be understood as a subclass of Moms functions, which stands apart as the best achievable compromise between approximation quality and speed [20]. By setting such kinds of probability distribution functions, the proposed method can characterize more complex distribution functions because every continuous function on a closed interval can be approximated uniformly to any prescribed accuracy by a spline function [23].

B. Q-Learning

Besides the probability distribution function, the updating rule is also important to the learning quality of EDAs. For distribution functions in traditional EDAs, updating rules should be specially designed with consideration of probability function type, learning rate and so on. To update it autonomously, the idea of RL is employed in this paper.

In RL, the optimal action selection method or policy is obtained by maximizing the numerical reward signal. Q-learning is a commonly used off-policy temporal difference control algorithm that approximates the optimal action-value function independent of the policy being followed. The Q-value, which is updated after every state transition, is calculated by

$$Q(s,a) = Q(s,a) + \alpha\left(r + \gamma \max Q(s',a') - Q(s,a)\right) \quad (7)$$

where s' is the following state of s after action a and a' is the corresponding action set for state s. r is the scalar feedback provided by the critic. Let Q^* be the global optima, it is proved that if each action is executed in each state an infinite number of times on an infinite run and α is decayed appropriately, the Q-value will converge with probability one to Q^* [24].

The idea of updating the probability function with Q-learning is efficiently used here to enhance the intelligence of the proposed method. A multivariate distribution model $\text{Pro}_{i+1|i,j}$ is given to represent the conditional probability of the jth variable in the $i+1$th moment with the given value in the ith moment. An example of the conditional probability function $\text{Pro}_{2|1,j}$ for the relationship between the first and the second key poses is exhibited in Fig. 2. Points in conditional probability distributions are 2-D interpolated by sample points, which are updated by Q-learning method as

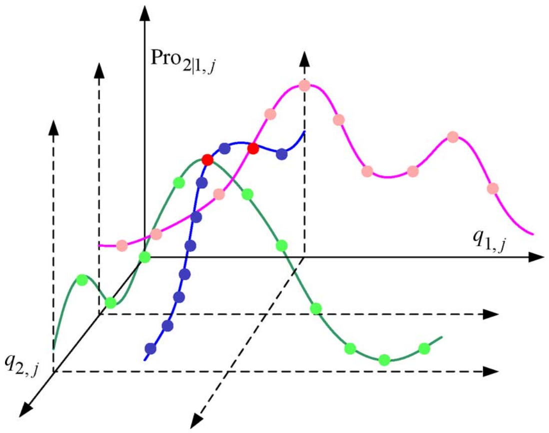

Fig. 2. Spline-based probability distribution function $\text{Pro}_{2|1,j}$.

$$\begin{aligned}\text{Pro}_{i+1|i,j} &= \text{Pro}_{i+1|i,j} \\ &\quad + \alpha\big(r(q_{i,j}) + \gamma \max \text{Pro}_{i+1,j} - \text{Pro}_{i+1|i,j}\big) \quad (8)\end{aligned}$$

where $i = 1, 2, \ldots, N_{\text{kp}} - 1, j = 1, 2, \ldots, N_{\text{q}}$. Reward r is specially designed with the same structure as that in the objective function. It is calculated by the selected N_{b} best samples as

$$r(q_{i,j}) = \beta_f r_{\text{local}}(q_{i,j}) + \beta_g r_{\text{global}}(q_{i,j}) \quad (9)$$

where local reward $r_{\text{local}}(q_{i,j}) = \mathbf{N}(\tau_{i,j}^{-1})$ focuses on energy cost at each actuated joint and it affects mainly on the corresponding joint; while the global reward $r_{\text{global}}(q_{i,j}) = \mathbf{N}(f_j^{-1})$ is the ZMP displacement at current sample moment. All joints have an effect on this factor. These two parts give an all-around estimation on reward and can provide proper feedback to Q-learning.

Remark 3: It can be seen that conditional probability is updated with consideration of the reward defined by current state parameters and the probability at the next pose. The rewards determine the immediate, intrinsic desirability of states. The probability at the next state indicates the long-term desirability of states after taking into account the states that are likely to follow, and the rewards available in those states. Such a kind of updating rule can be applied for various kinds of probability functions without foreknown knowledge.

C. Proposed Method EDA_S_Q

As mentioned in Section II, a total of $N_{\text{kp}} = 3$ key poses are determined in one complete gait cycle. The probability distribution functions of the joint angles at the first key pose are modeled by spline-based probability distribution functions. They are updated by

$$\text{Pro}_{1,j} = (1-\alpha)\text{Pro}_{1,j} + \alpha r(q_{1,j}). \quad (10)$$

The relationship between the joint angles at sequential poses is also formulated by the same kind of probability distribution function and updated with the Q-learning method as shown in (8). The information at current state and future state estimation is all taken into account in terms of Q-value. Thus, it is possible for a conditional probability distribution function to adjust the distribution autonomously in learning.

