Neural Processing Letters (2007) 26:121–131
DOI 10.1007/s11063-007-9046-9

A Global Minimization Algorithm Based on a Geodesic of a Lagrangian Formulation of Newtonian Dynamics

Joon Shik Kim · Jong Chan Kim · Jangmin O · Byoung-Tak Zhang

Published online: 27 July 2007

Abstract The global minimum search problem is important in neural networks because the error cost involved is formed as multiminima potential in weight parametric space. Therefore, parameters that produce a global minimum in a cost function are the best values for enhancing the performance of neural networks. Previously, a global minimum search based on a damped oscillator equation known as the heavy ball with friction (HBF) was studied. The kinetic energy overcomes a local minimum if the kinetic energy is sufficiently large or else the heavy ball will converge into a local minimum due to the action of friction. However, an appropriate damping coefficient has not been found in the HBF; therefore, the ball has to be shot again after it arrives at each local minimum until it finds a global minimum. In order to solve this problem, we determined an adaptive damping coefficient using the geodesic of Newtonian dynamics Lagrangian. This geometric method produces a second-order adaptively damped oscillator equation, the damping coefficient of which is the negative time derivative of the logarithmic function of the cost potential. Furthermore, we obtained a novel adaptive steepest descent by discretizing this second-order equation. To investigate the performance of this novel steepest descent, we applied our first-order update rule to the Rosenbrock- and Griewank-type potentials. The results show that our method determined the global minimum in most cases from various initial points. Our adaptive steepest descent may be applied in

J. S. Kim · J. C. Kim
Department of Physics and Astronomy, Seoul National University, San 56-1, Shillim-Dong, Kwanak-Gu,
Seoul 151-747, Republic of Korea
e-mail: shick@phya.snu.ac.kr

J. C. Kim
e-mail: jckim@phya.snu.ac.kr

Jangmin O. · B.-T. Zhang (✉)
School of Computer Science and Engineering, Seoul National University, San 56-1, Shillim-Dong,
Kwanak-Gu, Seoul 151-744, Republic of Korea
e-mail: btzhang@bi.snu.ac.kr

Jangmin O.
e-mail: jmoh@bi.snu.ac.kr

many fields related to global minimum search, such as neural networks, game theory, and economics.

Keywords Geodesic · Adaptively damped oscillator · Adaptive steepest descent · Global minimum search · Heavy ball with friction

1 Introduction

An error cost function is usually a multiminima potential function in a parametric space [1]. In this study, a global minimum search is found to be essential for good performance of neural network training. Motivated by this, we devised a novel global search algorithm based on a geodesic of the Newtonian dynamics Lagrangian. A Lagrangian is the kinetic energy minus the potential energy whose time integral is the action in classical dynamics [2].

Attouch et al. [3] developed a heavy ball with friction (HBF) method in which the inertia of a heavy ball is used to overcome a local minimum if the ball has sufficient kinetic energy. If the minimum is global, the frictional force damps the motion of the ball and the ball converges to a global minimum. However, an appropriate damping coefficient has not been found [3]. As such, the ball is required to be restarted several times to complete the global minimum search. Therefore, finding an appropriate adaptive damping coefficient is crucial for good performance of the HBF algorithm in a global minimum search. Cabot [4] contributed to stabilizing the HBF algorithm with a regularization term. Despite this, the need for a well-controlled damping coefficient remains.

A hint of this problem was obtained from Qian's paper [5]. He presented the convergence condition of critically damped oscillators as a quadratic potential by using approximate linear dynamics. If we extend this critically damped oscillator to a multiminima potential, we can formulate a novel efficient global search algorithm. However, Qian's learning rule uses a constant damping coefficient such that the damping force always opposes the velocity. We required an adaptive descent to stop the ball at a global minimum and an adaptive ascent rule to move the ball above a local minimum if it is not a global minimum. Therefore, the problem was finding an appropriate damping coefficient that recognizes the global minimum.

