Progress Report

Revisiting `ASkewSGD`: New Theoretical Guarantees for Quantization-Aware Deep Neural Network Optimization

Undergraduate Summer Research Internship (2025)

Faculty of Engineering, The Chinese University of Hong Kong

July 7$^{\text{th}}$, 2025

Author: WONG, Hok Fong (1155189917)
Under the supervision of Prof. WAI, Hoi-To and Mr. YAU, Chung-Yiu

Contents

Abstract

The question of enforcing quantization on the weights and activations in Deep Neural Networks (DNNs) has come to the forefront in recent years due to its relevance to restricted memory and/or computational resources. While low-precision fixed integer values could substantially reduce the memory footprint and latency, inaccurate quantization on model parameters is susceptible to significant accuracy drops. In this paper, we will continue to pave the way for quantization algorithms during neural network training and provide stronger guarantees for the Annealed Skewed Stochastic Gradient Descent algorithm (`ASkewSGD`) proposed by Leconte et al. [1]. In particular, we attempt to retain the algorithm's convergence guarantees without hinging on a highly differentiable loss function. Numerical experiments show that `ASkewSGD` are able to produce state-of-the-art results in classical benchmarks, justifying its effectiveness as a robust optimization algorithm.

Contributions

- We improve the understanding of `ASkewSGD`, a quantization-aware training algorithm, and provide new convergence insights for `ASkewSGD` proposed by Leconte et al. [1] under weaker assumptions. We identify a loophole in their proof on the loss function's Lipschitzness, and further weaken the differentiability requirement of the loss function. Several theoretical guarantees are provided to address this issue.
- (Multi-bit assessment to be completed.) We evaluate the performance of `ASkewSGD` along with other SOTA quantization-aware training methods (`BinaryConnect`, `ProxQuant`) by numerical experiments on `MNIST`, `CIFAR-10`, and `ImageNet`. We made the related codes available at `https://github.com/SWongHF/Experiment-2025-Summer`.

1 Background

We are interested in solving the optimization problem related to learning a quantized neural network (QNN),

$$\min_{\mathbf{w}\in\mathcal{Q}} \ell(\mathbf{w}), \text{ where } \ell(\mathbf{w}) = \mathbb{E}_{(\mathbf{x},y)\sim p_{\text{data}}}[\ell(f(\mathbf{x},\mathbf{w}),y)],$$

where $\ell : \mathbb{R}^d \to \mathbb{R}$ denotes the training loss, $\mathcal{Q} \subset \mathbb{R}^d$ is the set of quantization levels, d is the number of parameters in the neural network, p_{data} is the training distribution [1].

We do not necessarily aim to solve the above (hard, combinational) optimization problem globally and optimally. Instead, we want to find discrete weights $\mathbf{w} \in \mathcal{Q}$ that remain "satisfactory" when compared to the non-quantized continuous weights [2].

Goal. Given a continuously differentiable and reasonably smooth function ℓ, develop an algorithm with certain guarantees that converges to a close-by quantized weight of a locally, or globally, optimal continuous weight with similar performance.

2 Existing Works

2.1 BinaryConnect

Courbariaux et al. [3] considered an aggresive quantization scheme using binary networks, where $\mathcal{Q} = \{\pm 1\}^d$. The `BinaryConnect` (BC) algorithm updates the weights by the following scheme:

$$\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} - \gamma_k \widehat{\nabla}\ell(\mathsf{P}(\mathbf{w}^{(k)})), \quad \text{where } [\mathsf{P}(\mathbf{w})]_i = \begin{cases} +1 \text{ if } w_i \geq 0, \\ -1 \text{ otherwise.} \end{cases}$$

While the above method is deterministic, yet another scheme quantizes the weights stochastically, which is given by:

$$[\mathsf{P}(\mathbf{w})]_i = \begin{cases} +1 \text{ with probability } \sigma(w_i), \\ -1 \text{ with probability } 1 - \sigma(w_i), \end{cases} \qquad \text{where } \sigma(x) = \operatorname{clip}\left(\frac{x+1}{2}, 0, 1\right).$$

For the general quantized neural network, BC can easily be adapted by adding a quantization step that maps a real number input $w \in [0, 1]$ to a k-bit number output w_q [4]:

$$w_q = \operatorname{round}((2^k - 1)w) - (2^{k-1} - 1).$$

The behaviour of BC is analyzed by [2], stating that the updates of BC are formally the same as the dual averaging (DA) algorithm (as a non-convex counterpart). The work [2] also generalizes BC into `ProxConnect` with rigorous convergence guarantees.

2.2 ProxQuant

Bai et al. [5] proposed the proximal gradient method for quantization-aware training, which is a variant of the proximal operator

$$\mathbf{w}^{(k+1)} \leftarrow \mathsf{P}\left(\mathbf{w}^{(k)} - \eta_k \widehat{\nabla}\ell(\mathbf{w}^{(k)})\right),$$

where the proximal operator is defined as

$$\mathsf{P}(\mathbf{w}) = \operatorname*{arg\,min}_{\bar{\mathbf{w}} \in \mathbb{R}^n} \left(\frac{1}{2}\|\mathbf{w} - \bar{\mathbf{w}}\|_2^2 + \lambda R(\bar{\mathbf{w}})\right), \text{where } R(\bar{\mathbf{w}}) = \inf_{\hat{\mathbf{w}} \in \mathcal{Q}} \|\hat{\mathbf{w}} - \bar{\mathbf{w}}\|_2.$$

This method is theoretically sound, but the proximal operator is expensive to compute when $\mathcal{Q}$ is large.

3 The `ASkewSGD` Algorithm

We turn to consider the optimization problem $(\mathcal{P}_\varepsilon)$,

$$\min_{\mathbf{w}\in C_\varepsilon} \ell(\mathbf{w}) := \frac{1}{N}\sum_{j=1}^{N} \ell_j(\mathbf{w}), C_\varepsilon = \{\mathbf{w}\in\mathbb{R}^d : g_{\varepsilon,i}(\mathbf{w}) \geq 0, \text{ for all } i = 1,2,\ldots,n\},$$

where N is the size of the training set, $\ell_j : \mathbb{R}^d \to \mathbb{R}$ is the loss associated with the j-th observation, $g_{\varepsilon,i} : \mathbb{R}^d \to \mathbb{R}^n$ is the i-th constraint with dependence on ε, which is a given parameter controlling the landscape of $g_{\varepsilon,i}$.

As an attempt to solving the problem $\mathcal{P}_\varepsilon$, Muehlebach and Jordan [6] reformulated the position constraints into forward "velocity" constraints by considering a linear and convex approximation of the original feasible set under the Mangasarian-Fromovitz condition (see Definition 1). The parameters $\mathbf{w}^{(k)}$, are hence allowed to escape from the feasible set temporarily in search of better local basins, and thus is possible to achieve better results when compared to the classical projected gradient descent algorithm.

Definition 1 (**Mangasarian-Fromovitz Constraint Qualification**). *$\forall \mathbf{w}\in\mathbb{R}^d$, $\exists \mathbf{v}\in\mathbb{R}^n$ s.t. $\nabla g_i(\mathbf{w})^\top \mathbf{v} > 0$ for all $i \in I(\mathbf{w})$, where $I(\mathbf{w}) = \{i\in[d] \mid g_i(\mathbf{w}) \leq 0\}$.*

We define $V_\alpha(\mathbf{w}) = \{\mathbf{v}\in\mathbb{R}^n : \nabla g_i(\mathbf{w})^\top \mathbf{v} \geq -\alpha g_i(\mathbf{w})\}$, which can be identified as an extension of the tangent cone outside of the feasible set, and is always a convex polyhedron. This is a special property inherited from the assumption of MFCQ.

The MJ Algorithm [6] computes its iterates $\mathbf{w}\in\mathbb{R}^d$ as follows:

Algorithm 1: Muehlebach-Jordan (MJ) Algorithm

for $k = 1, 2, \ldots$ **do**
- $\mathbf{v}^{(k)} = \arg\min_{\mathbf{v}\in V_\alpha(\mathbf{w}^{(k)})}(1/2)\|\mathbf{v} + \nabla\ell(\mathbf{w}^{(k)})\|^2$.
- $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \gamma_k \mathbf{v}^{(k)}$.

We note that α controls the trade-off betweeen two objectives: for large α, the emphasis is on the convergence to the feasible set, while for small α, the focus is on reducing the objective function [6].

It is proven by Muehlebach and Jordan [6] that the iterate eventually converges to the set of stationary points under convexity of ℓ and strong convexity of g. As promising as it shows, Leconte et al. [1] has applied the idea on the quantization-aware training problem, which resulted in the algorithm `ASkewSGD`.

Definition 2 (**Remoteness Measurement of the Quantization Set**). *Let $\varepsilon\in(0,1]$, $w_i \in \mathcal{Q}_i, i\in[d]$, where $\mathcal{Q}_i = \{q_i^{(1)},\ldots,q_i^{(K_i)}\}$ are sets of quantization values defined coordinate-wise. We define the piecewise function*

$$\psi_i(w_i;\varepsilon) := \begin{cases} \varepsilon - (q_i^{(1)} - w_i)^2 & w_i < q_i^{(1)}, \\ \varepsilon - (w_i - q_i^{(j-1)})^2(w_i - q_i^{(j)})^2 & q_i^{(j-1)} \leq w_i < q_i^{(j)}, j = 2,\ldots,K, \\ \varepsilon - (w_i - q_i^{(K_i)})^2 & w_i \geq q_i^{(K_i)}, \end{cases}$$

for all $w_i\in\mathbb{R}$ and $i\in[n]$.

To see how we can draw connection from $\mathcal{P}_\varepsilon$ to our QNN optimization problem, we impose constraints on each parameter of the neural network, by setting $g_i(\mathbf{w}) = \psi_i(w_i;\varepsilon)$. This then forms the feasible set C_ε. Now, it should be clear that $\cap_{\varepsilon\in(0,1)} C_\varepsilon = \mathcal{Q}$. As a consequence, we can solve the QNN optimization problem by considering the smoothed sequence of interval-constrained optimization problems $(\mathcal{P}_{\varepsilon_j})_{j\in\mathbb{N}}$ with $\varepsilon_j \to 0$ and $\lim_{j\to\infty}\varepsilon_j = 0$, while each $(\mathcal{P}_{\varepsilon_j})$ is solvable by a variant of the MJ Algorithm.

To match with the MJ Algorithm, we define $V_{\varepsilon,\alpha} = \{\mathbf{v}\in\mathbb{R}^d : v_i\psi_i'(w_i;\varepsilon) \geq -\alpha\psi_i(w_i;\varepsilon) \text{ for } i \in I_\varepsilon(\mathbf{w})\}$, where $I_\varepsilon(\mathbf{w}) = \{i \in [d] : \psi_i(w_i;\varepsilon) \leq 0\}$, and the normal cone of C_ε is given by $N_{C_\varepsilon} = \{-\sum_{i\in I_\varepsilon(\mathbf{w})}\lambda_i\nabla g_i(\mathbf{w}), \lambda_i \in \mathbb{R}_+\}$. Now, we are ready to present the `ASkewSGD` algorithm [1].

Algorithm 2: `ASkewSGD` Algorithm for QNN Training

---

1 **for** $k = 1, 2, \ldots$ **do**
2   Obtain a stochastic gradient $\widehat{\nabla}\ell(\mathbf{w}^{(k)}) = 1/N_b\sum_{i=1}^{N_b}\nabla\ell_{j_i}(\mathbf{w}^{(k)})$.
3     $\widehat{\mathbf{v}}^{(k)} = \arg\min_{\mathbf{v}\in V_{\varepsilon,\alpha}(\mathbf{w}^{(k)})}(1/2)\|\mathbf{v}+\widehat{\nabla}\ell(\mathbf{w}^{(k)})\|^2$.
4     $\mathbf{w}^{(k+1)} = \mathbf{w}^{(k)} + \gamma_k\widehat{\mathbf{v}}^{(k)}$.

