Stochastic Learning Equation using Monotone Increasing Resolution of Quantization

Jinwuk Seok JNWSEOK@ETRI.RE.KR
Jeong-Si Kim SIKIM00@ETRI.RE.KR
High Performance Embedded System SW Research Section, Future Computing Research Division, Artificial Intelligence Research Lab. Electronics and Telecommunications Research Institute, Rep. Korea

Abstract

In this paper, we propose a quantized learning equation with a monotone increasing resolution of quantization and stochastic analysis for the proposed algorithm. According to the white noise hypothesis for the quantization error with dense and uniform distribution, we can regard the quantization error as i.i.d. white noise. Based on this, we show that the learning equation with monotonically increasing quantization resolution converges weakly as the distribution viewpoint. The analysis of this paper shows that global optimization is possible for a domain that satisfies the Lipschitz condition instead of local convergence properties such as the Hessian constraint of the objective function.

Keywords: Quantization, Machine Learning, Learning Equation, Stochastic Analysis

1. Introduction

A lot of researchers have regarded quantization as an efficient methodology enabling signal processing by reducing the amount of computation in small-scale hardware in the field of engineering(Boutalis et al. [3], Ljung and Ljung [12], Weiss and Mitra [16], including machine learning (Han et al. [7, 8], Wen et al. [17]). However, due to error generation and propagation by quantization, designing quantization in the signal processing has been updating the least significant bits corresponding directional derivative (Alistarh et al. [1], Hubara et al. [9], Seide et al. [14]). This type of quantization is vulnerable in local minima, occurring degradation of performance in signal processing, and the same problem occurs in the learning equation to machine learning. In this paper, we provide a novel quantized learning equation that can find an optimal point in a domain satisfying Lipschitz continuous to be robust to local minima. Increasing the resolution of quantization with respect to time, we note that the learning equation can find the global optimal point within such a domain by stochastic analysis. Numerical experiments show that the proposed methodology overcomes the weak point in conventional quantization, such as degradation of performance caused by quantization.

2. Fundamental Definition and Formulation for the Proposed Algorithm

We set the following definitions and assumptions, before beginning of our discussion.

Definition 1 *For $x \in \mathbf{R}$, we define the quantization as follows:*

$$x^Q \triangleq \frac{1}{Q_p}\lfloor Q_p \cdot (x + 0.5 \cdot Q_p^{-1})\rfloor = \frac{1}{Q_p}\left(Q_p \cdot x + \varepsilon\right) = x + \varepsilon Q_p^{-1}, \quad x^Q \in \mathbf{Q}. \tag{1}$$

, where $x^Q \in \mathbf{Z}$ is an integral part of $x \in \mathbf{R}$, $Q_p \in \mathbf{Z}^+$ is a quantization parameter, and ε is a quantization error such as $\varepsilon \in \mathbf{R}[-0.5, 0.5]$.

Assumption 1 *For $x_t \in B^o(x^*, \rho)$, there exist a positive value L w.r.t. a scalar field $f(x) : \mathbf{R}^n \to \mathbf{R}$ such that*

$$\|f(x_t) - f(x^*)\| \leq L\|x_t - x^*\|, \quad \forall t > t_0 \tag{2}$$

where $B^o(x^*, \rho)$ is an open ball such that $B^o(x^*, \rho) = \{x | \|x - x^*\| < \rho\}$, and $f(\cdot)$ is an objective function.

In (1), we replace the constant quantization parameter Q_p with a monotone increasing quantization parameter concerning time t such as $Q_p(t)$. Thereby, we obtain the quantization error term as a monotone decreasing function for time t. If a quantization error exists densely distributed and follows a uniform distribution, we can regard the quantization error to be a white noise according to Barnes et al. [2], Claasen and Jongepier [4], Gray and Neuhoff [6], Sripad and Snyder [15]. Additionally, in case of which a quantization error is a vector such as $\boldsymbol{\varepsilon} \in \mathbf{R}^n$, if it is pairwise independent and follows a uniform distribution asymptotically, Jiménez et al. [10] proves that the vector valued quantization error is a white noise as well. Therefore, we regard the quantization error vector as a white noise, without proof.

