Received: 7 September 2021 | Revised: 16 March 2022 | Accepted: 29 March 2022

DOI: 10.4218/etrij.2021-0320

ORIGINAL ARTICLE

ETRI Journal WILEY

Nonlinear optimization algorithm using monotonically increasing quantization resolution

Jinwuk Seok | **Jeong-Si Kim**

Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea

Correspondence
Jinwuk Seok, Artificial Intelligence Research Laboratory, Electronics and Telecommunications Research Institute, Daejeon, Republic of Korea.
Email: jnwseok@etri.re.kr

Funding information
Institute for Information and Communications Technology Promotion, Grant/Award Numbers: 2017-0-00142, 2021-0-00766

Abstract

We propose a quantized gradient search algorithm that can achieve global optimization by monotonically reducing the quantization step with respect to time when quantization is composed of integer or fixed-point fractional values applied to an optimization algorithm. According to the white noise hypothesis states, a quantization step is sufficiently small and the quantization is well defined, the round-off error caused by quantization can be regarded as a random variable with identically independent distribution. Thus, we rewrite the searching equation based on a gradient descent as a stochastic differential equation and obtain the monotonically decreasing rate of the quantization step, enabling the global optimization by stochastic analysis for deriving an objective function. Consequently, when the search equation is quantized by a monotonically decreasing quantization step, which suitably reduces the round-off error, we can derive the searching algorithm evolving from an optimization algorithm. Numerical simulations indicate that due to the property of quantization-based global optimization, the proposed algorithm shows better optimization performance on a search space to each iteration than the conventional algorithm with a higher success rate and fewer iterations.

KEYWORDS
nonlinear optimization, quantization, stochastic gradient descent

1 | INTRODUCTION

Recently, as artificial intelligence based on deep neural networks has shown successful results in several fields, many researchers have vigorously investigated learning algorithms induced by nonlinear optimization. In particular, in contrast to other applications, such as image recognition based on big data using algorithms based on stochastic gradient descent, many applications of artificial intelligence still use conventional optimization algorithms directly. Reinforcement learning is one prominent example of such applications.

Another viewpoint in contrast to the researching field is the requirement of artificial intelligence in tiny electronic devices (such as cell phones), innovative sportswear, and industrial control systems. Such hardwares typically have a small amount of computing power for minimizing power consumption. Nevertheless, users demand artificial intelligence that allows them to operate on such tiny computing devices. For developing such

artificial intelligence systems, a fast optimization algorithm under small computing power is required [1–3].

A quantized optimization algorithm is one of the methodologies satisfying such requirements, and we expect it to maintain optimization performance with comparably limited computing power. The Hogwild-style algorithm [4] fundamentally analyze the quantized machine learning based on an optimization theory. As the most extremely quantized algorithm, [5,6] provided the 1-bit quantization satisfying a strong convergence condition. In terms of increasing communication effectiveness of large-scale communication units, [7] proposed a quantized learning scheme considering the tradeoff between communication, computational efficiency, and convex optimization. However, as recent studies [8–10] report challenges in using optimization algorithms based on the SGD for machine learning, it is necessary to verify the consistency of the conventional quantized optimization algorithm.

Aiming at above mentioned challenges, we define a quantization error and propose a quantized optimization algorithm with weak convergence under the assumption that the defined quantization error satisfies the identically independent distribution condition. Toward this, we analyze the convergence condition of the proposed quantized optimization algorithm based on the stochastic differential equation and transition probability. In this analysis, we derive appropriate quantization conditions that ensure the convergence of the proposed algorithm using a schedule function for quantization. Further, we establish that the provided quantization condition ensures a global optimization property on a search space for each iteration in the viewpoint of distributional/weak convergence. Through numerical experiments, we show that the performance of the proposed algorithm is superior to conventional search algorithms.

2 | DEFINITION AND ANALYSIS OF QUANTIZATION

Before beginning discussion, we establish the following definitions and assumptions:

Definition 1. For $x \in \mathbf{R}$, the round-off for extraction of integral part is

$$x^Q \equiv \lfloor x \rfloor + \epsilon \ (\epsilon \in \mathbf{R}[0,1)), \tag{1}$$

where $x^Q \in \mathbf{Z}$ is an integral part of the real number x.

Definition 2. The greatest integer function or the Gauss function $[\cdot]$ is defined as follows:

$$[x] \equiv \lfloor x + 0.5 \rfloor = x + 0.5 - \epsilon \triangleq x + \varepsilon, \tag{2}$$

where $\varepsilon \in \mathbf{R}(-0.5, 0.5]$ is round-off error.

Assumption 1. For an objective function $f : \mathbf{R}^n \to \mathbf{R}$ with $f(x) \in C^2$, there exists a positive value $m \in \mathbf{R}$ such that

$$m \triangleq \inf_{\forall x, v \in \mathbf{R}^n} \frac{1}{\|v\|^2} \cdot \left\langle v, \frac{\partial^2 f}{\partial x^2} v \right\rangle, \tag{3}$$

for all $x, v \in \mathbf{R}^n$.

2.1 | Fundamental definition of quantization

The aim of this study is to extend a quantization error to a temporal perspective. In other words, in the initial stage, the proposed algorithm searches the optimal point at a low resolution, and over time, it searches the optimum point at a high resolution. Because decreasing the variance of the quantization error to the operation time ensures an increase in the resolution of quantization, we can control the resolution of quantization by the time-dependent variance and expect better optimization performance. For this reason, it is necessary to redefine the fundamental definition of quantization according to the Definitions 1 and 2, as follows:

$$x^Q \triangleq \frac{1}{Q_p} \lfloor Q_p \cdot (x + 0.5 \cdot Q_p^{-1}) \rfloor = \frac{1}{Q_p} [Q_p \cdot x] \in \mathbf{Q}. \tag{4}$$

Rewriting (4) with the definition of quantization, we obtain the following formulation, including the quantization error.

$$x^Q = \frac{1}{Q_p} [Q_p \cdot x] = \frac{1}{Q_p} \left(Q_p \cdot x + \varepsilon \right) = x + \varepsilon Q_p^{-1}. \tag{5}$$

In (5), we replace the constant quantization parameter Q_p with a monotonically increasing quantization parameter with respect to time t, such as $Q_p(t)$. Thus, we obtain the quantization error term as a monotonically decreasing function for time t.

In addition, Jimenes and others [11] proved that the quantization error is a white noise through asymptotic analysis, if the quantization error is an asymptotically pairwise independent and uniformly distributed within the error bound. Intuitively, quantization must be uniform for the quantization error to follow a uniform distribution. Therefore, we assume a uniform quantization representing equal quantized resolution for all $x \in \mathbf{R}$ at the same time $t \in \mathbf{R}$ without changing the quantized spatial resolution. Moreover, from the engineering perspective and following the binary number system, we define the quantization parameter as follows:

$$Q_p = \eta \cdot b^n \quad \eta \in \mathbf{Z}^+,\ \eta < b, \tag{6}$$

where the base b is $b \in \mathbf{Z}^+$, $b \geq 2$. Under the above assumptions and the proposition provided by Jimenes and others [11], we employ the following theorem, known as white noise hypothesis (WNH), without proof.