Hence, for input $q_{1,j}$, the probability distribution for $q_{2,j}$ and $q_{3,j}$ can be achieved by

$$\text{Pro}_{2,j} = \text{Pro}_{2|1,j}\text{Pro}_{1,j} \tag{11}$$

$$\text{Pro}_{3,j} = \text{Pro}_{3|2,j}\text{Pro}_{2,j}. \tag{12}$$

In conclusion, EDA_S_Q describes evolution as

$$q(t+1) = B^{N_\text{e}}\Upsilon R^{N_\text{b}}\text{Sq}(t) \tag{13}$$

where $\text{Sq}(t)$ defines the spline-based probability distribution function of the offspring. From this distribution, a population of N_e offspring is sampled via random selection R^{N_b} and evaluated by the fitness operator Υ. Proportional to the fitness, a population of N_b parents is selected by the selection method B^{N_e}.

The final structure of the proposed method EDA_S_Q can be outlined as follows:

1) **Initialization** Set $k = 1$. Randomly initialize $\text{Pro}_{i+1|i,j}(k)$ and $\text{Pro}_{l,j}(k)$, $i = 1, 2, \ldots, N_\text{kp} - 1$, $j = 1, 2, \ldots, N_\text{q}$, $l = 1, 2, \ldots, N_\text{kp}$.
2) **Sampling** Generate N_e samples q^{s_m} from $\text{Pro}_{i,j}(k)$ to form the current population $O(k)$ by (6), $m = 1, 2, \ldots, N_\text{e}$, $i = 1, 2, \ldots, N_\text{kp}$.
3) **Selection** Select the N_b best points q^{b_n} from q^{s_m} according to Υ calculated by (1), $N_\text{b} = \alpha_\text{s}N_\text{s}(0 < \alpha_\text{s} < 1)$, $n = 1, 2, \ldots, N_\text{b}$.
4) **Updating** Calculate the reward $r(q_{i,j})$ according to (9). Update $\text{Pro}_{1,j}(k+1)$ and $\text{Pro}_{i+1|i,j}(k+1)$ by (8) and (10), respectively. New probabilities of $\text{Pro}_{2,j}(k+1)$ and $\text{Pro}_{3,j}(k+1)$ are obtained by (11) and (12).
5) If stop condition is not met, go back to step 2) and let $k = k + 1$.

V. Experimental Results

To show the effectiveness of EDA_S_Q for biped gait generation and optimization, it is applied to the simulation model of the humanoid robot called RE. The height and weight of RE are 600 mm and 4.6 kg, respectively. A total of 23 DOFs are designed in RE with six per leg, four per arm, one for head, and two for waist. They are driven by a servomotor with maximum torque of 30 kg $\cdot$ cm. A laptop computer with a Windows XP operation system is set on RE for online control. The simplified model takes the basic structure parameters as the number of links $N_\text{l} = 9$, number of key poses $N_\text{kp} = 3$. A total of $N_\text{s} = 20$ poses are sampled in one gait cycle, and the first, the sixteenth and the eighteenth of them are key poses. The width of hip $l_h = 15$ cm, other links take the value of foot length $l_1 = 10$ cm, fore foot length $l_a = 5$ cm, heel length $l_b = 5$ cm, ankle length $l_2 = 4$ cm, crus and thigh length $l_3 = l_4 = 8$ cm, trunk length $l_5 = 20$ cm, step length $D_\text{s} = 30$ cm, and the max height of swing foot $D_h = 3$ cm. Masses of links $m_1 = m_2 = 0.1$ kg, $m_3 = m_4 = 0.2$ kg, $m_5 = 1$ kg. Desired ZMP is set as the middle line in the stable region.

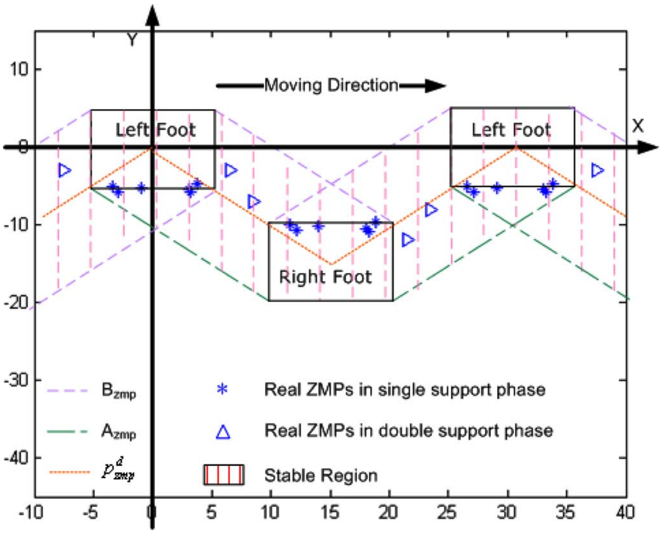

Fig. 3. ZMPs of the learned biped gait.