When the weight vector $w_t \in \mathbf{R}^n$, $w_t = \{w_t^1, w_t^2, \cdots w_t^n\}$ and a learning rate $\lambda_t = \alpha \in \mathbf{R}(0, 1)$, $\forall t \in \mathbf{Z}^+$ are given, , we can obtain a canonical formulation for quantized learning equation as follows:

$$w_{t+1} = w_t - \lambda_t \cdot h(w_t) \tag{3}$$

, where $h(w_t)$ is a directional derivative corresponding to the objective function $f(w_t, x_t)$ for machine learning such that $h(w_s) \triangleq (J \circ \nabla f)(w_s)$ for some function J . For example, if $J(w_s)$ is an identity function such that $J \circ f(x) = f(x)$, $h(w_s) = \nabla f(w_s)$, where $f(w_s)$ is an objective function for a machine learning algorithm.

Suppose that the parameter vector of the current step w_t and the next step w_{t+1} are quantized, we have

$$w_{t+1}^Q = \Big(w_t^Q - \lambda_t \cdot h(w_t)\Big)^Q = w_t^Q - (\lambda_t \cdot h(w_t))^Q \tag{4}$$

, where the $\boldsymbol{\varepsilon}$ is the vector valued quantization error so that the distribution of components are independent distribution defined $\boldsymbol{\varepsilon} \in \mathbf{R}^n$. If there exist a rational number $\alpha_t \in \mathbf{Q}(0, Q_p)$ instead of λ_t , we have the following quantized learning equation by simple calculation.

$$w_{t+1}^Q = w_t^Q - \frac{\alpha_t}{Q_p} \cdot Q_p h(w_t) + \varepsilon_t Q_p^{-1} = w_t^Q - \frac{\alpha_t}{Q_p}(Q_p h(w_t))^Q \quad \because \alpha_t \in \mathbf{Q}. \tag{5}$$

Hereby, we can obtain the search equation providing the quantized parameter vector for all steps $t \in \mathbf{N}$ by the mathematical induction.

3. Stochastic Analysis

3.1. Analysis of the Proposed Quantization

If the quantization error vector $\boldsymbol{\varepsilon}_t \in \mathbf{R}^n$ satisfying the WNH, Gray and Neuhoff [6], Klebaner [11] shows that the deviation can be calculated as follows:

$$\forall \boldsymbol{\varepsilon}_t \in \mathbf{R}^n, \mathbb{E}Q_p^{-2}\boldsymbol{\varepsilon}_t^2 = \mathbb{E}Q_p^{-2} \cdot tr(\boldsymbol{\varepsilon}_t \boldsymbol{\varepsilon}_t^T) = \frac{1}{12 \cdot Q_p^2} \cdot n. \tag{6}$$

Since the WNH establishes that the quantization error is a i.i.d. white noise, we can regard that weight vector $w_t^Q \in \mathbf{R}^n$ as a stochastic process $\{W_t\}_{t=0}^{\infty}$. Suppose that N is the number of needed data while the weight vector is updated. In other words, if t is an index of epoch, then N is total number of data, and if t is an index of mini-batch, N is the number of data in a unit mini-batch. Let a granular time index $s : \mathbf{Z}[0, N) \to \mathbf{R}[0, 1),\ s \in \mathbf{R}[t, t+1)$ such that $s(\tau) = \frac{1}{N}\tau$, where $\tau \in \mathbf{R}[0, N)$ so to set the time index between t and $t+1$. Additionally, we replace time index such that $t = t_1$ and $t+1 = t_N$, for convenience. Let $Z(s(\tau)) = w_{t_1} + s(\tau)(w_{t_N} - w_{t_1})$. By chain rule, we can obtain

$$\int_0^1 dZ(s) = \int_0^1 \frac{\partial Z(s)}{\partial s} ds = \int_t^{t+1} dw_s = \int_0^1 \left(-\alpha_t \nabla f(w_{t_1}) + \boldsymbol{\varepsilon_s} Q_p^{-1}(t_1)\right) ds. \tag{7}$$

Differentiate both sides in (7) to s. Additionally, letting $\boldsymbol{\varepsilon_s} ds = \sqrt{\frac{n}{12}} dB_s$ from (6) and WHN, we get the following stochastic differential equation to the proposed learning equation:

$$dW_s = -\alpha_t h(W_t) ds + \boldsymbol{\varepsilon_s} Q_p^{-1}(t) ds = -\alpha_t h(W_t) ds + \sqrt{\frac{n}{12}} Q_p^{-1}(s) d\boldsymbol{B_s} \tag{8}$$

where dB_s is the differential of a vector valued standard Wiener process with mean zero and variance one.