Proposition 1. If the quantization of x is uniform quantization at t with the quantization parameter defined as (4) and (6), the quantization error $\varepsilon Q_p(t) = x^Q - x$ is white noise.

Consequently, when the quantization parameter is a monotonically decreasing function with respect to time, the quantization error is white noise with the monotonically decreasing variance with respect to time. Furthermore, if the quantization error of a parameter vector ensures WNH, we consider the following independent assumption.

Assumption 2. For a $x \in \mathbf{R}^n$, and $x^Q \in \mathbf{R}^n$, we assume that the components of the quantization error $\vec{\varepsilon} Q_p = x^Q - x = \{\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_{n-1}\} \in \mathbf{R}^n$ are independent.

2.2 | Search equation providing the quantized parameter vector

First, we define the parameter vector $w_t \in \mathbf{R}^n$, a descent direction $h(w_t) \in \mathbf{R}^n$, and a search equation as follows:

$$w_{t+1} = w_t - \lambda_t \cdot h(w_t), \tag{7}$$

where $\lambda_t \in \mathbf{R}(0,1)$ is a step size such that $\lambda_t = \text{argmin}_{\lambda_t \in \mathbf{R}(0,1)} f(w_t - \lambda_t \cdot h(w_t))$. Suppose that the parameter vector of the current step w_t and next step w_{t+1} are quantized, we have

$$w_{t+1}^Q = \left(w_t^Q - \lambda_t \cdot h^Q(w_t)\right)^Q = w_t^Q - \left(\lambda_t \cdot h^Q(w_t)\right)^Q. \tag{8}$$

In (8), we set $g(x) \equiv \lambda_t \cdot h(x)$ and quantize it according to (5). Then, the quantized $g(x)$ is

$$g(x)^Q = \frac{1}{Q_p} \lfloor Q_p(g(x) + \vec{\varepsilon} Q_p^{-1}) \rfloor = \frac{1}{Q_p} \cdot Q_p g(x) + \varepsilon_t Q_p^{-1}, \tag{9}$$

where ε is the vector-valued quantization error so that the distribution of components are independent distribution defined $\varepsilon \in \mathbf{R}^n$.

If there exists a rational number $\alpha_t \in \mathbf{Q}(0, Q_p)$ to factorize $g(x)$ such that $g(x) = \alpha_t \overline{h}(x)$, we have

$$g(x)^Q = \frac{\alpha_t}{Q_p} \cdot Q_p \overline{h}(x) + \varepsilon_t Q_p^{-1}. \tag{10}$$

Replacing $\overline{h}(x)$ with $h(x)$ and substituting (10) in (8), the quantized search equation is obtained as follows:

$$\begin{aligned} w_{t+1}^Q &= w_t^Q - \frac{\alpha_t}{Q_p} \cdot Q_p \cdot h^Q(w_t) + \varepsilon_t Q_p^{-1} \\ &= w_t^Q - \frac{\alpha_t}{Q_p} [Q_p \cdot h^Q(w_t)] \quad \because \alpha_t \in \mathbf{Q}. \end{aligned}$$

Therefore, we can use mathematical induction to obtain the search equation that provides the quantized parameter vector for all steps $t \in \mathbf{N}$. In (11), if we set α_t and Q_p appropriately to the rational number system based on (6), we have the search equation suitable to a general hardware based on the binary system, as follows:

$$w_{t+1}^Q = w_t^Q - 2^{-(n-k)} \nabla f^Q(w_t),\ n,k \in \mathbf{Z}^+,\ n > k. \tag{11}$$

In (10), assuming that each component of $\vec{\varepsilon}_t \in \mathbf{R}^n$ is equal to the round-off error and the quantization error follows a uniform distribution, the deviation of the quantization error is given as follows:

$$\forall \vec{\varepsilon}_t \in \mathbf{R}^n, \mathbb{E} Q_p^{-2} \vec{\varepsilon}_t^2 = \mathbb{E} Q_p^{-2} \cdot tr(\vec{\varepsilon}_t \vec{\varepsilon}_t^T) = \frac{1}{12 \cdot Q_p^2} \cdot n. \tag{12}$$

For all $t > 0$, $t \in \mathbf{R}$, when the deviation of the quantization error is equal to (12) and a standard Wiener process $dB_t \in \mathbf{R}$, we set $\varepsilon_t Q_p^{-1} ds = q \cdot dB_t$. Then,

$$\mathbb{E} \varepsilon_t^2 Q_p^{-2} ds = \mathbb{E} q^2 dB_t^2 = q^2 ds \Rightarrow q = \sqrt{\frac{1}{12}} \cdot Q_p^{-1}. \tag{13}$$

Similarly, we set a vector-valued Wiener proces $d\vec{B}_s = \vec{\varepsilon} ds \in \mathbf{R}^n$ and $\vec{\varepsilon}_t Q_p^{-1} ds = q \cdot d\vec{B}_t$. Then, the deviation of the quantization error q is evaluated as $q = \sqrt{n/12} \cdot Q_p^{-1}$. When it holds, if we regard the deviation of the quantization error as the function depending on time, the only parameter that can be varied to the time index t is the quantization parameter Q_p. Therefore, we define it as a function of time, as follows:

$$\sigma(t) = \frac{n}{24} \cdot Q_p^{-2}(t). \tag{14}$$

Consequently, because we can regard the quantized weight vector $w_t^Q \in \mathbf{R}^n$ as a stochastic process $\{W_t\}_{t=0}^{\infty}$, the search (11) can be rewritten as a stochastic differential equation as follows:

$$\begin{aligned} dW_s &= -\lambda_t \cdot h(W_s) ds + \vec{\varepsilon}_s Q_p^{-1}(s) ds \\ &= -\lambda_t \cdot h(W_s) ds + \sqrt{\frac{n}{12}} Q_p^{-1}(s) d\vec{B}_s \\ &= -\lambda_t \cdot h(W_s) ds + \sqrt{2\sigma(s)} \cdot d\vec{B}_s. \end{aligned} \tag{15}$$

When the search algorithm is given as (15), the transition probability of the weight vector weakly converges to the following Gibb's distribution under suitable conditions [12]. Moreover, when the deviation of Gibb's distribution is a monotonically decreased to zero, that is, $\sigma(t) \to 0$, the transition probability of the weight vector converges to the global minima of the objective function $f(W_t)$ [13]. It means that

$$\lim_{t \uparrow \infty} \sigma(t) = \frac{n}{24} \cdot \lim_{t \uparrow \infty} Q_p^{-2}(t) = 0. \tag{16}$$