In every iteration, the iterate $\mathbf{w}^{(k)}$ is updated by the velocity. The velocity $\mathbf{v}^{(k)}$ is chosen to match as closely to the gradient flow $-\nabla\ell(\mathbf{w}^{(k)})$ as possible, subject to the velocity constraint $\mathbf{v}^{(k)} \in V_{\varepsilon,\alpha}(\mathbf{w}^{(k)})$. This motivates us to solve the following optimization problem

$$\arg\min_{\mathbf{v}\in V_{\varepsilon,\alpha}(\mathbf{w})} (1/2)\|\mathbf{v}+\mathbf{u}\|^2,$$

which has an explicit solution (we set $[s_{\varepsilon,\alpha}(\mathbf{g},\mathbf{w})]_i = M_c$ if $w_i = (q_i^{(j)} + q_i^{(j+1)})/2$ by convention):

$$[\mathbf{s}_{\varepsilon,\alpha}(\mathbf{g},\mathbf{w})]_i := \begin{cases} -g_i & \text{if } \psi_i(w_i;\varepsilon) > 0 \text{ or } -g_i\cdot\psi_i'(w_i;\varepsilon) \geq -\alpha\psi_i(w_i;\varepsilon) \geq 0, \\ \operatorname{clip}(-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon), M_c) & \text{otherwise.} \end{cases}$$

Furthermore, the set of stationary points are given by the Karush-Kuhn-Tucker condition:

$$\mathcal{Z}_\varepsilon := \{\mathbf{w}\in C_\varepsilon : \mathbf{0} \in -\nabla\ell(\mathbf{w}) - N_{C_\varepsilon}(\mathbf{w})\}.$$

That is, $\mathbf{w}\in\mathcal{Z}_\varepsilon$ if and only if $[\nabla\ell(\mathbf{w})]_i = 0$ when $\psi_i(w_i;\varepsilon) > 0$ and $\operatorname{sign}([\nabla\ell(\mathbf{w})]_i) = \operatorname{sign}(\psi_i'(w_i;\varepsilon))$ when $\psi_i(w_i;\varepsilon) = 0$.

4 Existing Convergence Analysis

To set up the analysis, we will introduce different assumptions on the loss function and step sizes that will be used in this paper.

Assumption 1. *The function ℓ is coercive, i.e.*

$$\lim_{\|\mathbf{w}\|\to\infty} \ell(\mathbf{w}) \to +\infty.$$

Assumption 2. *For $j \in \{1, \dots, N\}$, the function ℓ_j is continuously differentiable and M_ℓ-smooth, i.e.*

$$\|\nabla\ell_j(\mathbf{x}) - \nabla\ell_j(\mathbf{y})\|^2 \leq M_{\ell_j}\|\mathbf{x}-\mathbf{y}\|^2 \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Assumption 3. *For $j \in \{1, \dots, N\}$, the function ℓ_j is continuously differentiable and L_ℓ-Lipschitz continuous, i.e.*

$$|\ell_j(\mathbf{x}) - \ell_j(\mathbf{y})|^2 \leq L_{\ell_j}\|\mathbf{x}-\mathbf{y}\|^2 \text{ for any } \mathbf{x}, \mathbf{y} \in \mathbb{R}^d.$$

Assumption 4. *For $j \in \{1, \dots, N\}$, the function ℓ_j is d-times continuously differentiable.*

Assumption 5. *The stepsizes $(\gamma_k)_{k\geq 0}$ are positive, non-summable and square-summable, i.e.*

$$\sum_{k=0}^{\infty} \gamma_k = \infty, \sum_{k=0}^{\infty} \gamma_k^2 < \infty.$$

Lemma 1 (Convergence to the Feasible Set). *Under A3 and A5, it holds that* $\limsup_{k\to\infty} d(\mathbf{w}^{(k)}, C_\varepsilon) = 0$ *almost surely.*

Proof. See Leconte et al. [1], an asymptotic argument. □

Theorem 1 (Asymptotic Convergence Guarantees with Lipschitz Continuity). *Assume A3, A4, A5 and* $0 < \varepsilon \leq \inf_{1\leq i\leq d} \inf_{1\leq j<K_i} |q_i^{(j)} - q_i^{(j+1)}|^4/16$ *holds, where* $\{q_i^{(j)}\}$ *are the quantization levels,* $\ell(\mathbf{w}^{(k)})$ *converges and* $\lim_{k\to\infty} d(\mathbf{w}^{(k)}, \mathcal{Z}_\varepsilon) = 0$ *almost surely.*

Proof. See Leconte et al. [1]. Sard's Theorem makes A4 necessary for Leconte's proof. □

Remark. A3 is a special case of A2 and it rules out many functions (such as the quadratic function) as a consequence. A4 is non-standard. It represents strong differentiability of the loss function's differentiability, which is not always the case for NNs. Our work is to remove A3 (replaced by A2), A4. We also attempt to investigate the general behavior of `ASkewSGD` without A5.

5 Our Work

We provide four convergence guarantees different from Leconte et al.'s [1] for a modified variant of `ASkewSGD` (on the remoteness measurement by ψ) using deterministic and stochastic oracles. Note that Leconte et al.'s proof still holds in our modified version.

Definition 3 (**Quadratic Remoteness Measurement of the Quantization Set**)**.** *Let $\varepsilon \in (0, 1]$, $w_i \in \mathcal{Q}_i, i \in [d]$, where $\mathcal{Q}_i = \{q_i^{(1)}, \ldots, q_i^{(K_i)}\}$ are sets of quantization values defined coordinate-wise. We define the piecewise function*

$$\psi_i(w_i; \varepsilon) := \begin{cases} \varepsilon - (q_i^{(1)} - w_i) & \text{if } w_i < q_i^{(1)}, \\ \varepsilon - (q_i^{(j-1)} - w_i)(w_i - q_i^{(j)}) & \text{if } q_i^{(j-1)} \leq w_i < q_i^{(j)}, j = 2, \ldots, K_i, \\ \varepsilon - (w_i - q_i^{(K_i)}) & \text{if } w_i \geq q_i^{(K_i)}, \end{cases}$$

for all $w_i \in \mathbb{R}$.

Remark. This function has a lower order compared to Leconte et al.'s (see Definition 2), as we shall see in Lemma 2.

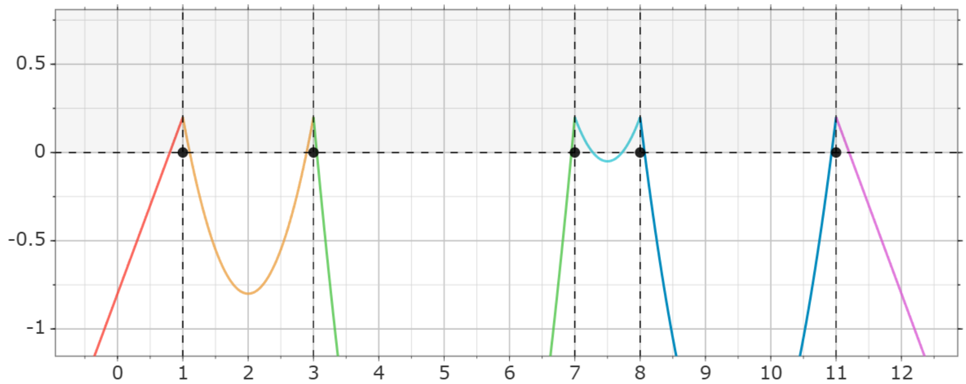

Figure 1: $\psi_i(w_i; \varepsilon)$ where $\varepsilon = 0.2$ and $\mathcal{Q}_i = \{1, 3, 7, 8, 11\}$.

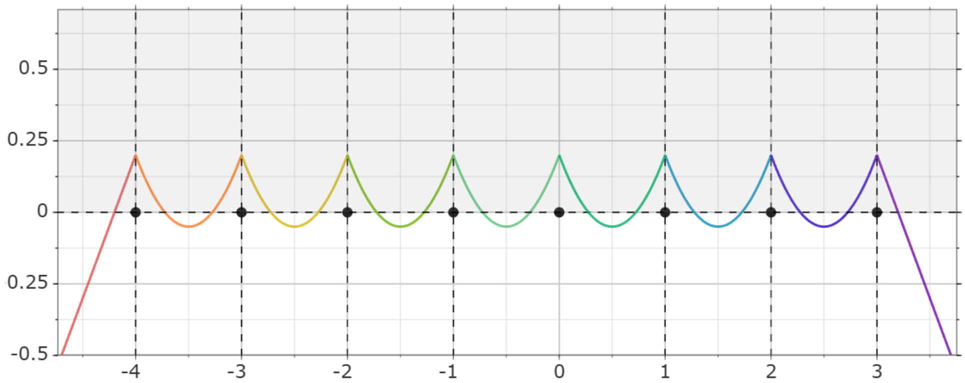

Figure 2: The INT3 uniform quantization scheme, where $\varepsilon = 0.2$ and $\mathcal{Q}_i = \{-4, -3, -2, -1, 0, 1, 2, 3\}$.

Lemma 2 (**A Distance Bound on the Skewing Force**)**.** *Fix an arbitrary* $0 < \varepsilon \leq \inf_{1\leq i\leq d} \inf_{1\leq j<K_i} |q_i^{(j)} - q_i^{(j+1)}|^2/4$*, where* $\{q_i^{(j)}\}$ *are the quantization levels. For* $j \in \{2, \dots, K_i - 1\}$*, denote* $[q_-, q_+]$ *as the set* $C_{\varepsilon,i} \cap [(q_i^{(j)} + q_i^{(j-1)})/2, (q_i^{(j)} + q_i^{(j+1)})/2)$*, where* $C_{\varepsilon,i}$ *is the projection of* C_ε *on the i-th coordinate. Let* $0 < \delta_1 < (q_i^{(j)} + q_i^{(j+1)} - 2q_+)/3$ *and* $0 < \delta_2 < (2q_- - q_i^{(j)} - q_i^{(j-1)})/3$ *be some small perturbations on a quantization level. Then,*

(a)

$$\left|\frac{-\alpha\psi_i(q_+ + \delta_1; \varepsilon)}{\psi_i'(q_+ + \delta_1; \varepsilon)}\right| \leq 2\alpha\delta_1;$$

(b)

$$\left|\frac{-\alpha\psi_i(q_- - \delta_2; \varepsilon)}{\psi_i'(q_- - \delta_2; \varepsilon)}\right| \leq 2\alpha\delta_2.$$

For $j \in \{1, K_i\}$*, denote* $q_- = q_i^{(1)} - \varepsilon, q_+ = q_i^{(K_i)} + \varepsilon$*, and let* $\delta_1 > 0, \delta_2 > 0$*, we have*

$$\left|\frac{-\alpha\psi_i(q_+ + \delta_1; \varepsilon)}{\psi_i'(q_+ + \delta_1; \varepsilon)}\right| = \alpha\delta_1, \left|\frac{-\alpha\psi_i(q_- - \delta_2; \varepsilon)}{\psi_i'(q_- - \delta_2; \varepsilon)}\right| = \alpha\delta_2.$$

Proof. By symmetry, we only need to prove the statement as in (a). Note that

$$\psi_i(q_+; \varepsilon) = \varepsilon - (q_+ - q_i^{(j-1)})(q_i^{(j)} - q_+) = 0 \text{ and } \psi_i'(w; \varepsilon) = 2w - q_i^{(j+1)} - q_i^{(j)}.$$

Using (1), we expand the numerator of the left-hand side of (a) and obtain

$$\begin{aligned}
\psi_i(q_+ + \delta_1; \varepsilon) &= \varepsilon + (q_+ + \delta_1 - q_i^{(j)})(q_+ + \delta_1 - q_i^{(j+1)}) \\
&= \varepsilon + (q_+ - q_i^{(j)})(q_+ - q_i^{(j+1)}) + \delta_1(2q_+ - q_i^{(j+1)} - q_i^{(j)}) + \delta_1^2 \\
&= \delta_1(2q_+ - q_i^{(j+1)} - q_i^{(j)}) + \delta_1^2 < 0.
\end{aligned}$$