After considerable research, we derived this adaptive damping coefficient from the concept of a geodesic, which is the shortest path connecting two points in a curved manifold. Edelman et al. [6] and Nishmori and Akaho [7] developed a geodesic formulation in the orthogonal constraint conditions and derived a learning rule based on orthogonal group theory. However, their problem was an image-processing problem, namely, independent component analysis (ICA) [7]. In contrast to their orthogonal group theoretical approach, we used this geodesic concept for Newtonian dynamics applicable to a heavy ball. Newton's second law is obtained by minimizing the action, which is the time integral of the classical Lagrangian. This is known as Hamilton's principle in classical mechanics [2]. However, our approach is to find the second-order equation by minimizing the time integral of the square root of the Newtonian dynamics Lagrangian. In this way, we obtained an adaptively damped second-order equation whose convergence to a global minimum is guaranteed by the singular crashing behavior of the heavy ball in an affine parameter coordinate. This affine parameter was introduced to define a new Lagrangian whose value is constant, as in ordinary differential geometry, to find the shortest curve connecting the two points on a sphere [8].

We discretized the second-order rule obtained to derive a first-order adaptive steepest descent using the procedure presented in Qian's paper [5]. Further, we attempted to solve Rosenbrock- and Griewank-type potentials that were used as examples in Attouch et al. [3].

Our adaptive learning rule exhibited good performance for global minimization in these problems and we anticipate that it will be applied to neural networks, economics, and game theory. Note that our rule is valid only when the global minimum value is zero. This condition is necessary to produce a singularity in an affine parameter coordinate to guarantee the convergence of a heavy ball.

We derive the adaptively damped oscillator equation in Sect. 2. Then, we prove the convergence of an adaptively damped oscillator in Sect. 3. In Sect. 4, we derive a novel adaptive steepest descent by discretizing a second-order equation. In Sect. 5, we apply our adaptive rule to a global minimum search. We present our conclusions in Sect. 6.

2 Derivation of an Adaptively Damped Oscillator

The Lagrangian for classical mechanics is defined as:

$$L_c = \frac{m}{2} \sum_{1 \le i \le n} \left(\frac{dw_i}{dt} \right)^2 - V, \tag{1}$$

where m is the mass of the heavy ball; w_i, the ith coordinate in a weight space; t, time; and n, the dimension of a weight space. V is an error cost and is referred to as the potential function in this paper. Solving the Euler–Lagrange equation using classical mechanics Lagrangian produces Newton's second law, which states that mass multiplied by acceleration is equal to the negative gradient of a potential [2].

However, this classical dynamics Lagrangian is not constant, as is that used for obtaining the geodesic of two points on a sphere [8]. Therefore, we are interested in defining a new Lagrangian whose value is a constant over time. This approach is similar to the geometrization of Newtonian dynamics [9] in defining the metric from the classical Newtonian Lagrangian. The detailed procedure is as:

$$\begin{aligned} ds^2 &= L_N d\sigma^2 \\ &= L_C dt^2 \\ &= -V dt^2 + \frac{m}{2} \sum_{1 \le i \le n} dw_i{}^2, \end{aligned} \tag{2}$$

where σ is an affine parameter; ds, an infinitesimal distance in a manifold; L_N, a new Lagrangian; and L_C, a classical mechanics Lagrangian. Based on the above relationship, we define L_N as:

$$L_N = \frac{1}{2} m \sum_{1 \le i \le n} \left(\frac{dw_i}{d\sigma} \right)^2 - V \left(\frac{dt}{d\sigma} \right)^2. \tag{3}$$

Our approach in defining a metric and obtaining a geodesic is similar to that of the procedure in general relativity theory [8,9] in that it considers both time and weight as equivalent variables. By solving the Euler–Lagrange equation of L_N, we obtain the following geodesic equations in the affine parameter coordinate. The Euler–Lagrange equations are:

$$-\frac{\partial L_N}{\partial t} + \frac{d}{d\sigma} \frac{\partial L_N}{\partial \dot{t}} = 0 \tag{4}$$

and

$$-\frac{\partial L_N}{\partial w_i} + \frac{d}{d\sigma} \frac{\partial L_N}{\partial \dot{w}_i} = 0, \tag{5}$$