Equation (16) illustrates that with monotonically decreasing deviation of the quantization error $\sigma(t)$, the quantization parameter increases monotonically and the resolution of quantization increases with time. Consequently, we claim that it is possible to find the global minima or the best optimal point on a finite domain with increasing resolution of quantization by increasing the quantization parameter Q_p under a suitable schedule depending on the time index from a low resolution caused by a low Q_p. In addition, we propose a feasible scheduling function for the resolution of quantization, given by the following theorem:

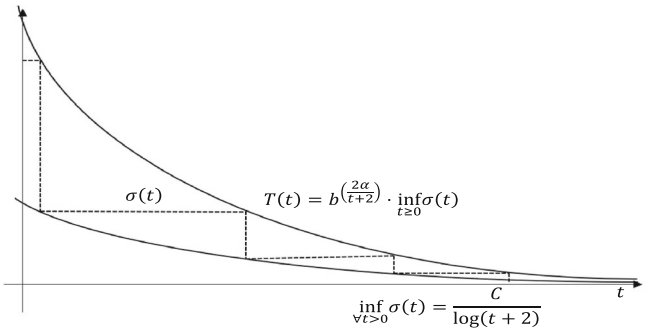

FIGURE 1 Conceptual diagram of quantization schedule

TABLE 1 The benchmark functions and corresponding difficulty scores for finding the global optimum

Benchmark Function	Equation	Known difficulty score
Ackley	$f(\mathbf{x}) = -a \cdot \exp\left(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d} \cos(c \cdot x_i)\right) + a + \exp(1)$	48.25
Whitley	$f(\mathbf{x}) = 1 + \frac{1}{4000}\sum_{i=1}^{n} x_i^2 - \prod_{i=1}^{n} \cos\left(\frac{x_i}{\sqrt{i}}\right)$	4.92
Rosenbrock 2D	$f(\mathbf{x}) = \sum_{i=1}^{d-1} [b(x_{i+1} - x_i^2)^2 + (a - x_i)^2]$, $\mathbf{x} \in \mathbf{R}^2$	44.17
Rosenbrock 100D	$f(\mathbf{x}) = \sum_{i=1}^{d-1} [b(x_{i+1} - x_i^2)^2 + (a - x_i)^2]$, $\mathbf{x} \in \mathbf{R}^{100}$	None
EggHolder	$f(x,y) = 977 - (y+47)\sin\left(\sqrt{\lvert y + 0.5y + 47\rvert}\right) - x\sin\left(\sqrt{\lvert x - (y+47)\rvert}\right)$	18.92
Xin-She Yang N.4	$f(\mathbf{x}) = 2.0 + \left(\sum_{i=1}^{d} \sin^2(x_i) - \exp\left(-\sum_{i=1}^{d} x_i^2\right)\right)\exp\left(-\sum_{i=1}^{d} \sin^2\sqrt{\lvert x_i\rvert}\right)$	26.33
Rosenbrock Modification	$f(\mathbf{x}) = 74 + 100(x_2 - x_1^2)^2 + (1 - x_1)^2 - 400e^{-\frac{(x_1+1)^2 + (x_2+1)^2}{0.1}}$	8.42
Salomon	$f(\mathbf{x}) = 1 - \cos\left(2\pi\sqrt{\sum_{i=1}^{d} x_i^2}\right) + 0.1\sqrt{\sum_{i=1}^{d} x_i^2}$	10.33
Drop-Wave	$f(x,y) = 1 - \frac{1 + \cos\left(12\sqrt{x^2+y^2}\right)}{(0.5(x^2+y^2)+2)}$	21.25
Powell D4	$f(\mathbf{x}) = \sum_{i=1}^{d} \lvert x_i\rvert^{i+1}$	32.58
Schaffel N. 2	$f(x,y) = 0.5 + \frac{\sin^2(x^2 - y^2) - 0.5}{(1 + 0.001(x^2+y^2))^2}$	39.58