Since $\delta_1 > 0$, we have $\psi_i'(q_+ + \delta_1; \varepsilon) = 2q_+ + \delta_1 - q_i^{(j+1)} - q_i^{(j)} < 0$ by (2). Then,

$$\begin{aligned}
\left|\frac{-\alpha\psi_i(q_+ + \delta_1; \varepsilon)}{\psi_i'(q_+ + \delta_1; \varepsilon)}\right| &= \frac{\alpha\delta_1(2q_+ - q_i^{(j)} - q_i^{(j+1)} + \delta_1)}{(2q_+ - q_i^{(j)} - q_i^{(j+1)}) + 2\delta_1} \\
&\leq \frac{\alpha\delta_1(2q_+ - q_i^{(j)} - q_i^{(j+1)})}{(2q_+ - q_i^{(j)} - q_i^{(j+1)}) + 2\delta_1} \\
&= \alpha\delta_1 - \frac{\alpha\delta_1^2}{(2q_+ - q_i^{(j)} - q_i^{(j+1)}) + 2\delta_1}.
\end{aligned}$$

Since $\delta_1 \leq (q_i^{(j)} + q_i^{(j+1)} - 2q_+)/3$, we have

$$\left|\frac{-\alpha\psi_i(q_+ + \delta_1; \varepsilon)}{\psi_i'(q_+ + \delta_1; \varepsilon)}\right| \leq 2\alpha\delta_1.$$

□

Remark. If we have $\sup_{k\geq k_0} \mathrm{d}(\mathbf{w}^{(k)}, C_\varepsilon) \leq \delta$, then $|-\alpha\psi_i(w_i^{(k)}; \varepsilon)/\psi_i'(w_i^{(k)}; \varepsilon)| \leq 2\alpha\delta$ for all $k \geq k_0$.

5.1 Deterministic Full Gradient Convergence Proof

Recall the update

$$[\mathbf{s}_{\varepsilon,\alpha}(\nabla\ell(\mathbf{w}),\mathbf{w})]_i := \begin{cases} -[\nabla\ell(\mathbf{w})]_i & \text{if } \psi_i(w_i;\varepsilon) > 0 \text{ or} \\ & \quad -[\nabla\ell(\mathbf{w})]_i\cdot\psi_i'(w_i;\varepsilon) \geq -\alpha\psi_i(w_i;\varepsilon) \geq 0, \\ \mathrm{clip}(-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon), M_c) & \text{otherwise.} \end{cases}$$

We simplify the "otherwise" direction with the notation $\mathbf{u}$ and refer this to the "skewing force."

Definition 4 (Categorization of the Update Direction). *For the update on the i-th coordinate, we define*

$$\begin{aligned} i \in S_\varepsilon^+(\mathbf{w}) &\iff \psi_i(w_i;\varepsilon) \leq 0 \textit{ and } -\alpha\psi_i(w_i;\varepsilon) > -[\nabla\ell(\mathbf{w})]_i\cdot\psi_i'(w_i;\varepsilon) \geq 0, \\ i \in S_\varepsilon^-(\mathbf{w}) &\iff \psi_i(w_i;\varepsilon) \leq 0 \textit{ and } -\alpha\psi_i(w_i;\varepsilon) \geq 0 > -[\nabla\ell(\mathbf{w})]_i\cdot\psi_i'(w_i;\varepsilon), \\ i \notin S_\varepsilon(\mathbf{w}) := S_\varepsilon^+(\mathbf{w})\cup S_\varepsilon^-(\mathbf{w}) &\iff \psi_i(w_i;\varepsilon) > 0 \textit{ or } -[\nabla\ell(\mathbf{w})]_i\cdot\psi_i'(w_i;\varepsilon) \geq -\alpha\psi_i(w_i;\varepsilon) \geq 0. \end{aligned}$$

5.1.1 Finite Time Convergence with Lipschitz Continuity and Constant Stepsize

Theorem 2 (Finite Time Convergence with Lipschitz Continuity and General Constant Stepsize). *Let $(\gamma_k)_{k\geq 0}$ be a γ-constant ($\gamma < 1/L$) sequence. Under assumption A3, we have*

$$\min_{k=0}^{T}\|\mathbf{v}^{(k)}\|^2 \leq \frac{2(\ell(\mathbf{w}^{(0)}) - \ell(\mathbf{w}^{(T+1)}))}{\gamma(T+1)} + 4\alpha d\gamma L_\ell^2 + 2M_cL_\ell d + \gamma L_\ell M_c^2 d + M_c^2 d.$$

with a convergence rate of $\mathcal{O}(1/T)$ and

$$\min_{k=0}^{T}\|\mathbf{v}^{(k)}\|^2 \to \mathcal{O}(\gamma + M_c d(1+L_\ell)),$$

as T tends to infinity.

Proof. First, apply the descent lemma on the L_ℓ-smooth function ℓ. Then, we discuss the two cases of the update coordinate-wisely. We obtain the following inequality:

$$\begin{aligned} \ell(\mathbf{w}^{(k+1)}) &\leq \ell(\mathbf{w}^{(k)}) + \gamma(\mathbf{v}^{(k)})^\top\nabla\ell(\mathbf{w}^{(k)}) + \frac{\gamma^2L_\ell}{2}\|\mathbf{v}^{(k)}\|^2 \\ &= \ell(\mathbf{w}^{(k)}) + \gamma\sum_{i\in S_\epsilon(\mathbf{w}^{(k)})} u_i^{(k)}[\nabla\ell(\mathbf{w}^{(k)})]_i - \gamma\sum_{i\notin S_\epsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 \\ &\quad + \frac{\gamma^2L_\ell}{2}\sum_{i\in S_\epsilon(\mathbf{w}^{(k)})}(u_i^{(k)})^2 + \frac{\gamma^2L_\ell}{2}\sum_{i\notin S_\epsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^k)]_i^2. \end{aligned}$$

Since $\gamma < 1/L_\ell$, we have $1-\gamma L_\ell/2 > 1/2$, then

$$\ell(\mathbf{w}^{(k+1)}) \leq \ell(\mathbf{w}^{(k)}) - \frac{\gamma}{2}\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 + \gamma\sum_{i\in S_\varepsilon(\mathbf{w}^{(k)})}u_i^{(k)}[\nabla\ell(\mathbf{w}^{(k)})]_i + \frac{\gamma^2L_\ell}{2}\sum_{i\in S_\varepsilon(\mathbf{w}^{(k)})}(u_i^{(k)})^2.$$

Since the gradient of ℓ is bounded by L_ℓ and $|u_i^{(k)}|$ is bounded by M_c, we have

$$\begin{aligned}\ell(\mathbf{w}^{(k+1)}) &\le \ell(\mathbf{w}^{(k)}) - \frac{\gamma}{2}\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 + \gamma\sum_{i\in S_\epsilon(\mathbf{w}^{(k)})} M_c L_\ell + \frac{\gamma^2 L_\ell}{2}\sum_{i\in S_\varepsilon(\mathbf{w}^{(k)})} M_c^2 \\ &\le \ell(\mathbf{w}^{(k)}) - \frac{\gamma}{2}\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 + \gamma M_c L_\ell d + \frac{\gamma^2 L_\ell M_c^2 d}{2}.\end{aligned}$$

Rearrange the inequality and summing over $T+1$ epochs from 0 to T, we obtain

$$\frac{\gamma}{2}\sum_{k=0}^{T}\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 \le \ell(\mathbf{w}^{(0)}) - \ell(\mathbf{w}^{(T+1)}) + (T+1)\gamma M_c L_\ell d + (T+1)\frac{\gamma^2 L_\ell M_c^2 d}{2}.$$

Completing the entire update direction with the skewing force, we can bound $\|\mathbf{v}^{(k)}\|$ as we have

$$\begin{aligned}\frac{\gamma}{2}\sum_{k=0}^{T}\left(\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 + \sum_{i\in S_\varepsilon(\mathbf{w}^{(k)})}(u_i^{(k)})^2\right) &\le \ell(\mathbf{w}^{(0)}) - \ell(\mathbf{w}^{(T+1)}) + 2(T+1)\alpha d\gamma^2 L_\ell^2 \\ &\quad + (T+1)\gamma M_c L_\ell d + (T+1)\frac{\gamma^2 L_\ell M_c^2 d}{2} + \frac{\gamma}{2}(T+1)M_c^2 d, \\ \min_{k=0}^{T}\left(\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 + \sum_{i\in S_\varepsilon(\mathbf{w}^{(k)})}(u_i^{(k)})^2\right) &\le \frac{2(\ell(\mathbf{w}^{(0)}) - \ell(\mathbf{w}^{(T+1)}))}{\gamma(T+1)} + 4\alpha d\gamma L_\ell^2 \\ &\quad + 2M_c L_\ell d + \gamma L_\ell M_c^2 d + M_c^2 d.\end{aligned}$$

Finally,

$$\min_{k=0}^{T}\|\mathbf{v}^{(k)}\|^2 \le \frac{2(\ell(\mathbf{w}^{(0)}) - \ell(\mathbf{w}^{(T+1)}))}{\gamma(T+1)} + 4\alpha d\gamma L_\ell^2 + 2M_c L_\ell d + \gamma L_\ell M_c^2 d + M_c^2 d.$$

As T tends to infinity, we have

$$\min_{k=0}^{T}\|\mathbf{v}^{(k)}\|^2 \to \mathcal{O}(\gamma + M_c d(1+L_\ell)),$$

with convergence rate of $\mathcal{O}(1/T)$. □

Why is an extra upper bound on γ necessary? Consider the case where $\mathcal{Q} = \{\pm 1\}$ and ℓ is decreasing away from $(-1,1)$. Let $[q_-, q_+] = \{w : \psi(w;\varepsilon) \ge 0\}$. Let $w < q_-$. There exists $\gamma = 2w/(\alpha(\varepsilon + w + 1)))$ such that w enters a cycle.

To see this, first note that $\psi(w;\varepsilon) = \varepsilon - (-1-w) < 0$. We compute that

$$u = \frac{-\alpha\psi(w)}{\psi'(w)} = -\alpha(\varepsilon + w + 1).$$

Then, $w' = w + \gamma u = w - \gamma\cdot\alpha(\varepsilon + w + 1)) = w - 2w = -w$. By symmetry, we have $w'' = -w' = w$.

Why do we have an extra M_c term? We cannot be sure whether $u_i^{(k)}$ converges to $C_{\varepsilon,i}$. Large gradient could disrupt the distance to feasible set for general γ.

In the sequel, we introduce three lemmata that provides non-asymptotic guarantees for ASkewSGD under Assumption A3 and a small enough step size γ (which will be regulated by the lemmata).

Lemma 3 (**Confinement to the Quantization Interval, General Case**). *Assume that A3 holds. Fix an* $i \in [d]$ *and* $2 \le j \le K_i - 1$. *Let* $[q_-, q_+] = C_{\varepsilon,i} \cap [(q_i^{(j-1)} + q_i^{(j)})/2, (q_i^{(j)} + q_i^{(j+1)})/2)$, *where* $C_{\varepsilon,i}$ *is the projection of* C_ε *on the* i*-th coordinate, and* $0 < \gamma \cdot \max\{L_\ell, M_c\} < \min\{q_+ - q_-, q_- - (q_i^{(j-1)} + q_i^{(j)})/2, (q_i^{(j-1)} + q_i^{(j)})/2 - q_+\}$. *If there is a moment* k_0 *such that* $w_i^{(k_0)} \in ((q_i^{(j)} + q_i^{(j-1)})/2, (q_i^{(j)} + q_i^{(j+1)})/2)$, *then* $w_i^{(k)}$ *is confined to the interval* $((q_i^{(j)} + q_i^{(j-1)})/2, (q_i^{(j)} + q_i^{(j+1)})/2)$ *for every* $k \ge k_0$.