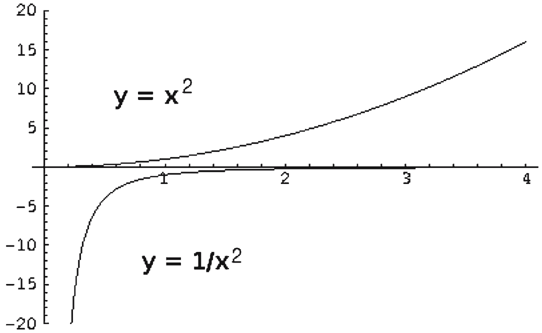

Fig. 1 Plot of potential V and negative inverse potential $-\frac{1}{V}$ when $V = x^2$

where the dot implies a derivative with respect to the affine parameter σ. Further, they yield the following geodesic equations:

$$\frac{dt}{d\sigma} = \frac{a}{V} \tag{6}$$

and

$$\frac{d^2 w_i}{d\sigma^2} = -\eta a^2 \frac{\partial}{\partial w_i}\left(-\frac{1}{V}\right), \tag{7}$$

where a is a constant and η is the learning rate, which is equal to the reciprocal of the mass, i.e., $\frac{1}{m}$. The above equations are in affine parameter coordinates; however, it is important to obtain the equation of motion in time t for the real motion of a heavy ball. To change the variable from σ to t, we use the chain rule:

$$\frac{dw_i}{d\sigma} = \frac{dt}{d\sigma}\frac{dw_i}{dt}, \tag{8}$$

$$\frac{d^2 w_i}{d\sigma^2} = \frac{dt}{d\sigma}\frac{d}{dt}\left(\frac{dt}{d\sigma}\right)\frac{dw_i}{dt} + \left(\frac{dt}{d\sigma}\right)^2 \frac{d^2 w_i}{dt^2}. \tag{9}$$

Therefore, from Eqs. 6, 7, 8, and 9 we obtain an adaptively damped oscillator equation in the time domain as:

$$\frac{d^2 w_i}{dt^2} = \frac{d}{dt}(\ln V)\frac{dw_i}{dt} - \frac{1}{m}\frac{\partial V}{\partial w_i}. \tag{10}$$

3 Proof of Convergence

We obtained two geodesic equations in terms of the affine parameter (7) and time (10). A geodesic equation in time t was derived from the geodesic equations in the affine parameter σ. We used the chain rule to change the independent variable from t to σ.

In Eq. 7, the left hand side represents the acceleration and the right hand side represents the force of a singular potential $-\frac{1}{V}$. In Fig. 1, we plot V and $-\frac{1}{V}$ when the potential $V = x^2$. Therefore, a mass will be accelerated toward a weight coordinate, which satisfies $V = 0$, with an infinite speed in an affine parameter coordinate because a negative inverse potential leads to negative infinity at weights satisfying $V = 0$. If this mass crash were observed in time, the mass would appear to accelerate toward an equilibrium weight coordinate and then

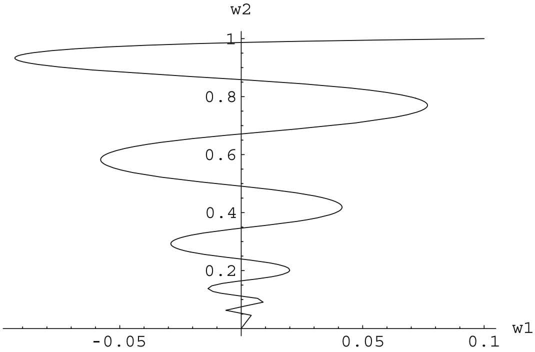

Fig. 2 A converging trajectory in the affine parameter coordinate for a quadratic potential

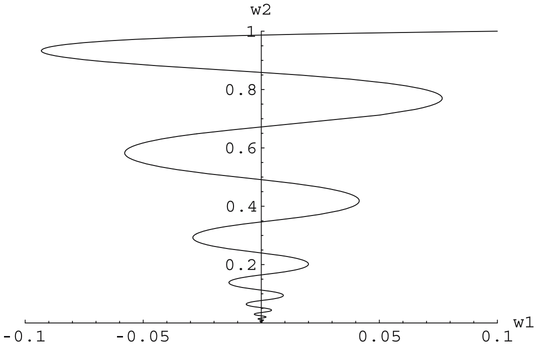

Fig. 3 A converging trajectory in time for a quadratic potential

be decelerated by an adaptive damping force represented by the first term on the right hand side of Eq. 10. Finally, if the minimum value of the potential were zero, then the heavy ball would stop at the minimum weight position. If the potential value were not zero, the adaptive damping coefficient would change sign; therefore, the motion of a heavy ball would change from descent to ascent and escape from the local minimum.