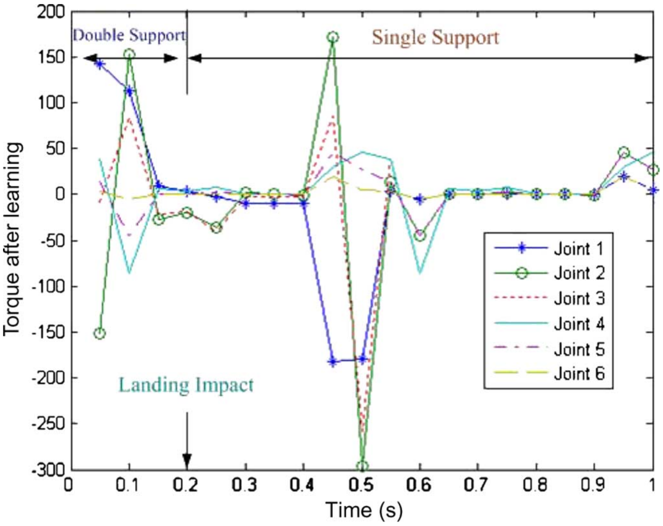

Fig. 4. Torques of the biped gait after learning.

Parameters used in EDA_S_Q are $N_\text{e} = 8$, $N_\text{b} = 4$, $\beta_f = \beta_g = 0.5$, $\alpha = 0.01$, $\gamma = 0.1$. The number of sample points for the spline function is chosen as $N_w = 20$ empirically [22]. EDA_S_Q stops learning when $\Upsilon(k) < 2$ for four continuous optimization epochs or the learning index $k > 200$.

Fig. 6 shows the objective function value obtained by EDA_S_Q. It reduces from about 4.5 to less than 3 in 100 epochs and finally converges to 2.9 in the desired 200 iterations. ZMP trajectory and torques of the learned gait are shown in Figs. 3 and 4, respectively. The torque before learning is exhibited in Fig. 5 for comparison. It can be seen that ZMP trajectory stays in a stable region without too much margin because it will cost too much energy. Since the objective function considers both stability and energy cost, only those gaits that are not only stable but also efficient will be finally selected. On average, both criteria are significantly lower in comparison to that before learning.

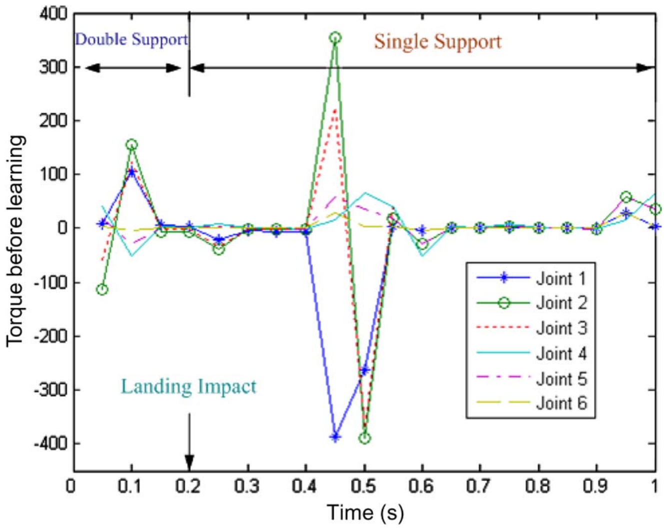

Fig. 5. Torques of the biped gait before learning.

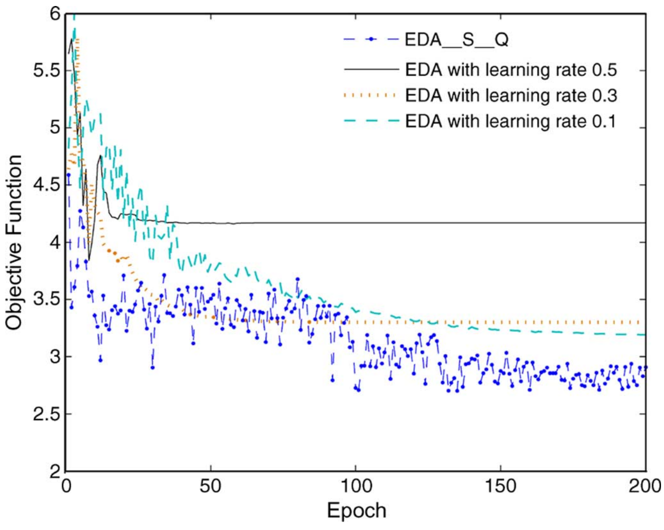

Fig. 6. Objective function variation obtained by EDA.