Theorem 2 *If the stochastic differential equation induced by the learning equation (5) satisfies (8), the stochastic process $\{W_t\}_{t=0}^{\infty}$ generated by the learning equation weakly converges to the global minimum on the domain defined in Assumption 1 when the deviation of the quantization error is given as follows:*

$$\inf_{t \geq 0} \sigma(t) = \frac{C}{\log(t+2)}, \quad C \in \mathbf{R},\ C \gg 0 \tag{9}$$

where, $\sigma(t) = \sqrt{\frac{n}{24}} Q_p(s)^{-1}$.

Theorem 2 means that if we properly increase the quantization resolution $Q(s)$, the proposed quantization learning equation can find the global minima of an objective function within the domain satisfying Lipschitz continuous is satisfied. We provide a detailed proof of Theorem in Appendix.

3.2. Scheduler function

Since $\sigma(t) \in \mathbf{R}$ is a proportional value to $Q_p(t) \in \mathbf{Z}$, we can't apply the result of Theorem 2 to the proposed quantization. However, if there exists a feasible $\sigma(t) \in \mathbf{Z}$ such that $\sigma(t) \geq \inf \sigma(t) \triangleq c/\log(2+t)$, it satisfies the Theorem 2. Furthermore, if there exists the supremum of $\sigma(t)$ such that $\inf \sigma(t) \leq \sigma(t) \leq T(t)$, the proposed quantization that satisfies the condition for global optimization is possible avoiding a extreme 1-bit quantization at early stage. For this, we define the quantization parameter $Q_p(t)$ depending on a monotone increasing function $\bar{h}(t) \in \mathbf{Z}^+$ with respect to time t.

$$Q_p(t) = \eta \cdot b^{\bar{h}(t)}, \quad \text{such that} \quad \bar{h}(t) \uparrow \infty \ \text{ as } \ t \to \infty. \tag{10}$$

By a simple calculation using the results in the previous section, $\bar{h}(t)$ satisfying (10) has the following supremum and the infimum.

$$\frac{1}{2}\log_b\left(\frac{n}{24 \cdot \eta^2} \cdot T(t)^{-1}\right) \leq \bar{h}(t) \leq \frac{1}{2}\log_b\left(\frac{n \log(t+2)}{24 \cdot \eta^2 \cdot C}\right) \tag{11}$$

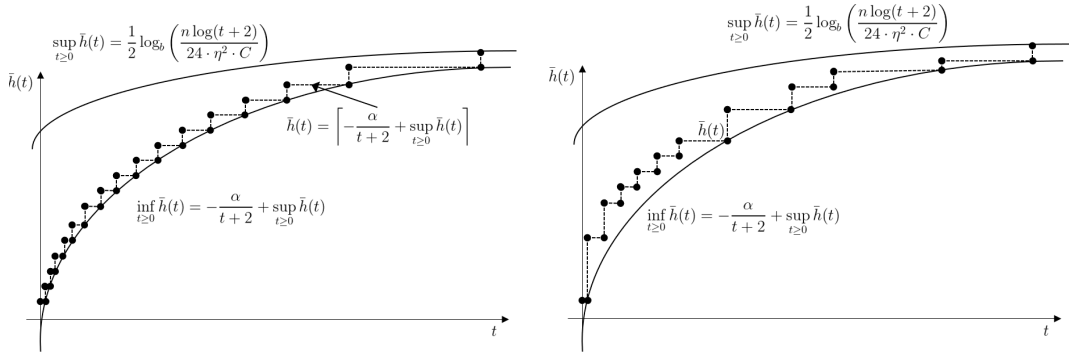

Figure 1: (a) Theoretical trend of $\sigma(t)$ based on the infimum (b) Practical trend of $\sigma(t)$ to avoid the vanishing gradient by the significant figure owing to quantization