We can guarantee the convergence of the second-order learning equation of an affine parameter because of its singular crashing behavior toward the weight position satisfying $V = 0$. Because the trajectory is the same in both the affine parameter and the time coordinate, we can confirm that a time geodesic equation Eq. 10 converges to the global minimum.

The two trajectories are shown in Figs. 2 and 3. In both figures, we set the quadratic potential as $V = \frac{1}{2}\mathbf{w}^T H\mathbf{w}$, where H is a Hessian. The exact value of H is given as:

$$H = \begin{pmatrix} 10 & 0 \\ 0 & 0.1 \end{pmatrix}, \tag{11}$$

and $\mathbf{w} = (w_1, w_2)^T$. We set the learning rate $\eta = 0.01$, i.e., we set mass $m = 100$. In addition, we set the initial conditions to be $w_1(0) = 0.1$, $w_2(0) = 1$, and $\dot{w_1}(0) = \dot{w_2}(0) = 0$ in Eq. 7 and Eq. 10. In addition, we set $a = 1$ in Eq. 7 since a represents the proportional constant of an affine parameter σ. We observe a similar convergence pattern between the two trajectories in Figs. 2 and 3. The difference near the origin in the two figures arises from numerical error caused by the singularity at the origin in Eq. 7.

As a result, the mass always crashes to the weight coordinate in which $V = 0$ if its initial velocity is sufficiently small so as not to escape from the negative inverse potential $-\frac{1}{V}$.

4 Derivation of a Novel Adaptive Steepest Descent in a Discrete Time Case

In this section, we derive a novel steepest descent with adaptive momentum. We refer to the procedure in Qian's paper [5]. Qian derived the momentum term and the learning rate in quadratic potential for a critically damped oscillator. We discretize Eq. [3] as follows:

$$\frac{\mathbf{w}_{t+\Delta t} + \mathbf{w}_{t-\Delta t} - 2\mathbf{w}_t}{(\Delta t)^2} = \frac{\ln V_t - \ln V_{t-\Delta t}}{\Delta t}\frac{\mathbf{w}_{t+\Delta t} - \mathbf{w}_t}{\Delta t} - \frac{1}{m}\frac{\partial V}{\partial \mathbf{w}}. \tag{12}$$

After rearranging the terms, we arrive at the first-order update rule, which is:

$$\begin{aligned}\mathbf{w}_{t+\Delta t} - \mathbf{w}_t = & \frac{1}{1-(\ln V_t - \ln V_{t-\Delta t})}(\mathbf{w}_t - \mathbf{w}_{t-\Delta t}) \\ & -\frac{(\Delta t)^2}{1-(\ln V_t - \ln V_{t-\Delta t})}\frac{1}{m}\frac{\partial V}{\partial \mathbf{w}}.\end{aligned} \tag{13}$$

The above learning rule is an adaptive steepest descent rule and m is the mass of the particle, i.e., the reciprocal of learning rate η. We can compare our learning rule with that from Qian's paper [5] as follows:

$$\begin{aligned}\mathbf{w}_{t+\Delta t} - \mathbf{w}_t = & \frac{m}{m+\mu\Delta t}(\mathbf{w}_t - \mathbf{w}_{t-\Delta t}) \\ & -\frac{(\Delta t)^2}{m+\mu\Delta t}\frac{\partial V}{\partial \mathbf{w}},\end{aligned} \tag{14}$$

where μ is a damping constant; m, a particle's mass; and the constant before the gradient is the momentum parameter.

We can determine the fixed learning rate and momentum parameter in Eq. 14, which are given as:

$$\eta_f = \frac{(\Delta t)^2}{m+\mu\Delta t}, \tag{15}$$

$$p_f = \frac{m}{m+\mu\Delta t}, \tag{16}$$

where η_f is the learning rate and p_f is the momentum parameter in Qian's paper on a critically damped oscillator [5]. In addition, we can obtain adaptively the learning rate η_a and the momentum parameter p_a in Eq. 13, which are given by:

$$\eta_a = \frac{(\Delta t)^2}{1-(\ln V_t - \ln V_{t-\Delta t})}\frac{1}{m} \tag{17}$$

and

$$p_a = \frac{1}{1-(\ln V_t - \ln V_{t-\Delta t})}. \tag{18}$$

Note that Qian's steepest descent is optimal for a quadratic potential; however, our adaptive steepest descent is applicable for a multiminima potential with a zero global minimum value.