VI. Discussion

For biped gait generation and optimization, if only the stability criterion is considered, traditional gait planning methods [13] can deal with them. However, for the multiobjective problem described in (1), heuristic methods are necessary for intelligent searching. Moreover, as indicated in Section II, traditional GAs depend on a lot of parameters, which is unfit for the multi-DOFs structure learning. To emphasize the precise description brought by the special kind of probability distribution function, biped gait optimization is performed on traditional EDA with Gaussian function $G(\mu, \sigma)$-based probability distribution function for comparison. Means μ and covariance σ are updated by (14) and (15), respectively [6].

$$\begin{aligned}\mu_{i,j}(k+1) = {} & (1-\alpha)\mu_{i,j}(k) \\ & + \alpha(\mu_{i,j,b}(k) + \mu_{i,j,2b}(k) - \mu_{i,j,w}(k))\end{aligned} \tag{14}$$

$$\begin{aligned}\sigma_{i,j}(k+1) = {} & (1-\alpha)\sigma_{i,j}(k) + \alpha \\ & \times \sqrt{\frac{1}{N_\text{b}} \sum_{i=1}^{N_\text{b}} (\mu_{i,j,l}(k) - \tilde{\mu}_{i,j,l}(k))^2}\end{aligned} \tag{15}$$

where $\mu_{i,j,b}(k)$, $\mu_{i,j,2b}(k)$ and $\mu_{i,j,w}(k)$ are values of the best, second best and worst individual (with respect to the objective function Υ) for $q_{i,j}$ at iteration k, $\mu_{i,j,l}$ are the N_b best individuals and $\tilde{\mu}_{i,j,l}$ is their mean, $l = 1, 2, \ldots, N_\text{b}$. α is the learning rate. Other parameters in EDA are the same as that in EDA_S_Q.

Objective function values obtained by EDA with different learning rates are shown in Fig. 6. It can be seen that an EDA with smaller learning rate (0.1) has longer convergent ability but with slower learning speed, while large learning rate (0.5) has worse convergent precision. However, even the preferable choice of learning rate (0.3) cannot get the same precision after 200 learning iterations. Moreover, the function value of traditional EDAs decreases along learning but from higher initial values (about 5.0 in average) with comparison to that (about 4.6) of EDA_S_Q. The spline-based probability model improves the description precision for EDA_S_Q and thus can start the optimization from a lower initial value. Table I compares the experimental results of EDA_S_Q and that of traditional EDA with $\alpha = 0.3$, which is the best in traditional ones. The quantitative comparison demonstrates that EDA_S_Q is better than the traditional EDA in terms of convergence speed.

TABLE I
EDA and EDA_S_Q

	EDA (0.3)	EDA_S_Q
Probability Model	Gaussian function	Spline function
Updating Rule	Partial Replacing	Q-learning
$\Upsilon(1)$	4.7	4.6
First k for $\Upsilon(k) \leq 3.5$	31	11
Minimum $\Upsilon(k)$	3.36	2.76

Details of the efficiency for the spline-based probability model have been proved in [7]. In this paper, we will focus more on the function of Q-learning-based updating rule. In the special application for biped gait optimization and learning, the probability model of each joint is modified not only by selected solutions but also affected by the probability model of the same joint at the next key pose. Instead of assuming joint angles as independent variables, EDA_S_Q employs the inner function between them to formulate the correct probability function. The interrelationship of joint angles at successive moments is also taken into consideration to update each probability model. Therefore, as shown in Table I, the permutation of joint angles at key poses with a smaller objective function value ($\Upsilon = 3.5$) can be achieved more quickly by EDA_S_Q (about ten iterations) than traditional EDA (at least 30 iterations).

Giving an example of the conditional probability distribution function $\text{Pro}_{2|1,2}$ of joint 2 between the first and the second key poses, variation of $\text{Pro}_{2|1,2}$ is shown in Fig. 7. The four figures record the distribution at the initial time [Fig. 7(a)], the 30th iteration [Fig. 7(b)], the 60th iteration [Fig. 7(c)], and the 100th iteration [Fig. 7(d)] in sequence.