To specify the $\bar{h}(t)$, we let $T(t)$ for $b > 1$ such that $T(t) = b^{\left(\frac{2\beta}{t+2}\right)} \cdot \inf_{t\geq 0}\sigma(t)$. Based on such $T(t)$, evaluating the supremum and the infimum of $\bar{h}(t)$, we can get the $\bar{h}(t)$ as follows:

$$\bar{h}(t) \geq \frac{1}{2}\log_b\left(\frac{n}{24\cdot\eta^2}\cdot T(t)^{-1}\right) = -\frac{\beta}{t+2}+\sup_{t\geq 0}\bar{h}(t), \quad \because \sup_{t\geq 0}\bar{h}(t) = \frac{1}{2}\log_b\left(\frac{n\log(t+2)}{24\cdot\eta^2\cdot C}\right) \tag{12}$$

Using (12), calculating the quantization parameter, we can obtain the $Q_p(t) = \eta\cdot b^{\bar{h}(t)}$ satisfying the result of Theorem 2. Figure 1 presents the conceptual diagram of the proposed schedule function.

4. Numerical Experiments

We conduct numerical experiments on the image classification problem to verify the empirical validation of the proposed algorithm and the analysis. The data set we employ in the experiment is the well-known CIFAR-10 data. The test network for the experiment is a ResNet with 32 Layers. The number of total samples to training is 50000, the testing data is 10000. We use the cross-entropy loss as the objective function provided by the PyTorch that is an A. I. framework based on python. We perform the ten training times with every 100 epochs, and we yield the classification accuracy from the average of Top-1 accuracy to the training and testing data set. The algorithms used in the experiments are the general stochastic gradient descent(SGD) algorithm and the ADAM (ADaptive Moment Estimation) algorithm widely used in machine learning. We compared the data classification accuracy by combining the proposed quantization algorithm with each conventional algorithm.

Furthermore, in the proposed quantization, we set the quantization parameter to be $Q_p(0) = 2^2$ at an initial stage, and the scheduler function calculates the quantization parameter every epoch. In the result of the numerical experiments, the proposed algorithm shows better classification performance than both ADAM and SGD's, as represented in Table 1. We can regard the result of the conventional quantization learning equation using only the lower 1 bit as the performance of the existing algorithm when it has the lowest learning rate in Table 1. In particular, when applied to ADAM, the performance improvement is higher than that of SGD. We consider that such a result is based on ADAM's feasible search domain is wider than SGD's.

	ADAM		QtADAM		SGD		QSGD	
Learning rate	Training	Test	Training	Test	Training	Test	Training	Test
0.25	78.73	69.26	78.16	71.28	95.78	77.27	95.95	77.27
0.125	79.47	69.00	83.86	72.75	93.86	74.42	94.89	75.69
0.0625	88.23	74.46	89.12	76.63	92.34	73.42	91.63	73.21
0.03125	93.91	78.91	94.62	79.13	86.45	68.65	86.06	67.36
0.015625	95.62	80.24	95.57	80.54	77.84	65.73	78.00	66.41
0.0078125	95.95	80.77	96.78	81.39	71.24	65.74	70.10	65.01
0.00390625	96.23	81.20	96.43	81.74	60.47	57.56	61.17	59.07
0.001953125	96.13	80.99	95.77	79.59	50.59	49.73	50.17	48.83
0.0009765625	95.71	78.28	94.84	77.12	43.49	42.66	44.18	43.19
Average	91.11	77.01	91.68	77.79	74.67	63.91	74.68	64.00

Table 1: The results of the numerical experiments about the classification of CIFAR-10 data set with various learning rates

Algorithm 1: Learning equation with the proposed quantization scheme

Data: Data-set needed classification such as the CIFAR-10
Result: Learned Network for Input Data
initialization
Set Parameters as $n \in \mathbf{Z}, \eta = 1, b = 2, \alpha \in \mathbf{Q}(0, 1]$
$t \leftarrow 0$ and Compute $\bar{h}(0)$ and $Q_p(0)$
while *meet stopping criterion* **do**
  $h_t \leftarrow (J \circ \nabla f)(w_t)$ `// Common weight update`
  $h^Q \leftarrow \frac{1}{Q_p}(Q_p \cdot h_t)^Q$ `// Quantization`
  Avoid Gradient Vanishing and Check the Limit of $\bar{h}_t$ `// Quantization`
  Check the bound of $\bar{h}(t)$ `// Quantization`
  $w_t \leftarrow w_t - \alpha \cdot h_t^Q$ `// Common weight update`
  $t \leftarrow t + 1$ `// to unit mini-batch or an epoch`
end

5. Conclusion

We present a quantization learning algorithm that monotonically increases the quantization resolution with respect to time and stochastic analysis of the proposed algorithm. The stochastic analysis of the proposed algorithm shows that the proposed quantization methodology can find the global optimum under the input domain satisfying Lipschitz continuous without any convex condition such as the limitation of Hessian. Therefore, we expect that the better the search capability of the conventional learning equation, the proposed algorithm shows better performance. We verify that the proposed algorithm shows superior classification performance without any degradation by quantization from the result of the numerical experiments.