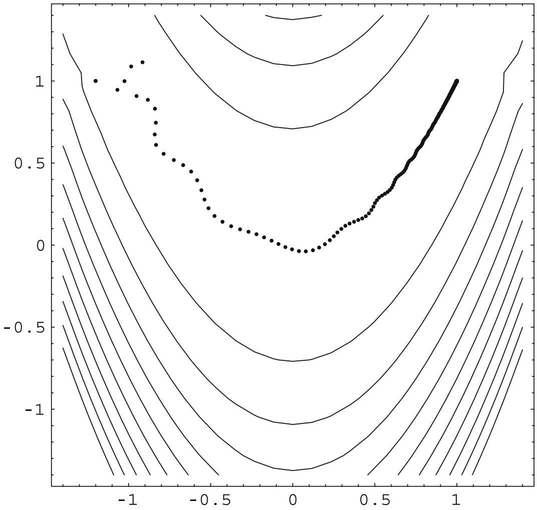

Fig. 4 A global minimum search for the Rosenbrock potential. The two initial points are $(-1.2, 1)$ and the global minimum is located at $(1, 1)$. We observe excellent convergence of a learning point to the global minimum. Particle mass $m = 1,000$

5 Global Minimum Search Problems

We applied the adaptive steepest descent obtained (13) to global minimum search problems. We took two examples from Attouch et al.'s paper [3] for a global minimum exploration. Mathematica 5.0 was used for the numerical calculations.

First, we applied our adaptive steepest descent rule (13) to the Rosenbrock potential. The exact form of the potential is as follows:

$$V(x, y) = 100(y - x^2)^2 + (1 - x)^2, \tag{19}$$

and the global minimum of the Rosenbrock potential is located at $(1, 1)$. We set the particle mass $m = 1,000$ and time increment $\Delta t = 1$. The initial two points were set at $(-1.2, 1)$. The global search result based on our rule (13) is presented in Fig. 4. We observe that the heavy ball arrives at the global minimum at $(1, 1)$ directly.

We observe that our method provides some control over the damping coefficient to obtain a rapid oscillation-free trajectory, a feature that was absent in Attouch et al. [3]. Their path depended on a damping coefficient. Higher values of the damping coefficient were observed to generate slow trajectories resembling those of steepest descent, whereas lower values generated rapid trajectories with oscillations that grew as the coefficient decreased.

Second, we checked the convergence of the heavy ball in a Griewank-type potential. The exact form of the potential is:

$$V(x, y) = (2x^2 + y^2 - xy)/50 - \cos(x)\cos(y/\sqrt{2}) + 1. \tag{20}$$

We simulated the global search with a particle mass $m = 100$ and time increment $\Delta t = 1$. The initial two points were $(10, -10)$, as shown in Fig. 5. We observe the direct convergence of a learning point to the global minimum. The potentials and initial points in Figs. 4 and 5 are the same as those in Attouch et al. [3].

In Fig. 6, we observe a global search for the Griewank-type potential with the initial two points at $(-6, -10)$. We observe that the ball escapes from two local minima located at

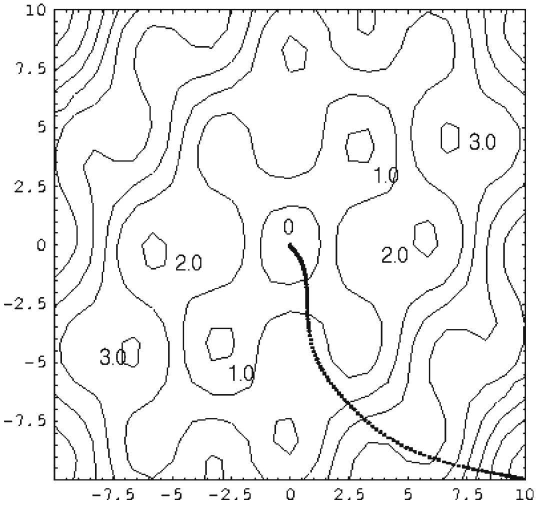

Fig. 5 A global exploration in the Griewank-type potential. The two initial points are (10, −10) and the global minimum is located at (0, 0). We observe the direct convergence of a learning point to the global minimum. Particle mass $m = 100$