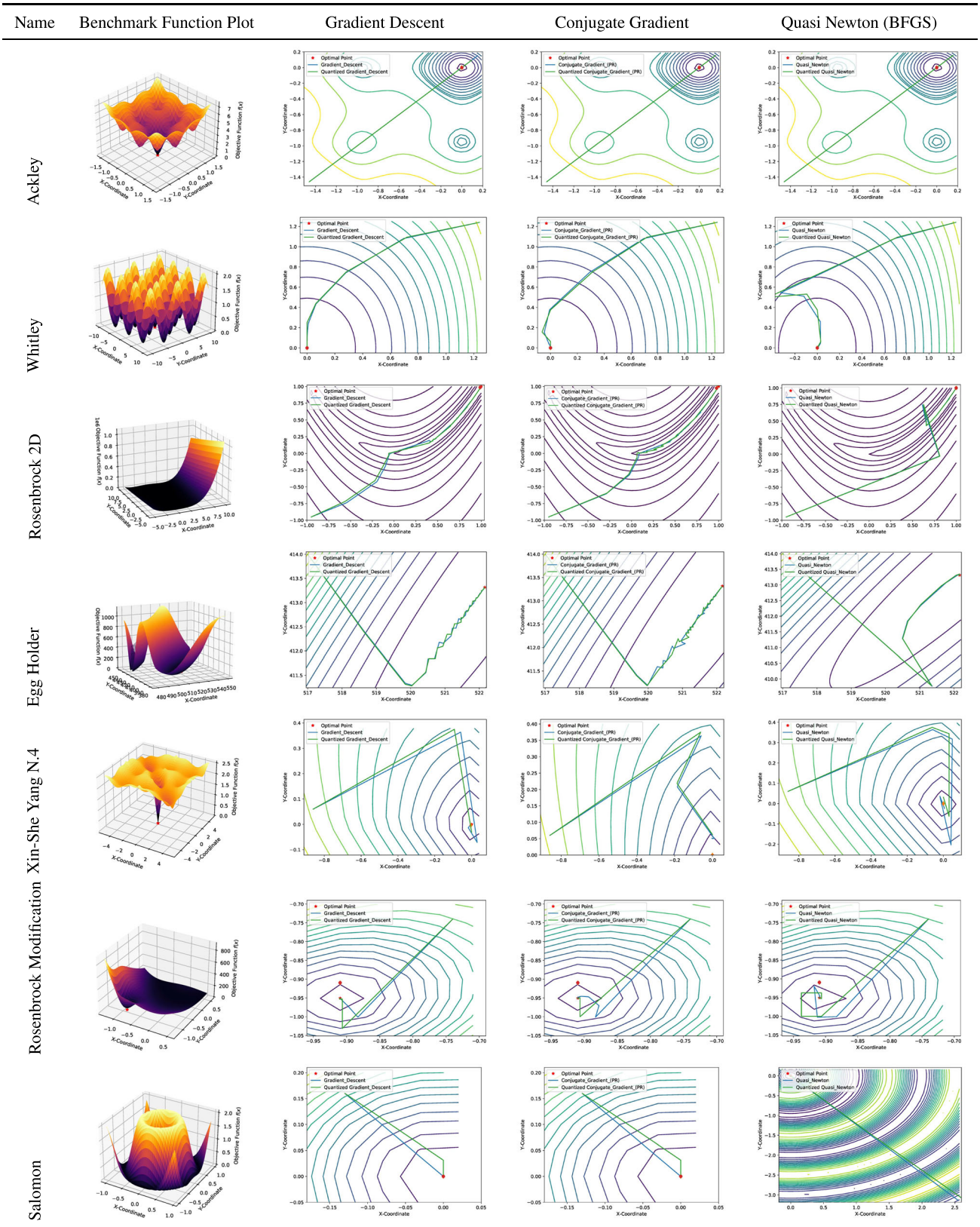

FIGURE 2 Results of the successful trace to each algorithm for benchmark functions. For a fair comparison, we select search results of a conventional and the proposed algorithms, both being successful in finding an optimal point. The blue line represents searching traces of conventional algorithm and the green line represents those of proposed algorithm. Some conventional searching traces represent a straight line and those look like the results of a single iteration. However, those took a number of iterations on the straight-like line. Alternatively, the proposed algorithm with bent lines took fewer iterations than the conventional algorithm

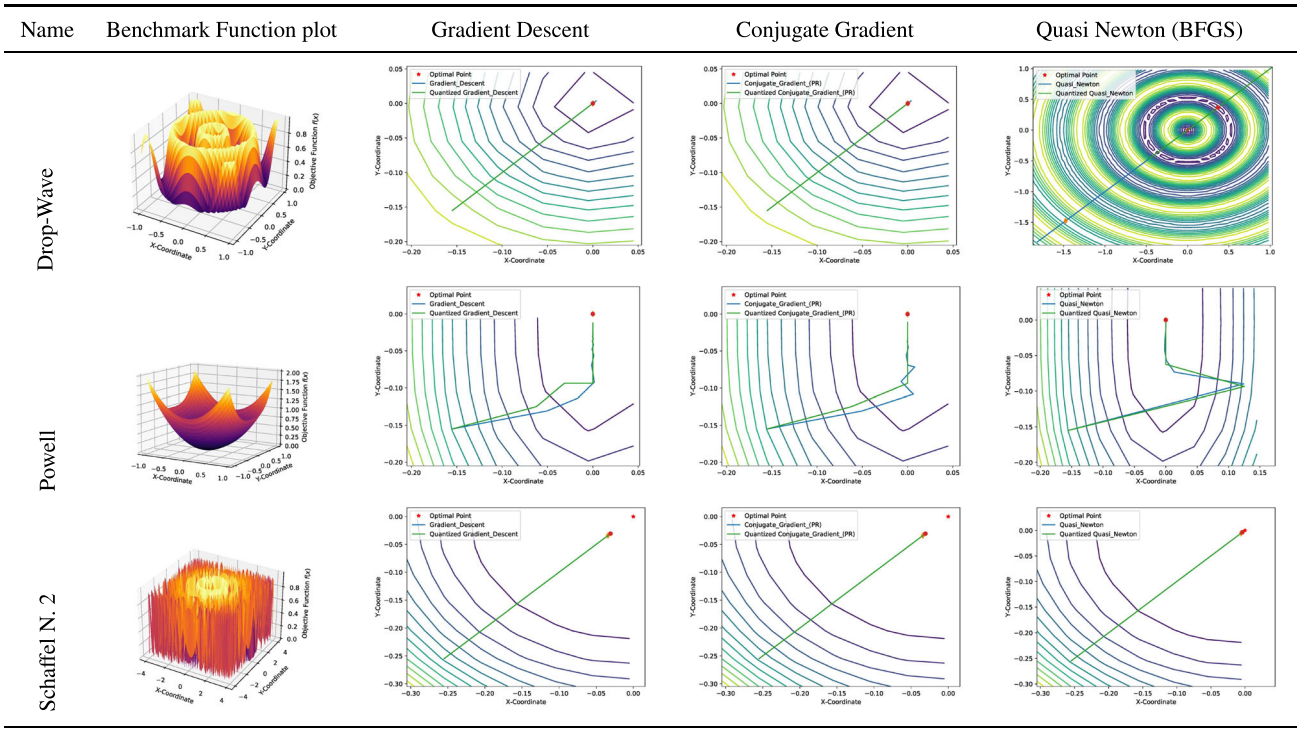

FIGURE 2 (Continued)

Theorem 1. If the search equation satisfies (15) and (16), the stochastic process $\{W_t\}_{t=0}^{\infty}$ generated by the search equation weakly converges to the global minimum when the deviation of the quantization error is given as follows:

$$\inf_{t\geq 0}\sigma(t)=\frac{C}{\log(t+2)},\ C\in\mathbf{R},C>>0, \tag{17}$$

where C is a hyperparameter representing the largest difference in the objective function, such as $\sup_{w_t\in D}f(w_t)-\min_{w_t\in D}f(w_t)$, on a domain D defined as $w_t\in D\subset\mathbf{R}^n$.

3 | DECISION SCHEME OF THE QUANTIZATION PARAMETER

Although we can use the appropriate scheduler provided in Theorem 1 for finding the global minimum, it is impossible to apply this scheduler directly to the quantized search algorithm. Because the deviation $\sigma(t)$ is a function proportional to $Q_p(t)$, which is an integer value, the deviation evaluated by (17) is not coincidental with the quantized search equation.