Appendix A. Acknowledgement

This work was supported by Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2017-0-00142, Development of Acceleration SW Platform Technology for On-device Intelligent Information Processing in Smart Devices)

Appendix B. Supplementary Information

We provide the proof of the Theorem 2, the detailed procedure of the sub-functions for the proposed algorithm, and the information of hyper-parameters to each algorithm employed in numerical experiments.

B.1. Proof of Theorem 2

Proof For the proof of the theorem, we depend on the lemmas in works of Geman and Hwang [5]. The aim of this section is to prove the following convergence of the transition probability:

$$\lim_{\tau\to\infty} \sup_{w_t, w_{t+\tau}\in\mathbf{R}^n} \|p(t, \bar{w}_t, t+\tau, w^*) - p(t, w_t, t+\tau, w^*)\| = 0 \tag{13}$$

, where t and τ is the epoch index and the iteration to a single data index, respectively. w^* represents an optimal weight vector.

Let the infimum of the transition probability from t to $t+1$ such that

$$\delta_t = \inf_{x,y\in\mathbf{R}^n} p(t, x, t+1, y) \tag{14}$$

By the lemma in Geman and Hwang [5], the upper bound of (30) is

$$\overline{\lim_{\tau\to\infty}} \sup_{w_t, w_{t+\tau}\in\mathbf{R}^n} \|p(t, \bar{w}_t, t+\tau, w^*) - p(t, w_t, t+\tau, w^*)\| \le 2\|w^*\|_\infty \prod_{k=0}^{\infty}(1-\delta_{t+k}). \tag{15}$$

Since $\prod_{k=0}^{\infty}(1-\delta_{t+k}) \le \exp(-\sum_{k=0}^{\infty}\delta_{t+k})$, we rewrite (15) as follows:

$$\overline{\lim_{\tau\to\infty}} \sup_{w_t, w_{t+\tau}\in\mathbf{R}^n} \|p(t, \bar{w}_t, t+\tau, w^*) - p(t, w_t, t+\tau, w^*)\| \le 2\|w^*\|_\infty \exp(-\sum_{k=0}^{\infty}\delta_{t+k})). \tag{16}$$

Herein, to obtain the bound of δ_{t+k}, we rewrite the stochastic differential form derived from Theorem 2 as follows:

$$dW_s = -\nabla H(W_s)ds + \sigma(s)\sqrt{G}dB_s, \quad s\in\mathbf{R}(t, t+1). \tag{17}$$

, where $\sigma(s) \triangleq Q_p^{-1}(s)$, $G = \frac{n}{12}$, and $\nabla H(W_s) = \lambda_s h(W_s) \triangleq \lambda_s \cdot (J\circ\nabla f)(W_s)$ for a function J such that $\nabla H(W^*) = \lambda_* h(W^*) = (J\circ\nabla f)(W^*) = 0$ as represented in (3).

Define a domain $\mathcal{F}\{f : [t, t+1] \to \mathbf{R}^n, f \text{ continuous }\}$, Let P_x be the probability measures on $\mathcal{F}$ induced by (17) and Q_x derived by the following equation:

$$d\bar{W}_\tau = \sigma(\tau)\sqrt{G}dB_\tau, \quad \tau\in\mathbf{R}(t, t+1). \tag{18}$$

By the Girsanov theorem (introduced in [11, 13]), we obtain

$$\frac{dP_w}{dQ_w} = \exp\left\{\int_t^{t+1} \frac{G^{-1}}{\sigma^2(\tau)}\langle -\nabla H(W_\tau), d\bar{W}_\tau\rangle - \frac{1}{2}\int_t^{t+1} \frac{G^{-1}}{\sigma^2(\tau)}\|\nabla H(W_\tau)\|^2 d\tau\right\}. \tag{19}$$

To compute the upper bound of (19), we will check the upper bound of $\|\nabla H\|$. Whereas, since $\|G\|$ is not depending on time index s, we regard it as a constant value for all s. By definition, since the objective function is continuous, the gradient of $H(w_s)$ fulfills the Lipschitz continuous condition (2) too.