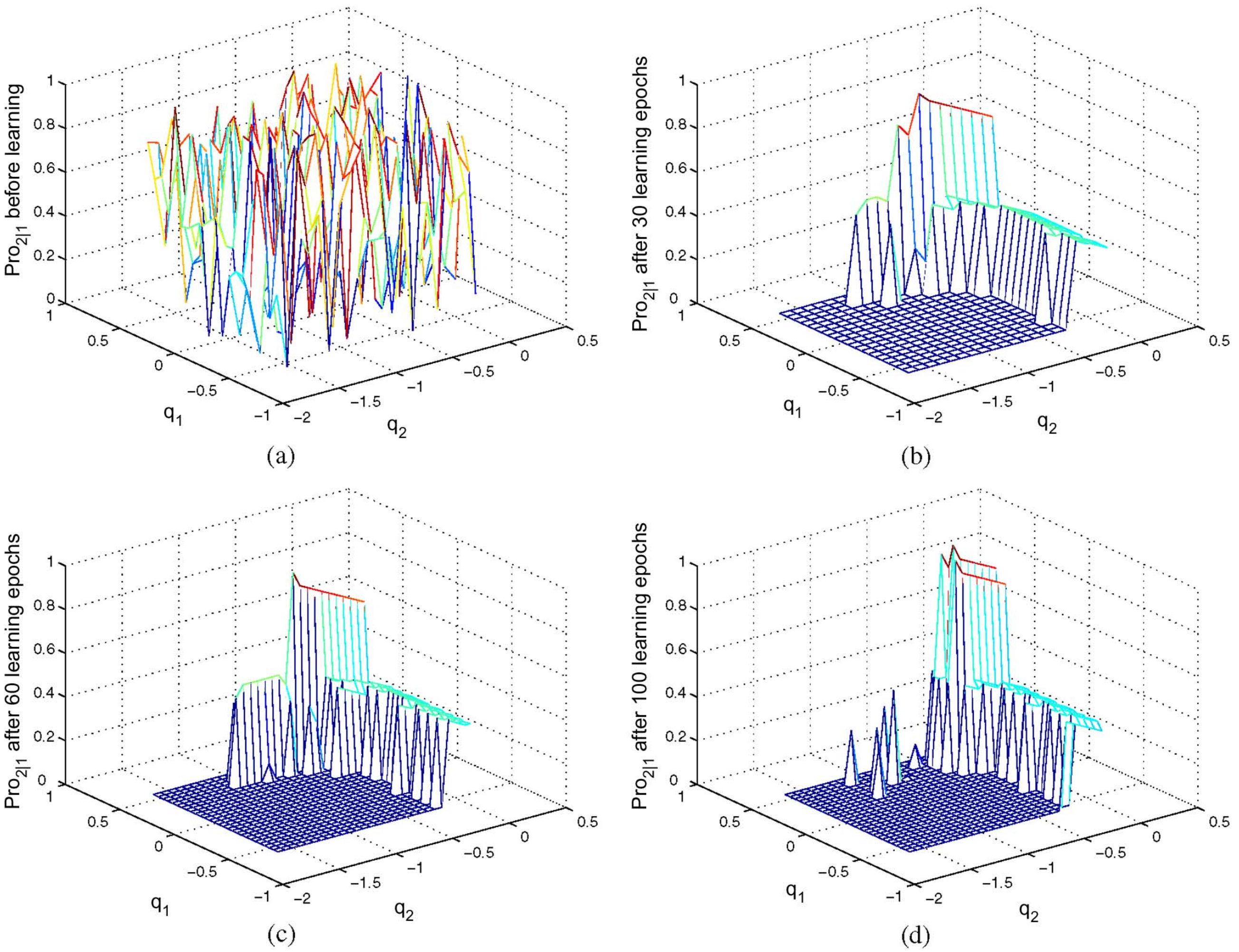

Fig. 7. $\mathrm{Pro}_{2|1,2}(k)$ during learning. (a) $k = 1$. (b) $k = 30$. (c) $k = 60$. (d) $k = 100$.

The peaks in Fig. 7 stand for the promising searching space. For example, for given joint angles of joint 2 at the first key pose $q_{1,2}(t) = 0.3$, the preferable solution for the same joint at the second key pose would be $q_{2,2}(t) = -0.2$ with the probability about 0.8. Here, $q_{2,2}(t)$ is not obtained directly from the corresponding probability distribution but from the conditional probability, which reflects the interrelationship between $q_{1,2}(t)$ and $q_{2,2}(t)$.

Thus, it can be concluded that: 1) the motor for joint 2 at the second key pose should work mainly during $[-0.5, 0]$ because $\mathrm{Pro}_{2|1,2}$ has a large probability value in this range whatever $q_{1,2}$ is. 2) [0, 0.5] is the desired region of motion for joint 2 at the first key pose because $\mathrm{Pro}_{2|1,2}$ is flat when $q_{1,2}$ alters in [0, 0.5]. This means that under such conditions, $q_{2,2}$ can vary in a correspondingly large region with legible poses. Similar conclusions can also be achieved through similar discussion on other conditional probability distribution functions.

The two merits, learning capability and precise description ability, that are inherited from Q-learning and spline-based EDA provide EDA_S_Q the remarkable ability to approximate a complex probability model without using prior knowledge in short learning epochs. It is also a useful tool to explore the relationship between interrelated parameters to be optimized, which may help us understand the biped locomotion and to control biped robots.

VII. Conclusion

There are two factors that affect the learning quality of EDA: the probability model and the updating rule. In this paper, we look at both factors to develop a new EDA with spline-based probability function and Q-learning-based updating rule (EDA_S_Q), which is able to more efficiently generate and optimize dynamically stable and low energy cost biped gaits. To deal with the relationship between the parameters of conjoint poses, their probability distribution functions are formulated without prior knowledge. Q-learning operates as a very efficient updating rule to improve the distribution function with simple rewards. By means of the proposed EDA_S_Q, desired biped gaits are generated and optimized with acceptable convergence time. Some suggestions on working scope for motors at joints can also be achieved through analyzing the conditional probability functions.