However, from the result of theorem 1, the optimal deviation $\sigma(t)$ should satisfy $\sigma(t)\geq\inf\sigma(t)\triangleq c/\log(2+t)$. Besides, it is a monotonically decreasing function. In addition, we let a supremum of $\sigma(t)$, satisfying the conditions for quantization (6) and (14), as another monotonically decreasing function $T(t)$ such that

$$\frac{C}{\log(t+2)}\leq\sigma(t)\leq T(t), \tag{18}$$

where $T(t)$ is a monotonically decreasing function, such as $T(t)\downarrow 0$ with respect to $t\uparrow\infty$. Moreover, when Δ is

TABLE 2 Input domain (represented with min and max per components) and optimal point corresponding benchmark functions

Benchmark function	Input domain	Optimal point
Ackley	$[-32,32]$	$[0,0]$
Whitley	$[-512,512]$	$[0,0]$
Rosenbrock 2D	$[-5,10]$	$[1,1]$
Rosenbrock 100D	$[-5,10]$	$[1,1,\ldots,1]\in\mathbf{R}^{100}$
EggHolder	$[400,600],[300,500]$	$[522.16,413.31]$
Xin-She Yang N.4	$[-5,5]$	$[0,\ldots,0]\in\mathbf{R}^4$
Rosenbrock Modification	$[-1.3,0.6]$	$[-0.91,-0.95]$
Salomon	$[-1,1]$	$[0,0]$
Drop-Wave	$[-1.0,1.0]$	$[0,0]$
Powell D4	$[-1,1]$	$[0,\ldots,0]\in\mathbf{R}^4$
Schaffel N. 2	$[-4,4]$	$[0,1.25]$

given as $\Delta \equiv \sup_{x,y \in \mathbf{R}^n}(f(x)-f(y))$, $T(t)$ has the following properties:

$$\frac{d}{dt}e^{-\frac{2\Delta}{T(t)}} = \frac{dT(t)}{dt}\cdot\frac{1}{T^2(t)}e^{-\frac{2\Delta}{T(t)}} \to 0 \text{ as } t \uparrow \infty. \quad (19)$$

Therefore, we obtain the schedule function $\sigma(t) \in \mathbf{Q}$, which has the infimum as the optimal schedule function $\inf \sigma(t)$ and the supremum as the monotonically decreasing function for quantization $T(t)$. The schedule function $\sigma(t)$ shown in Figure 1 satisfies (17).

TABLE 3 Experimental results of benchmark functions with three gradient-based search algorithms using line search

		Conventional algorithm		Proposed algorithm			
Benchmark Function	**Algorithm**	**Final step**	**Success**	**Final step**	**Success**	**Improvement ratio of steps**	**Improvement of succ. ratio**
Ackley	Gradient descent	8	43.0	3	49.0	62.50	13.95
	Conjugate gradient	16	35.0	4	32.0	75.00	−8.57
	Quasi newton (BFGS)	23	34.0	6	23.0	73.91	−32.35
Whitley	Gradient descent	13	54.0	12	54.0	7.69	0.00
	Conjugate gradient	9	53.0	7	53.0	22.22	0.00
	Quasi newton (BFGS)	6	26.0	6	30.0	0.00	15.38
Rosenbrock 2D	Gradient descent	3182	100.0	2427	95.0	23.73	−5.00
	Conjugate gradient	1601	83.0	1220	81.0	23.80	−2.41
	Quasi newton (BFGS)	47	87.0	49	89.0	-4.26	2.30
Rosenbrock 100D	Gradient descent	6845	82.0	2685	80.0	60.77	−2.44
	Conjugate gradient	4144	76.0	1262	76.0	69.55	0.00
	Quasi newton (BFGS)	839	76.0	77	80.0	90.82	5.26
EggHolder	Gradient descent	85	48.0	78	48.0	8.24	0.00
	Conjugate gradient	113	32.0	111	33.0	1.77	3.13
	Quasi newton (BFGS)	9	34.0	9	37.0	0.00	8.82
Xin-She Yang N.4	Gradient descent	17	3.0	10	32.0	41.18	966.67
	Conjugate gradient	0	0.0	8	41.0	inf	inf
	Quasi newton (BFGS)	17	7.0	4	26.0	76.47	271.43
Rosenbrock Modification	Gradient descent	7	11.0	8	12.0	−14.29	9.09
	Conjugate gradient	40	23.0	46	30.0	−15.00	30.43
	Quasi newton (BFGS)	7	8.0	10	8.0	−42.86	0.00
Salomon	Gradient descent	6	18.0	1	17.0	83.33	−5.56
	Conjugate gradient	5	18.0	1	18.0	80.00	0.00
	Quasi newton (BFGS)	6	5.0	2	4.0	60.00	−20.00
Drop-Wave	Gradient descent	5	5.0	2	4.0	60.00	−20.00
	Conjugate gradient	5	4.0	1	4.0	80.00	0.00
	Quasi newton (BFGS)	4	9.0	1	7.0	75.00	−22.22
Powell D4	Gradient descent	70	100.0	69	100.0	1.43	0.00
	Conjugate gradient	68	100.0	65	100.0	4.41	0.00
	Quasi newton (BFGS)	15	100.0	14	100.0	6.67	0.00
Schaffel N. 2	Gradient descent	9	66.0	10	63.0	−11.11	−4.55
	Conjugate gradient	10	58.0	11	58.0	−10.00	0.00
	Quasi newton (BFGS)	6	64.0	7	58.0	−16.67	−9.37
Average			44.30		46.73	32.48	0.18

To complete the quantized learning process, we define the quantization parameter $Q_p(t)$ using a monotonically decreasing function $\overline{h}(t) \in \mathbf{Z}^+$ with respect to t, as follows:

$$Q_p(t) = \eta \cdot b^{\overline{h}(t)}, \text{ such that } \overline{h}(t) \uparrow \infty \text{ as } t \uparrow \infty. \tag{20}$$

By virtue of (14), (18), and (20), the function of power parameter $\overline{h}(t)$ contains the following supremum and infimum:

$$\frac{1}{2}\log_b\left(\frac{n}{24\cdot\eta^2}\cdot T(t)^{-1}\right) \le \overline{h}(t) \le \frac{1}{2}\log_b\left(\frac{n\log(t+2)}{24\cdot\eta^2\cdot C}\right). \tag{21}$$

To illustrate (21) briefly, we let $n = 24$, $\eta = 1$, and the total number of datum to be 10^b. In addition, we assume that the time index t corresponds to one data. From the time index assumption, because t is equal to $\tau \cdot 10^b$ for an arbitrary epoch τ, the supremum of $h(t)$ is as follows:

$$\overline{h}(t) \le \frac{1}{2}\log_b C^{-1}(b + \log\tau) + \delta(t), \tag{22}$$

where $\delta(t) \in \mathbf{R}$ is an error corresponding to a remaining term such that $\delta(\tau) = \log_b[1 + \varepsilon/(b+\log\tau)]^{1/2}$. If the objective function is a probability distribution function, the maximum value of C is 1 for $\mathbf{R}^n$, and every 10^b epoch increases 1 bit of the resolution systematically.

In addition, for the quantized learning in a k-bit integer system, we can determine that the least significant bit serves as the optimal point in the range associated with $C < b^{-k}$. Therefore, we design the virtual infimum of $\overline{h}(t)$ satisfying quantized learning to be achieved on a low resolution for $\mathbf{R}^n$ and sufficiently large C at an initial stage. Moreover, for sufficiently small C, $\overline{h}(t)$ caused quantized learning to be processed on a high resolution near a feasible optimum after some time.