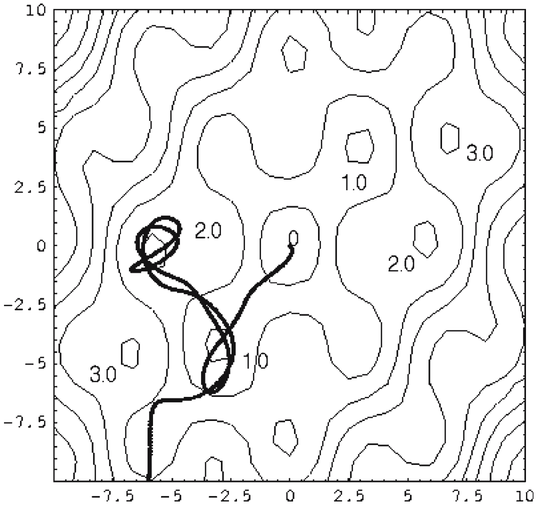

Fig. 6 A global exploration in Griewank-type potential. The two initial points are (−6, −10) and the global minimum is at (0, 0). We observe that a heavy ball escapes from two local minima, and then finally arrives at the global minimum. Particle mass $m = 100$

(−3, −5) and (−6, 0), then finally arrives at the global minimum at the origin. The inertia, adaptive descent, and adaptive ascent help the ball to escape from the local minima. Finally, the adaptive damping stops the ball at the origin, which is the global minimum.

In Attouch et al. [3], regardless of the initial conditions and the friction coefficient, the ball eventually rests at a critical point, which may be a local minimum. Therefore, the procedure was restarted using some other initial conditions together with the information gathered along

Table 1 Times elaped before the heavy ball arrives at the global minimum for 21 initial points in the Griewank-type potential (20). We observe that when the heavy ball starts from $(1, -10)$, the ball moves around a local minimum and cannot reach the global minimum. Particle mass $m = 100$

Initial point	$(-10, -10)$	$(-9, -10)$	$(-8, -10)$	$(-7, -10)$	$(-6, -10)$
Elapsed time	1,475	719	643	678	1,029
Initial point	$(-5, -10)$	$(-4, -10)$	$(-3, -10)$	$(-2, -10)$	$(-1, -10)$
Elapsed time	1,153	1,299	1,124	1,039	619
Initial point	$(0, -10)$	$(1, -10)$	$(2, -10)$	$(3, -10)$	$(4, -10)$
Elapsed time	1,163	Infinity	714	616	786
Initial point	$(5, -10)$	$(6, -10)$	$(7, -10)$	$(8, -10)$	$(9, -10)$
Elapsed time	329	322	743	1,269	1,141
Initial point	$(10, -10)$				
Elapsed time	634				

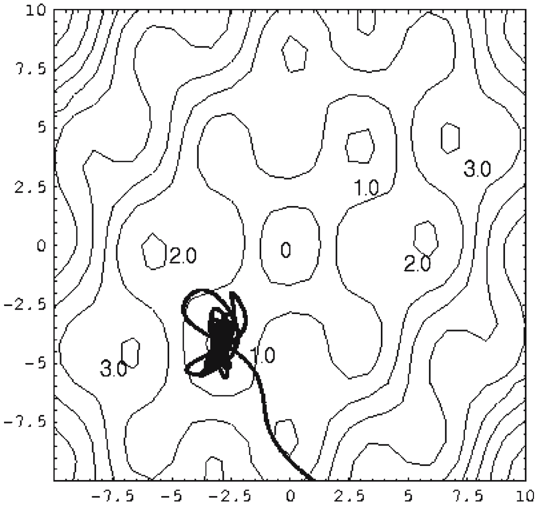

Fig. 7 Trajectory of a heavy ball with initial point $(1, -10)$. We observe the ball runs around a local minimum at $(-3, -4)$. Particle mass $m = 100$

the trajectory for global exploration. An advantage was that new trajectories could start from points on the trajectory that yielded a lower value of potential than the local minimum [3].