Proof. Notice for any $k > 0$, $\|\nabla \ell(\mathbf{w}^{(k)})\| \le L_\ell$ and $|v_i^{(k)}| \le \max\{L_\ell, M_c\}$. If $(q_i^{(j)} + q_i^{(j-1)})/2 \le w_i^{(k_0)} < q_-$, then $w_i^{(k_0)}$ is pushed to the right. We have $w_i^{(k_0)} \le w_i^{(k_0+1)} := w_i^{(k_0)} + \gamma \max\{L_\ell, M_c\} < q_+$. If $q_+ < w_i^{(k)} < (q_i^{(j)} + q_i^{(j+1)})/2$, then $w_i^{(k_0)}$ is pushed to the left, and $w_i^{(k_0)} \ge w_i^{(k_0+1)} := w_i^{(k_0)} - \gamma \max\{M_\ell, M_c\} \le q_-$. Finally, if $w_i^{(k_0)} \in [q_-, q_+]$, then $(q_i^{(j-1)} + q_i^{(j)})/2 < w_i^{(k_0)} - \gamma L_\ell \le w_i^{(k_0+1)} \le w_i^{(k_0)} + \gamma L_\ell < (q_i^{(j)} + q_i^{(j+1)})/2$. The same argument is inductively true for all $k \ge k_0$, thus concluding the proof. □

Lemma 4 (**Confinement to the Quantization Interval, Edge Case**). *Assume that A3 holds. Fix an* $i \in [d]$. *Let* $[q_-, q_+] = C_{\varepsilon,i} \cap (-\infty, (q_i^{(1)} + q_i^{(2)})/2)$, *where* $C_{\varepsilon,i}$ *is the projection of* C_ε *on the* i*-th coordinate, and* $0 < \gamma \cdot \max\{L_\ell, M_c\} < \min\{q_+ - q_-, (q_i^{(1)} + q_i^{(2)})/2 - q_+\}$. *If there is a moment* k_0 *such that* $w_i^{(k_0)} \in (-\infty, (q_i^{(1)} + q_i^{(2)})/2)$, *then* $w_i^{(k)}$ *is confined to the interval* $(-\infty, (q_i^{(1)} + q_i^{(2)})/2)$ *for every* $k \ge k_0$. *The same conclusion holds for* $[q_-, q_+] = C_{\varepsilon,i} \cap ((q_i^{(K_i-1)} + q_i^{(K_i)})/2, +\infty)$ *and* $0 < \gamma \cdot \max\{L_\ell, M_c\} < \min\{q_+ - q_-, (q_i^{(K_i-1)} + q_i^{(K_i)})/2 - q_-\}$ *with* $w_i^{(k)}$ *being confined to the interval* $((q_i^{(K_i-1)} + q_i^{(K_i)})/2, +\infty)$ *for every* $k \ge k_0$.

Proof. Similar to Lemma 3. If $w_i^{(k_0)} < q_-$, then $w_i^{(k_0)}$ is pushed to the right. We have $w_i^{(k_0)} \le w_i^{(k_0+1)} := w_i^{(k_0)} + \gamma \max\{L_\ell, M_c\} < q_+$. If $w_i^{(k_0)} \in [q_-, q_+]$, then $w_i^{(k_0+1)} \le w_i^{(k_0)} + \gamma L_\ell < (q_i^{(1)} + q_i^{(2)})/2$. □

Lemma 5 (**Convergence to Neighborhood of The Feasible Region**). *Assume that A3 holds. Let* γ *be such that each coordinate* i *of* $\mathbf{w}^{(k_0)}$ *is confined to an interval* I_i *as in Lemma 3 and 4. Let* $\tau = \min_i \inf_{\mathrm{d}(w_i, C_{\varepsilon,i}) \ge \gamma L_\ell} \{\mathrm{clip}(|\alpha \psi_i(w_i; \varepsilon)|/|\psi_i'(w_i; \varepsilon)|, M_c)\}$. *Then,* $\mathrm{d}(\mathbf{w}^{(k)}, C_\varepsilon) \le \gamma L_\ell \sqrt{d}$ *for all* $k \ge k_0$ $+\lceil (\mathrm{d}(\mathbf{w}^{(k_0)}, C_\varepsilon) - \gamma L_\ell)/(\eta\gamma) \rceil$.

Proof. Consider the case where $\mathbf{w}_i^{(k)} \in C_{\varepsilon,i}$. After an update, $\mathrm{d}(\mathbf{w}^{(k+1)}, C_\varepsilon) \le \sqrt{\sum_i (\gamma_k L_\ell)^2} = \gamma_k L_\ell \sqrt{d}$. Otherwise, the skewing force activates at $k+1$. For each infeasible coordinate i, the skewing correction is at least $\tau\gamma$ by definition (since the skewing force diminishes as $w_i^{(k)}$ gets closer to the feasible region, and the gradient pushes at least as strong as the skewing force). Hence, $\mathrm{d}_i(w_i^{(k+1)}, C_{\varepsilon,i}) \le \mathrm{d}_i(w_i^{(k)}, C_{\varepsilon,i}) - \tau\gamma$. Therefore, in at most an extra $\lceil (\mathrm{d}(\mathbf{w}^{(k_0)}, C_\varepsilon) - \gamma L_\ell)/(\tau\gamma) \rceil$ steps, we can guarantee that the distance to the feasible set is at an order of $\mathcal{O}(\gamma)$. □

Corollary 1 (**Bounding Intervals and Selection of Stepsize**). *Assume that A3 holds. Then, for any* $\mathbf{w}^{(k_0)}$*, there exists* γ *small enough and* k_1 *such that* $\mathrm{d}(\mathbf{w}^{(k)}, C_\varepsilon) \leq \gamma L_\ell \sqrt{d} =: \delta$ *and* $|u_i^{(k)}| \leq 2\alpha\delta$ *for all* $k \geq k_1 \geq k_0$ *and* $i \in [d]$*.*

Proof. Let $M = \{L_\ell, M_c\}$ and $\tau = \min_i \inf_{d(w_i, C_{\varepsilon,i}) \geq \gamma L_\ell} \{clip(|\alpha\psi_i(w_i;\varepsilon)|/|\psi_i'(w_i;\varepsilon)|, M_c)\}$. Consider the i-th coordinate of $\mathbf{w}^{(k)}$. There are three cases to consider:

(a) If $w_i^{(k)} < (q_i^{(1)} + q_i^{(2)})/2$, then let $[a_i, b_i] = C_{\varepsilon,i} \cap [-\infty, (q_i^{(1)} + q_i^{(2)})/2))$, we set $\eta_i := \sup\{\gamma : \gamma M < \min\{q_+ - q_-, (q_i^{(1)} + q_i^{(2)})/2 - q_+\}\}$;

(b) If $w_i^{(k)} \geq (q_i^{(K_i-1)} + q_i^{(K_i)})/2)$, then let $[a_i, b_i] = C_{\varepsilon,i} \cap [(q_i^{(K_i-1)} + q_i^{(K_i)})/2), +\infty)$, we set $\eta_i := \sup\{\gamma : \gamma M < \min\{q_+ - q_-, q_- - (q_i^{(K_i-1)} + q_i^{(K_i)})/2\}\}$;

(c) If $(q_i^{(j-1)} + q_i^{(j)})/2 \leq w_i^{(k)} < (q_i^{(j)} + q_i^{(j+1)})/2$, then let $[a_i, b_i] = C_{\varepsilon,i} \cap [(q_i^{(j-1)} + q_i^{(j)})/2, (q_i^{(j)} + q_i^{(j+1)})/2)$, we set $\eta_i := \min\{\sup\{\gamma : \gamma M < \min\{q_+ - q_-, q_- - (q_i^{(j-1)} + q_i^{(j)})/2, (q_i^{(j-1)} + q_i^{(j)})/2 - q_+\}\}, (q_i^{(j)} + q_i^{(j+1)} - 2q_+)/3, (2q_- - q_i^{(j)} - q_i^{(j-1)})/3\}$.

Then, the proof is immediately concluded by considering the selection of any γ that satisfies $0 < \gamma < \min_i \eta_i$ and $k_1 := k_0 + \lceil (\mathrm{d}(\mathbf{w}^{(k_0)}, C_\varepsilon) - \gamma L_\ell)/(\tau\gamma) \rceil$, which in fact satisfy Lemma 3, 4 and 5. $\square$

Theorem 3 (**Finite Time Convergence with Lipschitz Continuity and Small Constant Stepsize**). *Let* $(\gamma_k)_{k\geq 0}$ *be a* γ*-constant sequence, where* $\gamma < 1/L_\ell$ *and* k_1 *are given as in Corollary 1. Then, under assumption A3, we have*

$$\min_{k=k_1}^{k_1+T} \|\mathbf{v}^{(k)}\|^2 \leq \frac{2(\ell(\mathbf{w}^{(k_1)}) - \ell(\mathbf{w}^{(k_1+T+1)}))}{\gamma(T+1)} + 4\alpha d\gamma L_\ell^2 + 4\alpha^2 d\gamma^3 L_\ell^3 + 8\alpha^2 d\gamma L_\ell^2.$$

with a convergence rate of $\mathcal{O}(1/T)$ *and*

$$\min_{k=k_1}^{k_1+T} \|\mathbf{v}^{(k)}\|^2 \to \mathcal{O}(\gamma),$$

as T *tends to infinity.*

Proof. Starting from the descent inequality obtained previously, we have

$$\ell(\mathbf{w}^{(k+1)}) \leq \ell(\mathbf{w}^{(k)}) - \frac{\gamma}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \gamma \sum_{i \in S_\varepsilon(\mathbf{w}^{(k)})} u_i^{(k)} [\nabla \ell(\mathbf{w}^{(k)})]_i + \frac{\gamma^2 L_\ell}{2} \sum_{i \in S_\varepsilon(\mathbf{w}^{(k)})} (u_i^{(k)})^2.$$

Since the gradient of ℓ is bounded by L_ℓ and $|u_i^{(k)}| \leq 2\alpha\gamma L_\ell \sqrt{d}$ as a consequence, we have

$$\begin{aligned}
\ell(\mathbf{w}^{(k+1)}) &\leq \ell(\mathbf{w}^{(k)}) - \frac{\gamma}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \gamma \sum_{i \in S_\epsilon(\mathbf{w}^{(k)})} (2\alpha\gamma L_\ell) L_\ell + \frac{\gamma^2 L_\ell}{2} \sum_{i \in S_\varepsilon(\mathbf{w}^{(k)})} (2\alpha\gamma L_\ell)^2 \\
&\leq \ell(\mathbf{w}^{(k)}) - \frac{\gamma}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \gamma(2\alpha\gamma L_\ell) L_\ell d + \frac{\gamma^2 L_\ell}{2} (2\alpha\gamma L_\ell)^2 d \\
&\leq \ell(\mathbf{w}^{(k)}) - \frac{\gamma}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + 2\alpha d\gamma^2 L_\ell^2 + 2\alpha^2 d\gamma^4 L_\ell^3.
\end{aligned}$$

Rearrange the inequality and summing over $T+1$ epochs from k_1 to k_1+T, we obtain

$$\frac{\gamma}{2}\sum_{k=k_1}^{k_1+T}\sum_{i\notin T_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2 \leq \ell(\mathbf{w}^{(k_1)})-\ell(\mathbf{w}^{(k_1+T+1)})+2(T+1)\alpha d\gamma^2L_\ell^2+2(T+1)\alpha^2d\gamma^4L_\ell^3.$$