For instance, we suggest $T(t)$ satisfying (16) for the infimum of $\overline{h}(t)$, as follows:

$$T(t) = b^{\left(\frac{2\beta}{t+2}\right)} \cdot \inf_{t\ge 0}\sigma(t). \tag{23}$$

By (23), we can obtain the infimum of $\overline{h}(t)$ such that

$$\begin{aligned}\overline{h}(t) \ge \frac{1}{2}\log_b\left(\frac{n}{24\cdot\eta^2}\cdot T(t)^{-1}\right) &= -\frac{\beta}{t+2} + \sup_{t\ge 0}\overline{h}(t),\\ \because \sup_{t\ge 0}\overline{h}(t) &= \frac{1}{2}\log_b\left(\frac{n\log(t+2)}{24\cdot\eta^2\cdot C}\right),\end{aligned} \tag{24}$$

where α is a proportional constant, which controls the increasing speed of $\overline{h}(t)$ as t increases. Because $\overline{h}(t) \in \mathbf{Z}$ and $\overline{h}(t) > 0$, using an arbitrary small value $\gamma > 0$, we obtain a quantization parameter Q_p based on the infimum of $\overline{h}(t)$, as follows:

$$\begin{aligned}Q_p(t) &= \eta \cdot b^{\left\lceil -\frac{\beta}{t+2} + \frac{1}{2}\log_b\left(\frac{n\log(t+2)}{24\cdot\eta^2\cdot C}\right)\right\rceil}\\ \because \overline{h}(t) &= \left\lceil -\frac{\beta}{t+2} + \sup_{t\ge 0}\overline{h}(t)\right\rceil.\end{aligned} \tag{25}$$

Practically, we observe the vanishing gradient due to the problem of significant figures caused by quantization when $\overline{h}(t)$ is not sufficiently large. However, because the proposed algorithm improves the resolution of quantization by increasing the infimum of $\overline{h}(t)$ as t increases, the vanishing gradient problem can be addressed through quantization. In addition, when the norm of gradient is zero or minute, we can increase $\overline{h}(t)$ while holding the supremum of $h(t)$ to address the issue.

4 | NUMERICAL EXPERIMENTS

We evaluated the performance of the proposed algorithm with 10-benchmark functions represented in Table 1.

The following are the 10-benchmark functions we selected: the Rosenbrock, Ackley, Whitley, and Powell functions, which are traditional test functions, and six recently developed test functions, which are rated difficult to find the optimal point. In addition, to evaluate the performance of the proposed algorithm for high-order optimization problems, we tested with 100-dimensional Rosenbrock and 4-dimensional Xin-She-Yang and Powell functions. Figure 2 presents the diagram of each benchmark function and trace plots to each tested algorithm around the optimal points. Table 1 shows the difficulty scores in finding the global optimal point of each function, and Table 2 represents input domains and optimal points with respect to each test function. The quantization process starts at a 5-bit resolution and ends at a maximum of 17-bit resolution or less.

In experiments, we compared the proposed algorithm's search speed and performance with three conventional gradient-based algorithms using a line search method based on the Armijo-Wolf method. The conventional algorithms are the general gradient descent, conjugate gradient, and quasi-Newton with Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, and we apply the proposed quantization scheme to each gradient algorithm.

The line search algorithm effectively finds the minimum point on the search line that appears as a gradient

at every step. Therefore, the proposed algorithm performs the optimal line search compared with the conventional algorithm. As a result, the proposed algorithm increases the search speed even though its search performance is similar to that of the conventional searching algorithm without degradation by quantization error. Table 3 shows that global optimization performance is slightly better despite performance degradation caused by the quantization error, and the performance speed is improved by approximately 30% or more.

5 | CONCLUSION

We proposed a quantized gradient-based searching algorithm that can reduce the operation time by monotonically reducing the quantization step with respect to time without degrading optimization performance. We derived the quantization schedule based on weak convergence and evaluated appropriate quantization parameters. The proposed algorithm is more suitable for various embedded systems because quantization is composed of fixed-point fractional values. Consequently, it is possible to develop a large-scale parallel optimizer in an embedded system effectively for machine learning with limited computational capacitance.

The numerical experimental results for nonlinear benchmark functions show that the proposed algorithm achieves a fast searching speed without degrading optimization performance by quantization error. As a result, it is possible to apply the proposed algorithm to a general optimization field, such as reinforcement learning, as future work. Moreover, we will develop high-performance algorithms in embedded systems by exploiting better quantization step scheduling approaches.

ACKNOWLEDGEMENTS

This work was supported by Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (2017-0-00142, Development of Acceleration SW Platform Technology for On-device Intelligent Information Processing in Smart Devices, and 2021-0-00766, Development of Integrated Development Framework that supports Automatic Neural Network Generation and Deployment optimized for Runtime Environment).

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

ORCID

Jinwuk Seok https://orcid.org/0000-0001-5318-1237

REFERENCES

1. S. Garg, *Embedded systems market—global forecast to 2023*, MarketsandMarkets, 2017.
2. W. Han Yun, D. Kim, H.-S. Yoon, and J. Lee, *Disguised-face discriminator for embedded systems*, ETRI J. **32** (2010), no. 5, 761–765.
3. Y. C. Yoon, S. Y. Park, S. M. Park, and H. Lim, *Image classification and captioning model considering a cam-based disagreement loss*, ETRI J. **42** (2019), no. 1, 67–77.
4. C. M. De Sa, C. Zhang, K. Olukotun, C. Ré, and C. Ré, *Taming the wild: A unified analysis of hogwild-style algorithms*, *Advances in neural information processing systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, (eds.), Curran Associates, Inc., 2015, pp. 2674–2682.
5. F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, *1-bit stochastic gradient descent and application to data-parallel distributed training of speech DNNS*, (INTERSPEECH, Singapore), Sept. 2014, pp. 1058–1062.
6. N. Strom, *Scalable distributed dnn training using commodity GPU cloud computing*, (Sixteenth Annual Conference of the International Speech Communication Association), 2015, pp. 1488–1492.
7. D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, *Qsgd: Communication-efficient SGD via gradient quantization and encoding*, *Advances in neural information processing systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, (eds.), Curran Associates, Inc., 2017, pp. 1709–1720.
8. A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, *The marginal value of adaptive gradient methods in machine learning*, *Advances in neural information processing systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, (eds.), Curran Associates, Inc., 2017, pp. 4148–4158.
9. S. J. Osher, B. Wang, P. Yin, X. Luo, M. Pham, and A. T. Lin, *Laplacian smoothing gradient descent*, arXiv preprint, 2018. https://doi.org/10.48550/arXiv.1806.06317
10. K. Bae, H. Ryu, and H. Shin, *Does adam optimizer keep close to the optimal point?* arXiv preprint, 2019. https://doi.org/10.48550/arXiv.1911.00289
11. D. Jiménez, L. Wang, and Y. Wang, *White noise hypothesis for uniform quantization errors*, SIAM J. Math. Anal. **38** (2007), no. 6, 2042–2056.
12. C.-R. Hwang, *Laplace's method revisited: weak convergence of probability measures*, Ann. Probab. **8** (1980), no. 6, 1177–1182.
13. T.-S. Chiang, C.-R. Hwang, and S. J. Sheu, *Diffusion for global optimization in* $\mathbf{R}^n$, SIAM J. Control Optim. **25** (1987), no. 3, 737–753.
14. S. Geman and C.-R. Hwang, *Diffusions for global optimization*, SIAM J. Control Optim. **24** (1986), no. 5, 1031–1043.
15. F. C. Klebaner, *Introduction to stochastic calculus with applications*, Introduction to Stochastic Calculus with Applications, Imperial College Press, 2005.
16. B. Oksendal, *Stochastic differential equations: An introduction with applications*, Universitext, Springer Berlin Heidelberg, 2013.