Thereby, for $w_t \in B^o(w^*, \rho)$, there exist a positive value L' such that

$$\|\nabla f(w_\tau) - \nabla f(w^*)\| \le L'\|w_\tau - w^*\|, \quad \forall \tau > 0. \tag{20}$$

Successively, by the definition of $\nabla H(W^*)$ being equal to zero, the Lipschitz condition forms simply as follows :

$$\|\nabla H(W_t)\| \le L'\lambda_t \rho = C_0. \tag{21}$$

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

$$\begin{aligned}
&\left\|\int_t^{t+1} \frac{G^{-1}}{\sigma^2(s)}\langle \nabla H(W_s), d\bar{W}_s\rangle\right\| \le \int_t^{t+1} \left\|\frac{G^{-1}}{\sigma^2(s)}\langle \nabla H(W_s), d\bar{W}_s\rangle\right\| \\
&\le \int_t^{t+1} \frac{\|G^{-1}\|}{\sigma^2(s)}\|\nabla H(W_s)\|\,\sigma(s)\sqrt{\|G\|}dB_s \le \frac{\sqrt{\|G^{-1}\|}}{\sigma(s)}\sup\|\nabla H(W_s)\|\int_t^{t+1} dB_s \\
&\le \frac{\sqrt{\|G^{-1}\|}}{\sigma(s)}C_0\|B_t - \frac{1}{2}\| \le \frac{1}{\sigma(s)}C_0\sqrt{\|G^{-1}\|}(\rho + \frac{1}{2}).
\end{aligned} \tag{22}$$

It implies that

$$\left\|\int_t^{t+1} \frac{G^{-1}}{\sigma(s)}\langle -\nabla H(W_\tau, X_\tau), d\bar{W}_\tau\rangle\right\| \le \frac{C_1}{\sigma(s)} \tag{23}$$

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

In addition, the upper bound of the second term is

$$\begin{aligned}
&\frac{1}{2}\left\|\int_t^{t+1} \frac{G^{-1}}{\sigma^2(s)}\|\nabla H(W_s)\|^2 d\tau\right\| \le \frac{1}{2}\int_t^{t+1} \frac{\|G^{-1}\|}{\sigma^2(s)}\|\nabla H(W_s)\|^2 d\tau \\
&\le \frac{1}{2}\frac{\|G^{-1}\|}{\sigma^2(s)}\sup\|\nabla H(W_s)\|^2\int_t^{t+1} d\tau \le \frac{1}{2\sigma^2(s)}\|G^{-1}\|\cdot C_0^2 \le \frac{C_2}{2\sigma^2(s)}, \quad \because C_2 > \|G^{-1}\|\cdot C_0^2.
\end{aligned} \tag{24}$$

By assumption, since $\sigma(s)$ is monotone decreasing function, the supremum of $\sigma(s)$ is $\sigma(0)$ for all $s \in \mathbf{R}[0, \infty)$, i.e. $\sup_{s\in\mathbf{R}[0,\infty]}\sigma(s) = s(0)$. With the supremum of each term in (19), we can obtain the lower bound of the Radon-Nykodym derivative (19) such that

$$\frac{dP_w}{dQ_w} \ge \exp\left(-\frac{1}{\sigma(s)}\left(C_1 + \frac{C_2}{2\sigma(s)}\right)\right) \ge \exp\left(-\frac{C_3}{\sigma(s)}\right), \quad \because C_3 > 2\sigma(0)C_2 + C_1. \tag{25}$$

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

$$P_w(|W_{t+1} - w^*| < \varepsilon) \ge \exp\left(-\frac{C_3}{\sigma(s)}\right) Q_w(|W_{t+1} - w^*| < \varepsilon). \tag{26}$$