Completing the entire update direction with the skewing force, we can bound $\|\mathbf{v}^{(k)}\|$ as we have

$$\begin{aligned}\frac{\gamma}{2}\sum_{k=k_1}^{k_1+T}\left(\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2+\sum_{i\in S_\varepsilon(\mathbf{w}^{(k)})}(u_i^{(k)})^2\right) &\leq \ell(\mathbf{w}^{(k_1)})-\ell(\mathbf{w}^{(k_1+T+1)})+2(T+1)\alpha d\gamma^2L_\ell^2\\ &\quad+2(T+1)\alpha^2d\gamma^4L_\ell^3+4(T+1)\alpha^2d\gamma^2L_\ell^2\\ \min_{k=k_1}^{k_1+T}\left(\sum_{i\notin S_\varepsilon(\mathbf{w}^{(k)})}[\nabla\ell(\mathbf{w}^{(k)})]_i^2+\sum_{i\in S_\varepsilon(\mathbf{w}^{(k)})}(u_i^{(k)})^2\right) &\leq \frac{2(\ell(\mathbf{w}^{(k_1)})-\ell(\mathbf{w}^{(k_1+T+1)}))}{\gamma(T+1)}+4\alpha d\gamma L_\ell^2\\ &\quad+4\alpha^2d\gamma^3L_\ell^3+8\alpha^2d\gamma L_\ell^2.\end{aligned}$$

Finally,

$$\min_{k=k_1}^{k_1+T}\|\mathbf{v}^{(k)}\|^2 \leq \frac{2(\ell(\mathbf{w}^{(k_1)})-\ell(\mathbf{w}^{(k_1+T+1)}))}{\gamma(T+1)}+4\alpha d\gamma L_\ell^2+4\alpha^2d\gamma^3L_\ell^3+8\alpha^2d\gamma L_\ell^2.$$

As T tends to infinity, we have

$$\min_{k=k_1}^{k_1+T}\|\mathbf{v}^{(k)}\|^2 \to \mathcal{O}(\gamma),$$

with convergence rate of $\mathcal{O}(1/T)$. □

Corollary 2 (**Time Horizon Dependent Convergence with Lipschitz Continuity and Small Constant Stepsize**). *Given a sufficiently long time horizon T, there exists $\gamma := 1/\sqrt{T+1}$ (with $\gamma < 1/L_\ell$) and k_1 satisfying the conditions as in Corollary 1. Furthermore, under assumption A3, we have*

$$\min_{k=k_1}^{k_1+T}\|\mathbf{v}^{(k)}\|^2 \leq \frac{2(\ell(\mathbf{w}^{(k_1)})-\ell(\mathbf{w}^{(k_1+T+1)}))}{\sqrt{T+1}}+\frac{4\alpha dL_\ell^2}{\sqrt{T+1}}+\frac{4\alpha^2dL_\ell^3}{\sqrt{(T+1)^3}}+\frac{8\alpha^2dL_\ell^2}{\sqrt{T+1}}.$$

with a convergence rate of $\mathcal{O}(1/\sqrt{T})$ and

$$\min_{k=k_1}^{k_1+T}\|\mathbf{v}^{(k)}\|^2 \to 0,$$

as T tends to infinity.

Proof. Immediate from Theorem 3. □

Significance. We have derived a finite-time non-asymptotic convergence bound for constant stepsize full-gradient ASkewSGD on the update $\mathbf{v}^{(k)}$ (which is an indicator for stationarity), with removal of the strong differentiability assumption A4.

Implications. For $\gamma < 1/L_\ell$, we can bound the update due to bounded gradient and skewing forces. However, if we further restrict on γ, then the distance will become an upper bound for the skewing update and force the parameter to converge to the (γ-neighborhood of the) quantization values. We could even force the update to diminish to zero by setting a time-dependent step size schedule (as a balance between the convergence rate and bias).

5.1.2 Asymptotic Convergence with Lipschitz Smoothness and Robbins-Monro Stepsizes

Theorem 4 (**Asymptotic Convergence with Lipschitz Smoothness and Robbins-Monro Stepsizes**). *Let $\gamma_k < 1/L$ for all $k \geq k_0$. Under assumptions A2 and A5, we have*

$$\frac{1}{\sum_{k=k_0}^{T} \gamma_k} \sum_{k=k_0}^{T} \frac{\gamma_k}{2} \|\mathbf{v}^{(k)}\|^2 \longrightarrow \mathcal{O}(M_c^2 d), \qquad \textit{as } T \textit{ tends to infinity.}$$

Proof. Note that when $i \in S_\varepsilon^+(\mathbf{w})$, $|u_i^{(k)}| \geq |[\nabla \ell(\mathbf{w}^{(k)})]_i|$.

Starting from the descent inequality obtained previously, we have

$$\begin{aligned}
\ell(\mathbf{w}^{(k+1)}) \leq \ell(\mathbf{w}^{(k)}) &- \frac{\gamma_k}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \gamma_k \sum_{i \in S_\varepsilon^+(\mathbf{w}^{(k)})} u_i^{(k)} [\nabla \ell(\mathbf{w}^{(k)})]_i \\
&+ \gamma_k \sum_{i \in S_\varepsilon^-(\mathbf{w}^{(k)})} u_i^{(k)} [\nabla \ell(\mathbf{w}^{(k)})]_i + \frac{\gamma_k^2 M_\ell}{2} \sum_{i \in S_\varepsilon(\mathbf{w}^{(k)})} (u_i^{(k)})^2 \\
\leq \ell(\mathbf{w}^{(k)}) &- \frac{\gamma_k}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \gamma_k \sum_{i \in S_\epsilon^+(\mathbf{w}^{(k)})} |u_i^{(k)}|^2 \\
&- \gamma_k \sum_{i \in S_\epsilon^-(\mathbf{w}^{(k)})} |u_i^{(k)}| |[\nabla \ell(\mathbf{w}^{(k)})]_i| + \frac{\gamma_k^2 M_\ell}{2} \sum_{i \in S_\varepsilon(\mathbf{w}^{(k)})} M_c^2 \\
\leq \ell(\mathbf{w}^{(k)}) &- \frac{\gamma_k}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \gamma_k \sum_{i \in S_\epsilon^+(\mathbf{w}^{(k)})} |u_i^{(k)}|^2 + \frac{\gamma_k^2 M_\ell M_c^2 d}{2}.
\end{aligned}$$

Completing the entire update direction with the skewing force, we bound $\|\mathbf{v}^{(k)}\|^2$ by

$$\begin{aligned}
\frac{\gamma_k}{2} \|\mathbf{v}^{(k)}\|^2 &= \frac{\gamma_k}{2} \sum_{i \notin S_\varepsilon(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \frac{\gamma_k}{2} \sum_{i \in S_\varepsilon(\mathbf{w}^{(k)})} |u_i^{(k)}|^2 \\
&\leq \ell(\mathbf{w}^{(k)}) - \ell(\mathbf{w}^{(k+1)}) + \frac{3\gamma_k}{2} \sum_{i \in S_\epsilon(\mathbf{w}^{(k)})} |u_i^{(k)}|^2 + \frac{\gamma_k^2 M_\ell M_c^2 d}{2} \\
&\leq \ell(\mathbf{w}^{(k)}) - \ell(\mathbf{w}^{(k+1)}) + \frac{3\gamma_k d M_c^2}{2} + \frac{\gamma_k^2 M_\ell M_c^2 d}{2}.
\end{aligned}$$

Summing from epoch k_0 to T, we have

$$\begin{aligned}
\sum_{k=k_0}^{T} \frac{\gamma_k}{2} \|\mathbf{v}^{(k)}\|^2 &\leq \ell(\mathbf{w}^{(k_0)}) - \ell(\mathbf{w}^{(K+1)}) + \frac{3dM_c^2}{2} \sum_{k=k_0}^{T} \gamma_k + \frac{M_\ell M_c^2 d}{2} \sum_{k=k_0}^{T} \gamma_k^2 \\
\frac{1}{\sum_{k=k_0}^{T} \gamma_k} \sum_{k=k_0}^{T} \frac{\gamma_k}{2} \|\mathbf{v}^{(k)}\|^2 &\leq \frac{\ell(\mathbf{w}^{(k_0)}) - \ell(\mathbf{w}^{(T+1)})}{\sum_{k=k_0}^{T} \gamma_k} + \frac{3dM_c^2}{2} + \frac{M_\ell M_c^2 d}{2} \frac{\sum_{k=k_0}^{T} \gamma_k^2}{\sum_{k=k_0}^{T} \gamma_k}.
\end{aligned}$$

As T tends to infinity, we have

$$\frac{1}{\sum_{k=k_0}^{T}\gamma_k}\sum_{k=k_0}^{T}\frac{\gamma_k}{2}\|\mathbf{v}^{(k)}\|^2 \longrightarrow \mathcal{O}(M_c^2 d).$$

□

5.2 Convergence Analysis under Stochasticity

Let $\pi(\cdot)$ be a probability density function defined on the probability space $\mathtt{Z}$ and ξ be a random parameter, such that

$$\ell(\mathbf{x}) = \int_{\mathtt{Z}} \ell(\mathbf{x};\xi)\pi(\xi)\mathrm{d}\xi.$$

Assumption 6. *The stochastic gradient is unbiased, i.e.*

$$\mathbb{E}_{\xi\sim\pi}[\widehat{\nabla}\ell(\mathbf{x};\xi)] = \nabla\ell(\mathbf{x}), \forall \mathbf{x}\in\mathbb{R}^n.$$

Assumption 7. *The stochastic gradient has a bounded variance, i.e.*

$$\mathbb{E}_{\xi\sim\pi}\left[\left\|\widehat{\nabla}\ell(\mathbf{x};\xi) - \nabla\ell(\mathbf{x})\right\|_2^2\right] \le \sigma^2, \forall \mathbf{x}\in\mathbb{R}^n.$$

Recall the (stochastic version of) update

$$[\hat{\mathbf{s}}_{\varepsilon,\alpha}(\hat{\mathbf{g}},\mathbf{w})]_i := \begin{cases} -\hat{g}_i & \text{if } \psi_i(w_i;\varepsilon) > 0 \text{ or} \\ & \quad -\hat{g}_i\cdot\psi_i'(w_i;\varepsilon) \ge -\alpha\psi_i(w_i;\varepsilon) \ge 0, \\ \mathrm{clip}(-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon), M_c) & \text{otherwise.} \end{cases}$$

We simplify the "otherwise" direction with the notation $\mathbf{u}$ and refer this to the "skewing force."

Remark. For the stochastic case, we do not have full information for the true update direction and thus we need to discuss in total of four possible matches between the perturbed direction and the true update direction given by the true gradient. Furthermore, the magnitude of the skewing force does not depend on the (possibly perturbed) gradient.

Definition 5 (**Categorization of the Update Direction**)**.** *Consider the update in each iteration k, $\hat{\mathbf{v}}^{(k)}$, we categorize the update direction $\hat{\mathbf{v}}^{(k)}$ into four types (with the indices collected by $\hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)})$, $\hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})$, $\hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)})$, $\hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)})$ and $\hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)})\cup\hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})\cup\hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)})\cup\hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)}) = [d]$). Formally, for the update on i-th coordinate*

$$i\in\hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \Longleftrightarrow \hat{v}_i^{(k)} = -[\widehat{\nabla}\ell(\mathbf{w}^{(k)})]_i \textit{ and } v_i^{(k)} = -[\nabla\ell(\mathbf{w}^{(k)})]_i.$$
$$i\in\hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)}) \Longleftrightarrow \hat{v}_i^{(k)} = -[\widehat{\nabla}\ell(\mathbf{w}^{(k)})]_i \textit{ and } \mathbf{v}_i^{(k)} = u_i^{(k)}.$$
$$i\in\hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)}) \Longleftrightarrow \hat{v}_i^{(k)} = \hat{u}_i^{(k)} = v_i^{(k)} = u_i^{k}.$$
$$i\in\hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)}) \Longleftrightarrow \hat{v}_i^{(k)} = \hat{u}_i^{(k)} \textit{ and } v_i^{(k)} = -[\nabla\ell(\mathbf{w}^{(k)})]_i.$$

Remark. Our goal is to bound the true gradient. We know that the type 1 update is standard. The type 2 and 3 updates are captured by the fact that $\mathbf{u}$ can be infinitely small as δ progresses to zero. The type 4 update is tricky - as it represents a trade-off between the quantization value and decrease of function value. We alleviate this situation by hinging on the fact that the stationary point is reached after projecting the $\mathbf{w}$ to the quantization set on these coordinates (while the distance to this quantization value is merely δ). We allow (only) gradients on these coordinates to be non-zero because our problem is a constrained optimization problem.