AUTHOR BIOGRAPHIES

Jinwuk Seok received his BS and MS degrees in Electrical Control Engineering from Hong-Ik University, Seoul, Republic of Korea, in 1993 and 1995, respectively. Additionally, he received his PhD degree in Electrical Engineering from Hong-Ik University, Seoul, Republic of Korea, in 1998. He has been a principal member of engineering staff at Electronics and Telecommunications Research Institute in Korea since 2000, and an adjunct professor of Computer Software Engineering Department at the University of Science and Technology in Korea since 2009. His research interests include artificial intelligence, machine learning, and stochastic nonlinear control.

Jeong-Si Kim received her BS, MS, and PhD degrees in Computer Science from Gyeongsang National University, JinJu, Republic of Korea, in 1992, 1995, and 1999, respectively. Since 2000, she has worked in Electronics and Telecommunications Research Institute, Daejeon, South Korea, where she is now serving as a principal member of the engineering staff. She participated in "Development of Acceleration SW Platform Technology for On-device Intelligent Information Processing." Her research interests include on-device artificial intelligence, embedded systems, and debugging race conditions.

How to cite this article: J. Seok and J.-S. Kim, *Nonlinear optimization algorithm using monotonically increasing quantization resolution*, ETRI Journal **45** (2023), 119–130. https://doi.org/10.4218/etrij.2021-0320

APPENDIX A

Supplementary assumptions and lemmas

Assumption 3. For $w_t \in B^o(x^*,\rho)$, there exists a positive value L with respect to a scalar field $f(x):\mathbf{R}^n \to \mathbf{R}$ such that

$$\|f(x_t)-f(x^*)\| \le L\|x_t-x^*\|,\ \forall t > t_0, \tag{A1}$$

where $B^o(x^*,\rho)$ is an open ball $B^o(x^*,\rho) = \{x \mid \|x-x^*\| < \rho\}$.

Lemma 1. For all $x \in \mathbf{R}$,

$$(1-x) \le \exp(-x). \tag{A2}$$

Proof. By definition of the exponent, we write the exponential function as the following fundamental series:

$$\exp(-x) = \sum_{i=0}^{\infty} \frac{1}{i!}(-1)^i x^i = \sum_{k=0}^{\infty}\left(\frac{1}{2k!}x^{2k} - \frac{1}{(2k+1)!}x^{2k+1}\right). \tag{A3}$$

Let u_k be $u_k = (1/2k!)x^{2k}(1-(1/2k+1)x)$, and then we rewrite the series of exponent such that

$$\exp(-x) = u_0 + \sum_{k=1}^{\infty} u_k. \tag{A4}$$

For all $k > 0$, because each u_k is positive, we have

$$1-x = u_0 \le u_0 + \sum_{k=0}^{\infty} u_k. \tag{A5}$$

□

Proofs of Theorem 1

Proof. For the proof of the theorem, we depend on the lemmas in the works of [14]. First, we prove the following convergence of the transition probability:

$$\lim_{\tau\to\infty} \sup_{w_t, w_{t+\tau} \in \mathbf{R}^n} \|p(t,\overline{w}_t,t+\tau,w^*) - p(t,w_t,t+\tau,w^*)\| = 0, \tag{A6}$$

where t and τ are the epoch index and iteration to a single data index, respectively. w^* represents an optimal weight vector when the proposed algorithm selects a feasible learning rate. Let the infimum of the transition probability from t to $t+1$ such that

$$\delta_t = \inf_{x,y \in \mathbf{R}^n} p(t,x,t+1,y) \tag{A7}$$

Following the lemma in [14], the upper bound of (A6) is

$$\overline{\lim_{\tau\to\infty}} \sup_{w_t, w_{t+\tau}\in\mathbf{R}^n} \|p(t,\overline{w}_t,t+\tau,w^*)-p(t,w_t,t+\tau,w^*)\| \le 2\|w^*\|_\infty \prod_{k=0}^{\infty}(1-\delta_{t+k}). \tag{A8}$$

From the exponential approximation lemma (A2), we rewrite (A8), as follows:

$$\overline{\lim_{\tau\to\infty}} \sup_{w_t, w_{t+\tau}\in\mathbf{R}^n} \|p(t,\overline{w}_t,t+\tau,w^*)-p(t,w_t,t+\tau,w^*)\| \le 2\|w^*\|_\infty \exp\left(-\sum_{k=0}^{\infty}\delta_{t+k}\right)\Bigg). \tag{A9}$$

In this study, to obtain the bound of δ_{t+k}, we rewrite the stochastic differential form derived from (15), as follows:

$$dW_s = -\nabla H(W_s)ds + \sigma(s)\sqrt{G}dB_s,\ s\in\mathbf{R}(t,t+1), \tag{A10}$$

where $\sigma(s)\triangleq Q_p^{-1}(s)$, $G=\frac{n}{12}$, and $\nabla H(W_s)=\lambda_s\nabla f(W_s)$. Define a domain $\mathcal{F}\{f:[t,t+1]\to\mathbf{R}^n,\ f\in C^2\}$. Let P_x be the probability measures on $\mathcal{F}$ induced by (A10) and Q_x derived by the following equation:

$$d\overline{W}_\tau = \sigma(\tau)\sqrt{G}dB_\tau,\ \tau\in\mathbf{R}(t,t+1). \tag{A11}$$

Following the Girsanov theorem [15,16], we obtain

$$\frac{dP_w}{dQ_w} = \exp\left[\int_t^{t+1}\frac{G^{-1}}{\sigma^2(\tau)}\langle-\nabla H(W_\tau),d\overline{W}_\tau\rangle - \frac{1}{2}\int_t^{t+1}\frac{G^{-1}}{\sigma^2(\tau)}\|\nabla H(W_\tau)\|^2 d\tau\right]. \tag{A12}$$