The experimental results show that the proposed EDA_S_Q is significantly better than traditional EDA in terms of convergence speed to achieve dynamically stable and energy-efficient biped gaits. The generated gaits can also be used to drive our humanoid soccer robot, RE, which is one of the foremost leading soccer-playing humanoid robots in the RoboCup Humanoid League. To the best of our knowledge, the proposed EDA_S_Q is the first such kind of work in the framework of EDA and biped gait generation and optimization.

Based on the results of this paper, the body dynamics of the mechanical robot will be studied with transition probability models in future work. Our research challenge lies in the interpretation of transition probability models for biped locomotion so that we can progress toward a better understanding of human locomotion and extend the results to better control of humanoid robots.

Appendix I

Catmull–Rom cubic spline function f_{c} is a suboptimal realization of the cubic spline function f_{s} obtained by regularization theory.

For $v_j \le x \le v_{j+1}$, $1 \le j < N$,

$$\begin{aligned} f_{\mathrm{c}}(x) &= \sum_{i=1}^{N} \alpha_i |x - v_i|^3 \\ &= \alpha_1 |x - v_1|^3 + \cdots + \alpha_{j-1} |x - v_{j-1}|^3 \\ &\quad + \alpha_j |x - v_j|^3 - \alpha_{j+1} |x - v_{j+1}|^3 \\ &\quad - \alpha_{j+2} |x - v_{j+1}|^3 - \cdots - \alpha_N |x - v_N|^3 \end{aligned}$$

where α_i are coefficients and v_i are interpolation samples. $i = 1, \ldots, N$, N is the number of samples. Collecting the terms of equal degree, f_{s} can be expressed as

$$f_{\mathrm{s}}(x) = Ax^3 + Bx^2 + Cx + D.$$

To reproduce the interval $[v_j, v_{j+1}]$ with Catmull–Rom cubic spline, four control points W_1, W_2, W_3, W_4 are set on the curve equally as $\Delta x = v_{j+1} - v_j$, $w_{x_2} = v_j$, $w_{x_3} = v_{j+1}$. Thus

$$\begin{aligned} f_{\mathrm{c}}(x) &= \frac{1}{2}(-w_{y_1} + 3w_{y_2} - 3w_{y_3} + w_{y_4}) \left(\frac{x - v_j}{\Delta x} \right)^3 \\ &\quad + \frac{1}{2}(2w_{y_1} - 5w_{y_2} + 4w_{y_3} - w_{y_4}) \left(\frac{x - v_j}{\Delta x} \right)^2 \\ &\quad + \frac{1}{2}(-w_{y_1} + w_{y_3}) \left(\frac{x - v_j}{\Delta x} \right) + w_{y_2} \\ &= \bar{A}x^3 + \bar{B}x^2 + \bar{C}x + \bar{D} \\ &= f_{\mathrm{s}}(x). \end{aligned}$$

These calculations show that Catmull–Rom spline with properly chosen control points is equivalent to the optimal cubic spline everywhere between v_j and v_{j+1}.

Appendix II

This appendix considers the general problem of reconstruction of a function $D(x)$ on the continuous variable x from a discrete set of measurements collected on a uniform grid with step size Δx. The main interest is to quantify the difference between $D(x)$ and its approximated version $F_{\Delta x}(x)$.

Corollary: $\lim_{\Delta x \to 0} |F_{\Delta x}(x) - D(x)| = \lim_{\Delta x \to 0} |\varepsilon| = 0$ with the constraints that $F_{\Delta x}(k\Delta x) = D(k\Delta x)$. Where $F_{\Delta x} = \sum_{k \in \chi^d} w_k C((x/\Delta x) - k)$, $x \in R^d$, w_k are samples at integer coordinates. $C(x)$ is the interpolation function. $\chi_i = [\lfloor (x_i(t)/\Delta x) + (N/2) \rfloor, \ldots, \lfloor (x_i(t)/\Delta x) + (N/2) \rfloor + 3]$. N is the number of total samples. To satisfy the requirement of exact interpolation, $C(x)$ must vanish for all integer arguments except at the origin, where it must take a unit value.