Since Q_w is a normal distribution based on (18), we have

$$P_w(|W_{t+1}-w^*|<\varepsilon) \geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\int_{\|x-w^*\|<\varepsilon}\frac{1}{\sigma(s)\sqrt{2\pi\int_t^{t+1}G d\tau}}\exp\left(-\frac{(x-w^*)^2}{2\int_t^{t+1}G d\tau}\right)dx$$
$$\geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\frac{1}{\sigma(0)\sqrt{2\pi\|G\|}}\exp\left(-\frac{(\sqrt{\rho}+\varepsilon)^2}{2\|G\|}\right) \geq \exp\left(-\frac{C_3}{\sigma(s)}\right)\cdot C_4 \cdot \ \because C_4 = \frac{\sqrt{2}}{\sigma(0)\sqrt{\pi\|G\|}}. \tag{27}$$

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

$$\delta_t = \inf_{x,y\in\mathbf{R}^n} p(t,x,t+1,y)\Big|_{x=w_t,\ y=w^*} = \inf_{x,y\in\mathbf{R}^n}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}P_w(|W_{t+1}-w^*|<\varepsilon)$$
$$\geq \inf_{x,y\in\mathbf{R}^n}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\cdot C_4\cdot\exp\left(-\frac{C_3}{\sigma(s)}\right)\cdot\cdot\varepsilon \geq \exp\left(-\frac{C_5}{\sigma(s)}\right), \quad \because C_5 > C_3+\sigma(0)\cdot|\ln C_4| \tag{28}$$

Therefore, if there exists a monotone decreasing function such that $\sigma(s) \geq \frac{C_5}{\log(t+2)}$, it satisfies that the convergence condition derived by (16) such that

$$\sum_{k=0}^{\infty}\delta_{t+k} = \infty, \quad \forall k \geq 0. \tag{29}$$

It implies that

$$\overline{\lim_{\tau\to\infty}}\sup_{w_t,w_{t+\tau}\in\mathbf{R}^n}\|p(t,\bar{w}_t,t+\tau,w^*)-p(t,w_t,t+\tau,w^*)\| = 0. \tag{30}$$

■

B.2. Auxiliary Sub-Functions for Main Algorithm

Algorithm 2: Avoid Gradient Vanishing and Check the Limit of $\bar{h}_t$

Data: $h^Q,\ t$
Result: Re-quantized h^Q, $\sup_{t\geq 0}\bar{h}(t)$
$\sup_{t\geq 0}\bar{h}(t) \leftarrow \frac{1}{2}\log_b\left(\frac{n\log(t+2)}{24\cdot\eta^2\cdot C}\right)$
while $\|h^Q\|>0$ *or* $\bar{h}(t) > \sup\bar{h}(t)$ **do**
  if $\|h^Q\|=0$ *and* $\bar{h}(t)\leq\sup\bar{h}(t)$ **then**
    $\bar{h}(t)\leftarrow\bar{h}(t)+1$ `// Increase Resolution of Quantization by 1`
    $Q_p(t)\leftarrow\eta\cdot b^{\bar{h}(t)}$
    $h^Q\leftarrow\frac{1}{Q_p}(Q_p\cdot h_t)^Q$ `// Re-quantization with updated` Q_p
  else
    $h^Q\leftarrow h^Q$
  end
end

Herein, we provide auxiliary sub-algorithms for the proposed main procedure illustrated as Algorithm 1. The Algorithm 2 is a sub-function to avoid vanishing gradient by quantization. The Algorithm 3 is a sub-function to limit quantization parameter between $\sup \bar{h}(t)$ and $\inf \bar{h}(t)$.

Algorithm 3: Check the bound of $\bar{h}(t)$

Data: $\bar{h}(t),\ \sup_{t\geq 0} \bar{h}(t)$
Result: Updated $\bar{h}(t)$
$\inf \bar{h}(t) \leftarrow -\frac{\beta}{t+2} + \sup_{t\geq 0} \bar{h}(t)$
while $\bar{h}(t) \geq \inf \bar{h}(t)$ **do**
  if $\bar{h}(t) < \inf \bar{h}(t)$ **then**
    $\bar{h}(t) \leftarrow \bar{h}(t) + 1$ `// Increase Resolution of Quantization by 1`
  else
    $\bar{h}(t) \leftarrow \bar{h}(t)$
  end
end

B.3. Hyper-parameters for algorithms

We set the Hyper-parameters for each algorithms in numerical experiments as follows:

B.3.1. SGD

- Directional Derivative
$$h(w_t) = \nabla f(w_t) \tag{31}$$
- Hyper-Parameter $\lambda \in \mathbf{R}(0, 1)$