To compute the upper bound of (A12), we will evaluate the upper bound of $\|\nabla H\|$. However, because $\|G\|$ does not depend on the time index s, we regard it as a constant value for all s. By definition, because the objective function is continuous, the gradient of $H(w_s)$ fulfills the Lipschitz continuous condition (A1) as well. Therefore, for $w_t\in B^o(w^*,\rho)$, there exists a positive value L' such that

$$\|\nabla f(w_\tau)-\nabla f(w^*)\| \le L'\|w_\tau - w^*\|,\ \forall\tau>0. \tag{A13}$$

Successively, if the objective function $f(x)$ is strictly convex, the Lipschitz condition takes the following form:

$$\|\nabla H(w_t)\| \le L'\lambda_t\rho = C_0. \tag{A14}$$

Consequently, for all $s\in\mathbf{R}[t,t+1)$, we compute the upper bound of the first term in the exponential function, as follows:

$$\begin{aligned}&\left\|\int_t^{t+1}\frac{G^{-1}}{\sigma^2(s)}\langle\nabla H(W_s),d\overline{W}_s\rangle\right\| \\ &\le \int_t^{t+1}\left\|\frac{G^{-1}}{\sigma^2(s)}\langle\nabla H(W_s),d\overline{W}_s\rangle\right\| \\ &\le \int_t^{t+1}\frac{\|G^{-1}\|}{\sigma^2(s)}\|\nabla H(W_s)\|\sigma(s)\sqrt{\|G\|}dB_s \\ &\le \frac{\sqrt{\|G^{-1}\|}}{\sigma(s)}C_0\|B_t-\frac{1}{2}\| \le \frac{1}{\sigma(s)}C_0\sqrt{\|G^{-1}\|}\left(\rho+\frac{1}{2}\right).\end{aligned} \tag{A15}$$

It implies that

$$\left\|\int_t^{t+1}\frac{G^{-1}}{\sigma(s)}\langle-\nabla H(W_\tau,X_\tau),d\overline{W}_\tau\rangle\right\| \le \frac{C_1}{\sigma(s)}, \tag{A16}$$

where C_1 is a positive value such that $C_1 > C_0\sqrt{\|G^{-1}\|}(\rho+1/2)$.

In addition, the upper bound of the second term is

$$\begin{aligned}&\frac{1}{2}\left\|\int_t^{t+1}\frac{G^{-1}}{\sigma^2(s)}\|\nabla H(W_s)\|^2 d\tau\right\| \\ &\le \frac{1}{2}\int_t^{t+1}\frac{\|G^{-1}\|}{\sigma^2(s)}\|\nabla H(W_s)\|^2 d\tau \\ &\le \frac{1}{2\sigma^2(s)}\|G^{-1}\|\cdot C_0^2 \le \frac{C_2}{2\sigma^2(s)},\ \because C_2 > \|G^{-1}\|\cdot C_0^2.\end{aligned} \tag{A17}$$

By assumption, because $\sigma(s)$ is a monotonically decreasing function, the supremum of $\sigma(s)$ is $\sigma(0)$ for all $s\in\mathbf{R}[0,\infty)$, that is, $\sup_{s\in\mathbf{R}[0,\infty]}\sigma(s)=s(0)$. With the supremum of each term in (A12), we obtain the lower bound of the Radon-Nykodym derivative (A12) such that

$$\frac{dP_w}{dQ_w} \ge \exp\left(-\frac{1}{\sigma(s)}\left(C_1+\frac{C_2}{2\sigma(s)}\right)\right) \ge \exp\left(-\frac{C_3}{\sigma(s)}\right), \tag{A18}$$

where $C_3 > 2\sigma(0)C_2 + C_1$.

Consequently, for any $\varepsilon>0$ and $w_t, w^*\in\mathbf{R}^n$, the infimum of $P_w(|W_{t+1}-w^*|<\varepsilon)$ is

$$P_w(|W_{t+1}-w^*|<\varepsilon)\geq \exp\left(-\frac{C_3}{\sigma(s)}\right) \quad Q_w(|W_{t+1}-w^*|<\varepsilon). \tag{A19}$$

Because Q_w is a normal distribution based on (A11), we have

$$\begin{aligned}
&P_w(|W_{t+1}-w^*|<\varepsilon)\geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\cdot \\
&\int_{\|x-w^*\|<\varepsilon}\frac{1}{\sigma\sqrt{2\pi\int_t^{t+1}G d\tau}}\exp\left(-\frac{(x-w^*)^2}{2\int_t^{t+1}G d\tau}\right)dx \\
&\geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\cdot \\
&\int_{\|x-w^*\|<\varepsilon}\frac{1}{\sigma\sqrt{2\pi\|G\|\int_t^{t+1}d\tau}} \\
&\exp\left(-\frac{\left(\sqrt{\rho}+\varepsilon\right)^2}{2\|G\|\int_t^{t+1}d\tau}\right)dx \\
&\geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\frac{1}{\sigma(0)\sqrt{2\pi\|G\|}} \\
&\exp\left(-\frac{\left(\sqrt{\rho}+\varepsilon\right)^2}{2\|G\|}\right)\int_{\|x-w^*\|<\varepsilon}dx \\
&\geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\frac{1}{\sigma(0)\sqrt{2\pi\|G\|}}\left(1+\frac{\left(\sqrt{\rho}+\varepsilon\right)^2}{2\|G\|}\right)2\varepsilon \\
&\geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\cdot C_4\cdot\varepsilon,\ \because C_4=\frac{\sqrt{2}}{\sigma(0)\sqrt{\pi\|G\|}}.
\end{aligned} \tag{A20}$$

Finally, we obtain the lower bound of the transition probability such that

$$\begin{aligned}
\delta_t &= \inf_{x,y\in\mathbf{R}^n}p(t,x,t+1,y)|_{x=w_t,\ y=w^*} \\
&= \inf_{x,y\in\mathbf{R}^n}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}P_w(|W_{t+1}-w^*|<\varepsilon) \\
&\geq \inf_{x,y\in\mathbf{R}^n}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\cdot C_4\cdot\exp\left(-\frac{C_3}{\sigma(s)}\right)\cdot\cdot\varepsilon \\
&\geq \exp\left(-\frac{C_5}{\sigma(s)}\right),\ \because C_5>C_3+\sigma(0)\cdot|\ln C_4|.
\end{aligned}$$

Therefore, if there exists a monotonically decreasing function such that $\sigma(s)\geq\frac{C_5}{\log(t+2)}$, the convergence condition derived by (A9) is satisfied such that

$$\sum_{k=0}^{\infty}\delta_{t+k}=\infty,\ \forall k\geq 0. \tag{A21}$$

□