Note that our algorithm is valid only if the global minimum potential value is zero. This constraint guarantees singularity of the affine parameter coordinate for the negative inverse potential $-\frac{1}{V}$, as described in Sect. 3. If not, the trajectory does not stop, a phenomenon expected when the singularity disappears.

We measured the elapsed time when the heavy ball converges to the global minimum for 21 initial points; the results are shown in Table 1.

In Table 1, the elapsed time of the initial point $(1, -10)$ is infinity. The trajectory of the heavy ball in this case is shown in Fig. 7. We observe that the ball is moving around the local minimum $(-3, -4)$. To overcome this unfavorable behavior, we changed the particle's mass from 100 to 10,000, and observed the ball converge to the origin after 9,201 steps. The reason the heavy ball reached the global minimum from the same initial point is believed to be that the inertia of the ball increases with the mass and this large inertia helps the ball to escape from a local minimum against the attractive force toward the local minimum. Furthermore, the increased mass influences the adaptive damping force to produce a stronger ascent from a local minimum.

6 Conclusions

We have developed a novel algorithm for a global minimization based on the HBF method. Our geodesic of Newtonian dynamics Lagrangian produced a novel momentum parameter and learning rate. We confirmed the validity of our learning rule by applying it to Rosenbrock- and Griewank-type potentials.

The most significant finding of this research is that we devised a minimization rule for the global minimum search problem. The minimization rule for Newtonian dynamics is Hamilton's principle in which the time integral of the Newtonian dynamics Lagrangian is minimized. Furthermore, in our rule, the minimum search algorithm is obtained by minimizing the time integral of the square root of the Newtonian dynamics Lagrangian. The difference between the Newtonian dynamics rule and our optimization rule is whether we use the Newtonian dynamics Lagrangian or its square root.

The second important finding is that an appropriate damping coefficient for a damped oscillator equation for global minimization was determined. We propose the negative time derivative of the logarithmic function of the error cost as an adaptive damping coefficient. Finding this coefficient has been researched since the study by Attouch et al. [4].

Finally, we can interpret the meaning of this work by implementing artificial intelligence based on the Newtonian dynamics rule. This mechanical modeling provides us with a clear and easy understanding of the optimization by using the physical concept of the HBF. This understanding is useful in simplifying a complex problem and suggesting a most effective solution to this simplified problem using mechanical optimization. The applicable cost functions will be those of economics, game theory, and neural networks with a constraint on the potential; namely, the global minimum value must be zero.

Future work aims to devise a novel adaptive algorithm that works without these constraints. This includes developing a principle to change the potential incrementally so that it can have a singularity for future moves.

Acknowledgments We thank Bumseok Kyae, Seong Chan Park, Yung-Kyun Noh, and Ji-Woo Lee for their fruitful discussions. This work was supported by the BK21 Program of the Ministry of Education and Human Resources Development and the NRL program of the Ministry of Science and Technology of the Korean government.

References

1. Cho S-Y, Chow TWS (1999) Training multilayer neural networks using fast global learning algorithm—least-squares and penalized optimization methods. Neurocomputing 25:115–131
2. Marion JB, Thornton ST (1995) Classical dynamics of particles and systems. Saunders College Pub., Fort Worth
3. Attouch H, Goudou X, Redont P (2000) The heavy ball with friction method, I. The continuous dynamical system: global exploration of the local minima of a real-valued function by asymptotic analysis of a dissipative dynamical system. Commun Contemp Math 2(1):1–34
4. Cabot A (2004) Inertia gradient-like dynamics system controlled by a stabilizing term. J Optim Theory Appl 120(2):275–303
5. Qian N (1999) On the momentum term in gradient descent learning algorithms. Neural Netw 12(1):145–151

6. Edelman A, Arias TA, Smith ST (1998) The geometry of algorithms with orthogonality constraints. Siam J Matrix Anal Appl 20(2):303–353
7. Nishimori Y, Akaho S (2005) Learning algorithms utilizing quasi-geodesic flows on the Stiefel manifold. Neurocomputing 67:106–135
8. Martin JL (1995) General relativity: a first course for physicists. Prentice Hall Europe, Hertfordshire
9. Caiani L, Casetti L, Clementi C, Pettini M (1997) Geometry of dynamics, Lyapunov exponents, and phase transition. Phys Rev Lett 79:4361–4364