Proof: For the mean square

$$\begin{aligned} \varepsilon^2(\Delta x) &= |F_{\Delta x}(x) - D(x)|_{L_2}^2 \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (F_{\Delta x}(x) - D(x))^2 \, dx_1, \ldots, dx_d \end{aligned}$$

the following formula predicts the approximation error in the Fourier domain [25]:

$$\eta^2(\Delta x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \left| \bar{F}(\delta) \right|^2 E_{\mathrm{int}}(\delta \Delta x) d\delta_1, \ldots, d\delta_d$$

where $\bar{F}(\delta) = \int F(x) e^{2\pi i \delta x} dx$ is the Fourier transform of the arbitrary function $F(x)$, E_{int} is an interpolation error kernel that depends on the basis function only. It is given by

$$E_{\mathrm{int}}(\delta) = \frac{\left| \sum_{k \in \chi^d} \hat{C}(\delta + 2\pi k) \right|^2 + \sum_{k \in \chi^d} \left| \hat{C}(\delta + 2\pi k) \right|^2}{\left| \sum_{k \in \chi^d} \hat{C}(\delta + 2\pi k) \right|^2}.$$

$\varepsilon = \eta$ holds for band limited functions [25]. A decrease in the sampling width Δx will result in a decrease of the argument of E_{int}. Thus, the error kernel must vanish at the origin such that

$$\eta^2(\Delta x) = \lim_{\Delta x \to 0} (w_{\mathrm{int}})^2 \Delta x^2 \frac{1}{2\pi} \int_{-\infty}^{\infty} \left| \delta^L \hat{F}(\delta) \right|^2 d\delta.$$

The vanishing rate of error kernel is controlled by approximation order L and a constant

$$w_{\mathrm{int}} = \lim_{\delta \to 0} \frac{\sqrt{E_{\mathrm{int}}(\delta)}}{\delta^L}.$$

Finally, we have

$$\begin{aligned} \lim_{\Delta x \to 0} \eta(\Delta x) &= \lim_{\Delta x \to 0} |F_{\Delta x}(x) - F(x)| \\ &= w_{\mathrm{int}} \Delta x^L \left| \hat{F}^L(\delta) \right|_{L_2}. \end{aligned}$$

Therefore, for any smooth function $w(x)$ with approximation order L and a constant w_{int}, the approximation error ε predicted by η decreases like Δx^L, where Δx is sufficiently small.

Remark 4: Approximation order L gives a global estimation of approximation speed when the sampling width Δx gets finer.

Remark 5: The constant w_{int} ranks the quality of basis functions that having the same approximation order L. A smaller w_{int} corresponds to a better $w(x)$.

REFERENCES

[1] M. Vukobratovic and B. Borovac, "Zero moment point—Thirty five years of its life," *Int. J. Humanoid Robot.*, vol. 1, no. 1, pp. 157–173, 2004.

[2] T. McGeer, "Passive dynamic walking," *Int. J. Robot. Res.*, vol. 9, no. 2, pp. 62–82, 1990.

[3] F. Asano, M. Yamakita, N. Kamamichi, and Z. Luo, "Biped gait generation and control based on a unified property of passive dynamic walking," *IEEE Trans. Robot.*, vol. 21, no. 4, pp. 754–762, Aug. 2005.

[4] C. Zhou and P. K. Yue, "Robo-erectus: A low cost autonomous humanoid soccer robot," *Adv. Robot.*, vol. 18, no. 7, pp. 717–720, Aug. 2004.

[5] P. Larranaga and J. A. Lozano, *Estimation of Distribution Algorithms: A New Tool for Evolutionary Computation*. Boston, MA: Kluwer, 2001.

[6] L. Hu, C. Zhou, and Z. Sun, "Biped gait optimization using estimation of distribution algorithm," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Tsukuba, Japan, 2005, pp. 283–288.

[7] L. Hu, C. Zhou, and Z. Sun, "Biped gait optimization using spline function based probability model," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2006, pp. 830–835.

[8] L. Hu, C. Zhou, and Z. Sun, "Estimating probability distribution with Q-learning for biped gait generation and optimization," in *Proc. IEEE Int. Conf. Int. Robots Sys.*, Beijing, China, 2006, pp. 362–368.

[9] I. J. Schoenberg, "Contributions to the problem of approximation of equidistant data by analytic functions. Part A—On the problem of smoothing or graduation. A first class of analytic approximation formulas," *Q. Appl. Math.*, vol. IV, no. 1, pp. 45–99, 1946.

[10] I. J. Schoenberg, "Contributions to the problem of approximation of equidistant data by analytic functions. Part B—On the problem of osculatory interpolation. A second class of analytic approximation formulae," *Q. Appl. Math.*, vol. IV, no. 2, pp. 112–141, 1946.

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

[12] M. Vukobratovic, B. Borovac, D. Surla, and D. Stokic, *Biped Locomotion: Dynamics, Stability, Control, and Application*. New York: Springer-Verlag, 1990.

[13] Q. Huang, K. Yokoi, S. Kajita, K. Kaneko, H. Arai, N. Koyachi, and K. Yanie, "Planning walking patterns for a biped robot," *IEEE Trans. Robot. Autom.*, vol. 17, no. 3, pp. 280–289, Jun. 2001.

[14] G. Endo, J. Morimoto, J. Nakanishi, and G. M. W. Cheng, "An empirical exploration of a neural oscillator for biped locomotion control," in *Proc. Int. Conf. Robot. Autom.*, 2004, vol. 3, pp. 3036–3042.