B.3.2. ADAM

- Directional Derivative
$$h(w_t) = \frac{\sqrt{1-(\beta_2)^t}}{1-\beta^t} \cdot \frac{m_t^i}{\sqrt{v_t^i} + \varepsilon} \tag{32}$$
  - First order Momentum : $m_t^i = \beta_1 m_{t-1}^i + (1-\beta_1)\nabla f(w_t)^i$
  - Second order Momentum : $v_t^i = \beta_2 v_{t-1}^i + (1-\beta_2)(\nabla f(w_t)^i)^2$
- Hyper parameters
$$\lambda \in \mathbf{R}(0, 1), \eta_t = 0.001, \beta_1 = 0.9, \beta_2 = 0.999 \tag{33}$$

B.3.3. PROPOSED QUANTIZATION

For the proposed algorithm, since it is a quantization method, there is not a directional derivation.

- Hyper parameters
$$\bar{h}(0) = 2, b = 2, \eta = 1, \beta = 20.0, C = 10^6 \tag{34}$$
  Therefore, the initial value of the quantization parameter is $Q_p(0) = b^{\bar{h}(0)}$

References

[1] Dan Alistarh, Demjan Grubic, Jerry Li, Ryota Tomioka, and Milan Vojnovic. Qsgd: Communication-efficient sgd via gradient quantization and encoding. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 1709–1720. Curran Associates, Inc., 2017.

[2] Casper W. Barnes, Boi N. Tran, and Shu Hung Leung. On the statistics of fixed-point roundoff error. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 33(3):595–606, 1985. doi: 10.1109/TASSP.1985.1164611.

[3] Y. S. Boutalis, S. D. Kollias, and G. Carayannis. A fast multichannel approach to adaptive image estimation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(7): 1090–1098, 1989.

[4] T. Claasen and A. Jongepier. Model for the power spectral density of quantization noise. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(4):914–917, 1981.

[5] Stuart Geman and Chii-Ruey Hwang. Diffusions for global optimization. *SIAM Journal on Control and Optimization*, 24(5):1031–1043, 1986. doi: 10.1137/0324060.

[6] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44 (6):2325–2383, 2006.

[7] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1135–1143. Curran Associates, Inc., 2015.

[8] Song Han, Jeff Pool, Sharan Narang, Huizi Mao, Enhao Gong, Shijian Tang, Erich Elsen, Peter Vajda, Manohar Paluri, John Tran, Bryan Catanzaro, and William J. Dally. DSD: dense-sparse-dense training for deep neural networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

[9] Itay Hubara, Matthieu Courbariaux, Daniel Soudry, Ran El-Yaniv, and Yoshua Bengio. Quantized neural networks: Training neural networks with low precision weights and activations. *CoRR*, abs/1609.07061, 2016.

[10] David Jiménez, Long Wang, and Yang Wang. White noise hypothesis for uniform quantization errors. *SIAM J. Math. Analysis*, 38(6):2042–2056, 2007. doi: 10.1137/050636929.

[11] F.C. Klebaner. *Introduction to Stochastic Calculus with Applications*. Introduction to Stochastic Calculus with Applications. Imperial College Press, 2005. ISBN 9781860945557.

[12] S. Ljung and L. Ljung. Error propagation properties of recursive least-squares adaptation algorithms. *Automatica*, 21(2):157–167, 1985.

[13] B. Oksendal. *Stochastic Differential Equations: An Introduction with Applications*. Universitext. Springer Berlin Heidelberg, 2013. ISBN 9783662036204.

[14] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu. 1-bit stochastic gradient descent and application to data-parallel distributed training of speech dnns. In *Interspeech 2014*, September 2014.

[15] A. Sripad and D. Snyder. A necessary and sufficient condition for quantization errors to be uniform and white. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(5): 442–448, 1977.

[16] A. Weiss and D. Mitra. Digital adaptive filters: Conditions for convergence, rates of convergence, effects of noise and errors arising from the implementation. *IEEE Transactions on Information Theory*, 25(6):637–652, 1979.

[17] Wei Wen, Chunpeng Wu, Yandan Wang, Yiran Chen, and Hai Li. Learning structured sparsity in deep neural networks. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2074–2082. Curran Associates, Inc., 2016.