5.2.1 Asymptotic Convergence with Lipschitz Continuity and Robbins-Monro Stepsizes

Idea. We first construct a bound for $[\nabla \ell(\mathbf{w}_\tau)]_i$ summed over $i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})$ (note that $\mathbf{w}_\tau$ is a randomized variable). We then use Lemma 2 again for the gradient at coordinates in $\hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)})$. Coordinates in $\hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)})$ is justified by the smoothness of the gradient and the algorithm's special dynamics.

Theorem 5 (**Asymptotic Convergence with Lipschitz Continuity and Robbins-Monro Stepsizes**)**.** *Assuming that A3, A5, A6, A7,* $0 < \varepsilon \leq \inf_{1\leq i \leq d} \inf_{1 \leq j < K_i} |q_i^{(j)} - q_i^{(j+1)}|^2/4$ *holds, where* $\{q_i^{(j)}\}$ *are the quantization levels, then (i) there is a* k_0 *with* $\forall k \geq k_0, \gamma_{k_0} \leq \min\{1, 1/L_\ell\}, \mathrm{d}(\mathbf{w}^{(k)}, C_\varepsilon) \leq \delta, \|\mathbf{u}^{(k)}\| \leq 2\alpha\delta$*; (ii)* $\lim_{K\to\infty} d(w^{(\tau_K)}, \mathcal{Z}_\varepsilon) \to 0$ *almost surely, where* τ_K *is a randomized index at iteration* K *taking value from* k_0 *to* K*, generated with the probability distribution proportional to the stepsizes* $(\gamma_k)_{k_0 \leq k \leq K}$.

Proof. The first statement is justified by assumptions A3, A5 and Lemma 1, 2.

We focus on the second statement. Notice that we can break down the gradient norm coordinate-wisely. The first part of the second statement follows from the standard SGD analysis. First, apply the descent lemma on the L_ℓ-smooth function ℓ with care. We have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{w}^{(k)}}[\ell(\mathbf{w}^{(k+1)})] &\leq \ell(\mathbf{w}^{(k)}) + \gamma_k \mathbb{E}_{\mathbf{w}^{(k)}}[(\hat{\mathbf{v}}^{(k)})^\top] \nabla \ell(\mathbf{w}^{(k)}) + \frac{\gamma_k^2 L_\ell}{2} \mathbb{E}_{\mathbf{w}^{(k)}}[\|\hat{\mathbf{v}}^{(k)}\|^2] \\
&= \ell(\mathbf{w}^{(k)}) - \gamma_k \sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \frac{\gamma_k^2 L_\ell}{2} \mathbb{E}_{\mathbf{w}^{(k)}} \left[\sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\widehat{\nabla} \ell(\mathbf{w}^{(k)})]_i^2 \right] \\
&\quad + \gamma_k \sum_{i \in \hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)})} u_i^{(k)} [\nabla \ell(\mathbf{w}^{(k)})]_i + \frac{\gamma_k^2 L_\ell}{2} \sum_{i \in \hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)})} (u_i^{(k)})^2 \\
&\leq \ell(\mathbf{w}^{(k)}) - \gamma_k \sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 \\
&\quad + \frac{\gamma_k^2 L_\ell}{2} \left(\sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \sigma^2 \right) \\
&\quad + \gamma_k \sum_{i \in \hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)})} 2\alpha\delta L_\ell + \frac{\gamma_k^2 L_\ell}{2} \sum_{i \in \hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)})} 4\alpha^2\delta^2
\end{aligned}
$$

$$\leq \ell(\mathbf{w}^{(k)}) - \frac{\gamma_k}{2} \sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2 + \frac{L_\ell \gamma_k^2 \sigma^2}{2} + 2dL_\ell \gamma_k \alpha \delta + 2dL_\ell \gamma_k^2 \alpha^2 \delta^2,$$

where we have assumed $1 - \gamma_{k_0} L_\ell / 2 > 1/2$ in the last step.

Now, we can take full expectation with constant $\mathsf{a} \geq 2dL_\ell \alpha + 2d\alpha^2 \delta \geq 2dL_\ell(\alpha + \gamma_k \alpha^2 \delta)$ and derive

$$\begin{aligned}\frac{\gamma_k}{2} \mathbb{E}\left[\sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2\right] &\leq \mathbb{E}[\ell(\mathbf{w}^{(k)})] - \mathbb{E}[\ell(\mathbf{w}^{(k+1)})] + \frac{L_\ell \gamma_k^2 \sigma^2}{2} + 2dL_\ell \gamma_k \alpha \delta + 2dL_\ell \gamma_k^2 \alpha^2 \delta^2 \\ &\leq \mathbb{E}[\ell(\mathbf{w}^{(k)})] - \mathbb{E}[\ell(\mathbf{w}^{(k+1)})] + \frac{L_\ell \gamma_k^2 \sigma^2}{2} + \mathsf{a} \gamma_k \delta.\end{aligned}$$

Summing from epoch k_0 to K, we have

$$\sum_{k=k_0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[\sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2\right] \leq \left(\mathbb{E}[\ell(\mathbf{w}^{(k_0)})] - \mathbb{E}[\ell(\mathbf{w}^{(K+1)})]\right) + \frac{L_\ell \sigma^2}{2} \sum_{k=k_0}^{K} \gamma_k^2 + \mathsf{a}\delta \sum_{k=k_0}^{K} \gamma_k.$$

Let $H_K = \sum_{k=k_0}^{K} \gamma_k$ and hence $\tau_K = k$ with probability γ_k / H_K, we can take the expectation of the randomized squared-gradient:

$$\mathbb{E}\left[[\nabla \ell(\mathbf{w}^{(\tau_K)})]_i^2\right] = \sum_{k=k_0}^{K} \mathbb{P}(\tau_K = k) \mathbb{E}\left[[\nabla \ell(\mathbf{w}^{(k)})]_i^2\right] = \sum_{k=k_0}^{K} \frac{\gamma_k \mathbb{E}\left[[\nabla \ell(\mathbf{w}^{(k)})]_i^2\right]}{H_K} = \frac{2}{H_K} \sum_{k=k_0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[[\nabla \ell(\mathbf{w}^{(k)})]_i^2\right].$$

$$\begin{aligned}\mathbb{E}\left[\sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(\tau_K)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(\tau_K)})} [\nabla \ell(\mathbf{w}^{(\tau_K)})]_i^2\right] &= \frac{2}{H_K} \sum_{k=k_0}^{K} \frac{\gamma_k}{2} \mathbb{E}\left[\sum_{i \in \hat{S}_{1,\varepsilon}(\mathbf{w}^{(k)}) \cup \hat{S}_{2,\varepsilon}(\mathbf{w}^{(k)})} [\nabla \ell(\mathbf{w}^{(k)})]_i^2\right] \\ &\leq \frac{2}{H_K} \left(\mathbb{E}[\ell(\mathbf{w}^{(k_0)})] - \mathbb{E}[\ell(\mathbf{w}^{(K+1)})]\right) + \frac{L_\ell \sigma^2}{H_K} \sum_{k=k_0}^{K} \gamma_k^2 + \frac{2\mathsf{a}\delta}{H_K} \sum_{k=k_0}^{K} \gamma_k \\ &\overset{K \to \infty}{\longrightarrow} 2\mathsf{a}\delta,\end{aligned}$$

since assumption 5 ensures that $\sum_{k=k_0}^{\infty} \gamma_k = \infty, \sum_{k=k_0}^{\infty} \gamma_k^2 < \infty$.

In the second part of the second statement, we consider the case where $i \in \hat{S}_{3,\varepsilon}(\mathbf{w}^{(k)})$. Note that the true update direction is $v_i^{(k)} = u_i^{(k)} \leq 2\alpha\delta$ given that $k \geq k_0$.

In the third part of the second statement, we consider the case where $i \in \hat{S}_{4,\varepsilon}(\mathbf{w}^{(k)})$. For this case to happen, we must first have $\psi_i(w_i^{(k)}; \varepsilon) \leq 0$. Let $c_- < c_+$ be such that $\psi_i(c_-; \varepsilon) = \psi_i(c_+; \varepsilon) = 0$. By Lemma 1, for all k large enough, there are three possible cases: (a) $w_i^{(k)}$ stays at the left of $[c_-, c_+]$ or $w_i^{(k)}$ stays at the right of $[c_-, c_+]$, or (b) $w_i^{(k)}$ visits $[c_-, c_+]$ infinitely often.

If there is no k_1 such that $w_i^{(k)} \in (c_-, c_+)$ for all $k \geq k_1$, then there exists a subsequence $\{w_i^{(j)}\}_{j \geq 1}$ such that $w_i^{(j)} \to c_-$ or $w_i^{(j)} \to c_+$, since $\limsup_k w_i^{(k)} \leq c_+$, $\liminf_k w_i^{(k)} \geq c_-$. As a result, either $w_i^{\infty} \to c_-$ or $w_i^{\infty} \to c_+$, as $\mathrm{d}(\mathbf{w}^{(k)}, C_\varepsilon) \to 0$ (if w_i ever converges). Otherwise, $\hat{S}_{4,\varepsilon}(\mathbf{w}^k) = \emptyset$ for every $k \geq k_1$.

We verify the stationarity condition for w_i^{∞}. It suffices to consider the case where $w_i^{\infty} = c_-$ without loss of generality. Violation of the stationary condition only happens when $-[\nabla\ell(\mathbf{w}^{\infty})]_i \cdot \psi_i'(c_-;\varepsilon) > 0$. However, the algorithm will take an update with $-[\nabla\ell(\mathbf{w}^{\infty})]_i \neq 0$, which violates $w_i^{(k)}$'s convergence to c_-.

Denote $\bar{\bar{\mathbf{w}}}$ as a result of projecting $\mathbf{w}$ to C_ε for coordinates $i \in S_{4,\varepsilon}(\mathbf{w})$. It is clear that the smoothness condition gives $\|\nabla\ell(\mathbf{w}) - \nabla\ell(\bar{\bar{\mathbf{w}}})\| \le L_\ell\|\bar{\bar{\mathbf{w}}} - \mathbf{w}\| \le L_\ell\delta$. This implies $\|\nabla\ell(\mathbf{w})\| \le \|\nabla\ell(\bar{\bar{\mathbf{w}}})\| + L_\ell\delta$ and hence $\|\nabla\ell(\mathbf{w})\|^2 \le \|\nabla\ell(\bar{\bar{\mathbf{w}}})\|^2 + 2L_\ell\delta\|\nabla\ell(\bar{\bar{\mathbf{w}}})\| + L_\ell^2\delta^2$.