[15] Z. Liu and C. Li, "Fuzzy neural network quadratic stabilization output feedback control for biped robots via H_∞ approach," *IEEE Trans. Syst., Man, Cybern.*, vol. 33, no. 1, pp. 67–84, Feb. 2003.

[16] G. Capi, S. Kaneko, K. Mitobe, L. Barolli, and Y. Nasu, "Optimal trajectory generation for a prismatic joint biped robot using genetic algorithms," *Robot. Auton. Syst.*, vol. 38, no. 2, pp. 119–128, Feb. 2002.

[17] D. Katic and M. Vukobratovic, "Survey of intelligent control techniques for humanoid robots," *J. Intell. Robot. Syst.*, vol. 37, no. 2, pp. 117–141, Jun. 2003.

[18] Q. Zhang, J. Sun, E. Tsang, and J. Ford, "Hybrid estimation of distribution algorithm for global optimization," *Eng. Comput.*, vol. 21, no. 1, pp. 91–107, 2004.

[19] L. Hu, C. Zhou, and Z. Sun, "Optimizing biped gait using probability model," in *Proc. Int. Conf. Comput. Intell., Robot. Auton. Syst.*, Singapore, Dec. 2005.

[20] P. Thvenaz, T. Blu, and M. Unser, "Interpolation revisited," *IEEE Trans. Med. Imag.*, vol. 19, no. 7, pp. 739–758, Jul. 2000.

[21] J. W. Grizzle, G. Abba, and F. Plestan, "Asymptotically stable walking for biped robots: Analysis via systems with impulse effects," *IEEE Trans. Autom. Control*, vol. 46, no. 1, pp. 51–64, Jan. 2001.

[22] L. Vecci, F. Piazza, and A. Uncini, "Learning and approximation capabilities of adaptive spline activation function neural networks," *Neural Netw.*, vol. 11, no. 2, pp. 259–270, Mar. 1998.

[23] J. H. Ahlberg, E. N. Nilson, and J. L. Walsh, *The Theory of Splines and Their Applications*. New York: Academic, 1967.

[24] C. J. C. H. Watkins, "Learning from delayed rewards," Ph.D. dissertation, Univ. Cambridge, Psychol. Dept., Cambridge, U.K., 1989.

[25] T. Blu and M. Unser, "Quantitative Fourier analysis of approximation techniques: Part I—Interpolators and projectors," *IEEE Trans. Signal Process.*, vol. 47, no. 10, pp. 2783–2795, Oct. 1999.

Lingyun Hu received the M.S. and Ph.D. degrees in computer science from Tsinghua University, Beijing, China, in 2003 and 2007, respectively.

From 2005 to 2006, she was a Researcher Associate in the School of Electrical and Electronic Engineering at Singapore Polytechnic, and worked on the Singapore TB Model Project of Development of A Full-Scale Humanoid Robot. Now, she works as a Research Scientist at the Advanced Robotics and Intelligent Control Centre of Singapore Polytechnic, Singapore. She is also a member of the Organizing Committee of the Humanoid League at RoboCup 2008. Her research focuses on humanoid robotics, intelligent learning and optimization, evolutionary computation, and soft computing.

Changjiu Zhou received the B.Eng. and M.Eng. degrees in electrical engineering from Jilin University (formerly Jilin University of Technology), China, in 1985 and 1988, respectively, and the Ph.D. degree in control engineering from Dalian Maritime University, China, in 1997.

He is currently Director of the Advanced Robotics and Intelligent Control Center at Singapore Polytechnic, Singapore, where he founded humanoid robotics group Robo-Erectus (www.robo-erectus.org) which won first place in Humanoid Free Performance competition at RoboCup 2003 and second place in all the four humanoid competitions at RoboCup 2004. He has been serving as a Guest Professor at Dalian Maritime University, Dalian, China, since 2001. He is also an Adjunct Professor at Jilin University, China. He is an Associate Editor of the *International Journal of Humanoid Robotics* and the *IES Journal B: Intelligent Devices and Systems* and also served as Guest Editor for some journal special issues, e.g., *Fuzzy Sets and Systems*. He has about 150 research publications including three edited books published by Springer-Verlag. His current research interests include intelligent robotic systems, computational intelligence, humanoid robotics, multirobotic systems, machine learning, intelligent control, and educational robotics.

Dr. Zhou is currently a member of the RoboCup Executive Committee, where he was Chair of the technical committee for the Humanoid League from 2003 to 2004.

Zengqi Sun (SM'93) received the degree from the Department of Automatic Control, Tsinghua University, Beijing, China, in 1966, and the Ph.D. degree in control engineering from the Chalmers University of Technology, Göteborg, Sweden, in 1981.

He is currently a Professor of the Department of Computer Science and Technology, Tsinghua University, Beijing, China. His current research interests include intelligent control, robotics, networked control, fuzzy systems, neural networks, and evolution computing, etc.