$$
\begin{aligned}
\mathbb{E}\left[\|\nabla\ell(\mathbf{w}^{(\tau_K)})]\|^2\right] &= \mathbb{E}\left[\sum_{i\in[n]}[\nabla\ell(\mathbf{w}^{(\tau_K)})]_i^2\right] \\
&= \mathbb{E}\left[\sum_{i\in\hat{S}_{1,\varepsilon}(\mathbf{w}^{(\tau_K)})\cup\hat{S}_{2,\varepsilon}(\mathbf{w}^{(\tau_K)})}[\nabla\ell(\mathbf{w}^{(\tau_K)})]_i^2\right] \\
&\quad + \mathbb{E}\left[\sum_{i\in\hat{S}_{3,\varepsilon}(\mathbf{w}^{(\tau_K)})}[\nabla\ell(\mathbf{w}^{(\tau_K)})]_i^2\right] + \mathbb{E}\left[\sum_{i\in\hat{S}_{4,\varepsilon}(\mathbf{w}^{(\tau_K)})}[\nabla\ell(\mathbf{w}^{(\tau_K)})]_i^2\right] \\
&\le 2\mathrm{a}\delta + (2\alpha\delta)^2 + L_\ell^2\delta^2 + 2L_\ell\delta\mathbb{E}\left[\sum_{i\in\hat{S}_{4,\varepsilon}(\mathbf{w}^{(\tau_K)})}\left|\left[\nabla\ell(\bar{\bar{\mathbf{w}}}^{(\tau_K)})\right]_i\right|\right] \\
&\quad + \mathbb{E}\left[\sum_{i\in\hat{S}_{4,\varepsilon}(\mathbf{w}^{(\tau_K)})}\left[\nabla\ell(\bar{\bar{\mathbf{w}}}^{(\tau_K)})\right]_i^2\right] \quad \text{(as K approaches infinity)} \\
&\xrightarrow{K\to\infty\Rightarrow\delta\to0} \mathbb{E}\left[\sum_{i\in\hat{S}_{4,\varepsilon}(\mathbf{w}^{(\tau_K)})}\left[\nabla\ell(\bar{\bar{\mathbf{w}}}^{(\tau_K)})\right]_i^2\right].
\end{aligned}
$$

From the above we know that we can be arbitrarily close to a stationary point in $\mathcal{Z}_\varepsilon$. $\square$

5.2.2 Asymptotic Convergence with Lipschitz Smoothness and Robbins-Monro Stepsizes

Theorem 6 (**Asymptotic Convergence with Lipschitz Smoothness and Robbins-Monro Stepsizes**)**.** *Assuming that A1, A2, A5, A6, A7,* $0 < \varepsilon \le \inf_{1\le i\le d}\inf_{1\le j<K_i}|q_i^{(j)} - q_i^{(j+1)}|^2/4$ *holds, where* $\{q_i^{(j)}\}$ *are the quantization levels, then* $\lim_{k\to\infty} d(w^{(k)}, \mathcal{Z}_\varepsilon) \to 0$ *almost surely.*

Significance. This has largely improved the results of ASkewSGD's convergence with incorporation of the mainstream stochastic analysis method. We can even generalize the constraint function ψ. This has left us with a class of candidate functions for selection (given that ψ is Lipschitz-smooth).

Proof. To: I will read this. The ASkewSGD update is

$$\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} + \gamma_k\hat{\mathbf{v}}^{(k)},$$

where $\hat{\mathbf{v}}^{(k)} = \mathbf{s}_{\varepsilon,\alpha}(\widehat{\nabla\ell}(\mathbf{w}^{(k)}), \mathbf{w}^{(k)})$, with the mean direction is $\mathbf{h}(\mathbf{w}) = \mathbb{E}[\mathbf{s}_{\varepsilon,\alpha}(\widehat{\nabla\ell}(\mathbf{w}), \mathbf{w})|\mathbf{w}]$.

Consider the Stochastic Approximation (SA) Scheme, where we express the update as:

$$\mathbf{w}^{(k+1)} \leftarrow \mathbf{w}^{(k)} + \gamma_k(\mathbf{h}(\mathbf{w}^{(k)}) + \boldsymbol{\xi}^{(k)}),$$

where $\boldsymbol{\xi}^{(k)} = \hat{\mathbf{v}}^{(k)} - \mathbf{h}(\mathbf{w}^{(k)})$ is a martingale difference noise.

Define the Lyapunov function

$$L(\mathbf{w}) = \ell(\mathbf{w}) + \frac{\alpha}{2}\sum_{i=1}^{d}[\min(0, \psi_i(\mathbf{w}))]^2.$$

Since $\ell \in C^1$, and the penalty term $[\min(0, \psi_i(w_i; \varepsilon))]^2$ is C^1 due to the piecewise C^1 structure of ψ_i and the derivative continuity at boundaries, we know that L is C^1. Note that ℓ is also radially bounded (if ℓ is also coercive), with the gradient

$$\nabla L(\mathbf{w}) = \nabla \ell(\mathbf{w}) + \mathbf{r}(\mathbf{w}), \text{where } r_i(\mathbf{w}) = \begin{cases} 0, & \psi_i(w_i; \varepsilon) > 0, \\ \alpha\psi_i(w_i; \varepsilon)\psi_i'(w_i; \varepsilon), & \psi_i(w_i; \varepsilon) \leq 0. \end{cases}$$

Moreover, $\mathbf{h}(\mathbf{w}) = \mathbf{0}$ if and only if $\mathbf{w} \in \mathcal{Z}_\varepsilon$.

We first prove the "if" direction. When $\mathbf{w} \in \mathcal{Z}_\varepsilon$ (which is a subset of C_ε), $\psi_i(w_i; \varepsilon) \geq 0$. If $\psi_i(w_i; \varepsilon) > 0$, then $[\nabla \ell(\mathbf{w})]_i = 0$. Otherwise, $\alpha\psi_i(w_i; \varepsilon)/\psi_i'(w_i; \varepsilon) = 0$ and we just need to consider the case $-[\nabla \ell(\mathbf{w})]_i \cdot \psi_i'(w_i; \varepsilon) \geq 0$, this is impossible as $\operatorname{sign}([\nabla \ell(\mathbf{w})]_i) = \operatorname{sign}(\psi_i'(w_i; \varepsilon))$.

We then investigate the "only-if" direction. If $\psi_i(w_i; \varepsilon) > 0$, we can make use of the fact that $[\nabla \ell(\mathbf{w})]_i = [h(\mathbf{w})]_i = 0$. If $\psi_i(w_i; \varepsilon) < 0$ and $-[\nabla \ell(\mathbf{w})]_i \cdot \psi_i'(w_i; \varepsilon) \leq \alpha\psi_i(w_i; \varepsilon)$, then $h(\mathbf{w}) \neq 0$, which is a contradiction. This implies that $0 = -[\nabla \ell(\mathbf{w})]_i \cdot \psi_i'(w_i; \varepsilon) \geq \alpha\psi_i(w_i; \varepsilon) > 0$, which is also impossible. If $\psi_i(w_i; \varepsilon) = 0$ and $-[\nabla \ell(\mathbf{w})]_i \cdot \psi_i'(w_i; \varepsilon) \leq \alpha\psi_i(w_i; \varepsilon) = 0$, then indeed we have $\operatorname{sign}([\nabla \ell(\mathbf{w})]_i) = \operatorname{sign}(\psi_i'(w_i; \varepsilon))$. Now, we are left with the conditions that takes $[h(\mathbf{w})]_i = -[\nabla \ell(\mathbf{w})]_i = 0$ and $\psi_i(w_i; \varepsilon) \geq 0$, which is always satisfactory.

We need to prove the following claim.

Claim. Almost surely, $\langle \nabla L(\mathbf{w}), \mathbf{h}(\mathbf{w}) \rangle \leq -\mathtt{b}\|\mathbf{h}(\mathbf{w})\|^2$ for some $\mathtt{b} > 0$ when $\mathbf{h}(\mathbf{w}) \neq 0$.

Case 1. If w_i is strictly feasible, i.e. $\psi_i(\mathbf{w}) > 0$, then $[\nabla L(\mathbf{w})]_i = [\nabla \ell(\mathbf{w})]_i$ and $[h(\mathbf{w})]_i = -[\nabla \ell(\mathbf{w})]_i$. Hence,

$$[\nabla L(\mathbf{w})]_i[\mathbf{h}(\mathbf{w})]_i = -[\nabla \ell(\mathbf{w})]_i^2 = -[\mathbf{h}(\mathbf{w})]_i^2.$$

Case 2. If w_i is not strictly feasible, i.e. $\psi_i(\mathbf{w}) \leq 0$, then $[\nabla L(\mathbf{w})]_i = [\nabla \ell(\mathbf{w})]_i + \alpha\psi_i(w_i; \varepsilon)\psi_i'(w_i; \varepsilon)$.

(a) The gradient condition holds, i.e. $-[\nabla \ell(\mathbf{w})]_i \cdot \psi_i'(w_i; \varepsilon) \geq -\alpha\psi_i(w_i; \varepsilon) \geq 0$. Then, $[\mathbf{h}(\mathbf{w})]_i = -[\nabla \ell(\mathbf{w})]_i$,

$$\begin{aligned}[\nabla L(\mathbf{w})]_i \cdot [\mathbf{h}(\mathbf{w})]_i &= -[\nabla \ell(\mathbf{w})]_i^2 - \alpha\psi_i(w_i; \varepsilon)\psi_i'(w_i; \varepsilon) \cdot [\nabla \ell(\mathbf{w})]_i \\ &\leq -[\nabla \ell(\mathbf{w})]_i^2 - \alpha^2\psi_i^2(w_i; \varepsilon) \leq -[\mathbf{h}(\mathbf{w})]_i^2,\end{aligned}$$

where

$$-\alpha\psi_i(w_i; \varepsilon)\psi_i'(w_i; \varepsilon) \cdot [\nabla \ell(\mathbf{w})]_i \leq -\alpha\psi_i(w_i; \varepsilon) \cdot (\alpha\psi_i(w_i; \varepsilon)) = -\alpha^2\psi_i(w_i; \varepsilon)^2.$$

(b) The gradient condition fails, i.e. $-[\nabla \ell(\mathbf{w})]_i \cdot \psi_i'(w_i;\varepsilon) < -\alpha\psi_i(w_i;\varepsilon)$. Then, $[\mathbf{h}(\mathbf{w})]_i = \mathrm{clip}(-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon), M_c)$.
If $|[\mathbf{h}(\mathbf{w})]_i| \le M_c$, then

$$\begin{aligned}
[\nabla L(\mathbf{w})]_i \cdot [\mathbf{h}(\mathbf{w})]_i &= ([\nabla \ell(\mathbf{w})]_i + \alpha\psi_i(w_i;\varepsilon)\psi_i'(w_i;\varepsilon)) \cdot (-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon)) \\
&= -\alpha\frac{\psi_i(w_i;\varepsilon)\cdot[\nabla \ell(\mathbf{w})]_i}{\psi_i'(w_i;\varepsilon)} - \alpha^2\psi_i(w_i;\varepsilon) \\
&< -\alpha^2\psi_i^2(w_i;\varepsilon)\left(\frac{1}{(\psi_i'(w_i;\varepsilon))^2}+1\right) \\
&\le -\left(\frac{\alpha\psi_i(w_i;\varepsilon)}{\psi_i'(w_i;\varepsilon)}\right)^2 = -[\mathbf{h}(\mathbf{w})]_i^2.
\end{aligned}$$

Otherwise, assume that clipping is activated and $[\mathbf{h}(\mathbf{w})]_i = \varsigma M_c$, where $\varsigma = \mathrm{sign}(-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon))$. Since the case where $\psi_i'(w_i;\varepsilon) = 0$ is measure-zero, then almost surely,

$$\begin{aligned}
[\nabla L(\mathbf{w})]_i \cdot [\mathbf{h}(\mathbf{w})]_i &= [\nabla L(\mathbf{w})]_i \cdot (-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon)) \cdot \frac{\varsigma M_c}{-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon)} \\
&\le -(\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon))^2 \cdot \frac{\varsigma M_c}{-\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon)} \\
&= (\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon)) \cdot \varsigma M_c \\
&\le -\varsigma^2 M_c^2 \quad (\text{since } |\alpha\psi_i(w_i;\varepsilon)/\psi_i'(w_i;\varepsilon)| \ge M_c) \\
&= -M_c^2.
\end{aligned}$$

Thus, $L(\mathbf{w}(t))$ is almost surely non-increasing along trajectories. Moreover, L is radially unbounded, then $L(\mathbf{w}(t))$ converges to some $L^* \ge \inf L$ as $t \to \infty$.

Define the equilibrium set:

$$E = \left\{\mathbf{w} : \frac{\mathrm{d}}{\mathrm{d}t}L(\mathbf{w}) = 0\right\} \overset{\text{a.s.}}{=} \{\mathbf{w} : h(\mathbf{w}) = 0\} = \mathcal{Z}_\varepsilon$$

By LaSalle's Invariance Principle, the largest invariant set in E is $\mathcal{Z}_\varepsilon$ (since $h(\mathbf{w}) = 0$ implies $\mathbf{w}(t)$ is constant). Hence,

$$\lim_{t\to\infty} d(\mathbf{w}(t), Z_\varepsilon) = 0 \quad \text{a.s.}.$$

For any $\epsilon > 0$, choose $\xi > 0$ such that

$$\mathbf{w}(0) \in B_\xi(\mathcal{Z}_\varepsilon) \implies L(\mathbf{w}(0)) < \min_{\mathbf{w}\in\partial B_\epsilon(\mathcal{Z}_\varepsilon)} L(\mathbf{w}).$$

Since L decreases along trajectories, $\mathbf{w}(t) \in B_\epsilon(\mathcal{Z}_\varepsilon)$ for all $t \ge 0$, and the set Z_ε is almost surely globally asymptotically stable.

Via Kushner-Clark theorem, we verify that $\{\mathbf{w}^{(k)}\}$ bounded a.s., $\mathbb{E}[\boldsymbol{\xi}^{(k)} \mid \mathcal{F}_k] = 0$, $\mathbb{E}[\|\boldsymbol{\xi}^{(k)}\|^2 \mid \mathcal{F}_k] \le \sigma^2$, $\sum \gamma_k = \infty$, $\sum \gamma_k^2 < \infty$, and Z_ε is almost surely globally asymptotically stable for ODE, and conclude that

$$\lim_{k \to \infty} d(\mathbf{w}^{(k)}, \mathcal{Z}_\varepsilon) = 0 \quad \text{a.s..}$$

□

6 Experimental Results

We compare `ASkewSGD` [1] with 3 other approaches: a full precision NN, `BinaryConnect` [3], `ProxQuant` [5]. We denote [Wx/Ay] as a neural architecture with x-bit precision weights and y-bit activations. For x-bit quantization, we refer to the discrete integer values from -2^{x-1} to $2^{x-1}-1$.

Two Moons Classification. We train a five-layer MLP (shown as below) with a 2-dimensional input, and ReLU as activation function on the non-convex two moon classification task for $T = 60$ epochs. Our dataset consists of $n = 2000$ training samples (in batches of 50 per iteration, points are colored in blue and red) and $m = 500$ test samples (colored in black and white) in 2D, which is generated by `scikit-learn` library's `make_moons` data generator. We have also added a random Gaussian noise with the variance of $\xi = 0.1$ to offset the point from its original position. We apply the logistic loss function in this experiment. The learning rate for the full precision SGD, `BinaryConnect`, `ProxQuant`, `ASkewSGD` is set to 0.5, 0.05, 0.05 and 0.166, respectively. Then, we multiply the learning rate by 0.5 at epochs $15, 25$. For `ProxQuant`, we start the regularization at epoch $T_{PQ} = 11$ and set $\lambda_{it} = 0.00125 \cdot (it - 400)$, where it is the iteration number. For `ASkewSGD`, we start the annealing process at epoch $T_{AS} = 30$ and set $\alpha = 0.2$, $\varepsilon_t = 0.88^{t-T_{AS}}$.

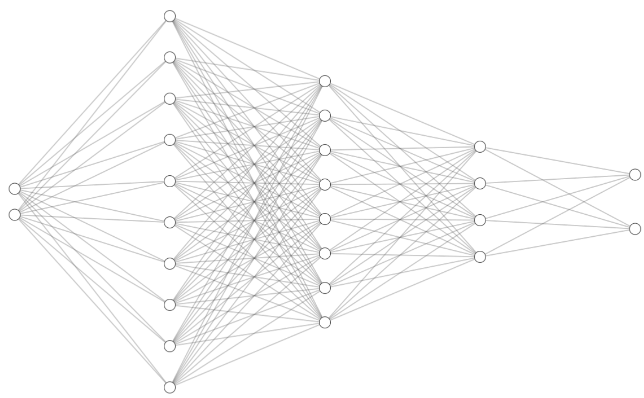

Figure 3: NN architecture for the Two Moons Classification Experiment.

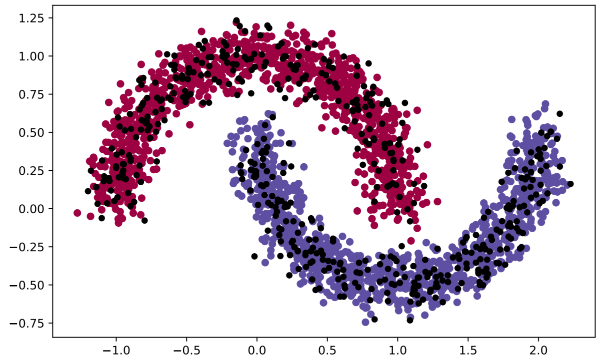

Figure 4: The Two Moons Dataset.

We obtain the following contour plot for the two moons classification problem. The first row shows the performance of different optimizers on the original net before quantization, and the second row depicts the quantized model with the quantization set $\mathcal{Q} = \{-1, +1\}$.

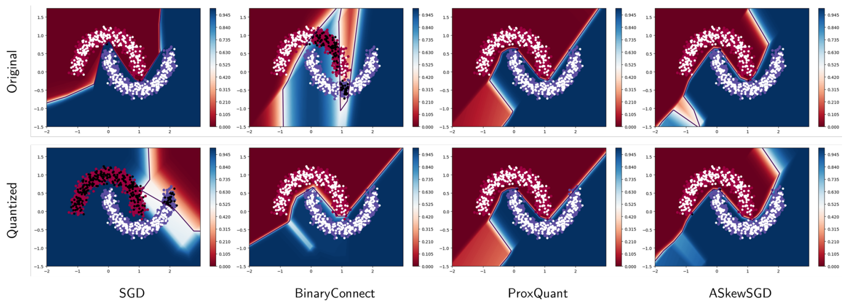

Figure 5: Contour plot with the Two Moons Dataset.

The original net has its contour completely perturbed by the weight quantization process. `BinaryConnect` is aware of the quantization scheme, thus generalizing well on the test set, but the original model is completely off from the groundtruth. `ASkewSGD` and `ProxQuant`, on the other hand, have only a slight skew away from the sensible region, and the quantized model is still able to generalize well on the test set, which demonstrates the strength of the two optimization methods. However, we observe from Table 1 that `ProxQuant` can still perform inconsistently even on this simple task.

Table 1: Performance after 60 epochs (average and standard deviation over 15 random experiments).

Method	Loss	Quantized Loss	Quantized Accuracy
Full Precision [W32/A32]	0.08 ± 0.0913	5.44 ± 2.6426	49.85 ± 5.162
Deterministic `BinaryConnect` [W1/A32]	0.80 ± 0.4973	$\mathbf{0.03 \pm 0.0731}$	98.77 ± 4.0333
`ProxQuant` [W1/A32]	$\mathbf{0.01 \pm 0.0058}$	0.28 ± 0.5932	96.03 ± 6.8739
`ASkewSGD` [W1/A32]	0.01 ± 0.0111	0.05 ± 0.0719	$\mathbf{98.80 \pm 2.3833}$

MNIST. We train a 3-layer MLP model without bias on MNIST for $T = 50$ epochs. Our MLP uses an input layer of dimension $d = 784$, two hidden layers (512 and 512 neurons), and an output layer (10 neurons). The dataset consists of $n = 50000$ training samples (in batches of 50) and $m = 10000$ test samples. We apply the multiclass cross entropy loss function in this experiment.

We first describe the hyperparameters for the 1-bit setting. The learning rate for the full precision SGD, `BinaryConnect`, `ProxQuant`, `ASkewSGD` is set to 0.06, 0.0002, 0.06 and 0.06, respectively. Then, we multiply the learning rate by 0.5 at epochs $7, 15, 30$. For `ProxQuant`, we start the regularization at epoch $T_{PQ} = 21$ and set $\lambda_{it} = 10^{-7} \cdot (it - 20000)$, where it is the iteration number. For `ASkewSGD`, we start the annealing process at epoch $T_{AS} = 21$, set $\alpha = 0.001$, and in each epoch t we have $\varepsilon_t = 0.88^{t-T_{AS}}$. For the 2-bit setting, the learning rate for the full precision SGD, `BinaryConnect`, `ProxQuant`, `ASkewSGD` is set to 0.006, 0.0075, 0.06 and 0.0002, respectively. Then, we multiply the learning rate by 0.5 at epochs $7, 15, 30$. For `ProxQuant`, we start the regularization at epoch $T_{PQ} = 1$ and set $\lambda_{it} = 10^{-7} \cdot it$, where it is the iteration number. For `ASkewSGD`, we start the annealing process at epoch $T_{AS} = 21$, set $\alpha = 0.02$, and in each epoch t we have $\varepsilon_t = 0.88^{t-T_{AS}}$.

Table 2: Performance on MNIST after 50 epochs over 5 random experiments.

Method	Test Loss	Test Accuracy	Quantized Test Accuracy	Top-1 (Quantized)
Full Precision [W32/A32]	0.089 ± 0.001	$97.27\% \pm 0.0.03$	$93.54\% \pm 1.270$ ([W1/A32])	94.94% ([W1/A32])
(Best: **98.26**%)	0.167 ± 0.004	$95.12\% \pm 0.157$	89.19 ± 2.300 ([W2/A32])	91.71% ([W2/A32])
`BinaryConnect` [W1/A32]	2.289 ± 0.001	$58.70\% \pm 4.058$	$93.26\% \pm 0.455$	93.66%
`ProxQuant` [W1/A32]	0.089 ± 0.001	$97.25\% \pm 0.034$	$93.75\% \pm 1.100$	94.94%
`ASkewSGD` [W1/A32]	143.501 ± 6.343	$95.78\% \pm 0.298$	**$94.47\% \pm 0.508$**	**95.14**%
`BinaryConnect` [W2/A32]	2.288 ± 0.002	$86.67\% \pm 0.855$	$94.25\% \pm 0.485$	94.80%
`ProxQuant` [W2/A32]	0.144 ± 0.003	$95.74\% \pm 0.068$	$87.48\% \pm 6.237$	93.33%
`ASkewSGD` [W2/A32]	0.068 ± 0.001	$97.94\% \pm 0.075$	**97.26 ± 0.252**	**97.55**%

We had to lower the learning rate for the full-precision setting as it acts as a trade-off between quantized test accuracy and final test accuracy, while we have reported the best performance using a larger learning rate of $lr = 0.09$ that has helped to generalize better. Even without quantization-aware training, the full-precision MLP can still provide an accuracy of 94.94% under 1-bit quantization, which is surprising. In this task, `ASkewSGD` demonstrates its strength for both the 1-bit and 2-bit settings (with an accuracy close to the full precision in the 2-bit setting), while `ProxQuant` continues to lag behind due to its inherent sensitivity to the regularizing parameter λ.

Computer Vision Task. We train `ResNet-18` without bias and tunable parameters on the Batch Normalization Layer on the batch normalization layer on the CIFAR-10 dataset for $T = 150$ epochs. The dataset consists of $n = 50000$ training samples and $m = 10000$ test samples. We apply the cross-entropy loss function in this experiment. We set the learning rate lr and the hyperparameter for ASkewSGD α to 0.06 and 0.2, respectively. The quantization set $\mathcal{Q}$ is set to $\{-1, +1\}$.

Table 3: Performance of ResNet-18 on CIFAR10 after 150 epochs.

Method	Best Test Accuracy	Test Loss
Full Precision [W32/A32]	89.25%	0.580
`BinaryConnect` [W1/A32]	86.24%	0506
`ProxQuant` [W1/A32]	.% (quantized)	0.3940 (quantized)
`ASkewSGD` [W1/A32]	86.91% (quantized)	0.2586 (quantized)

It is worthwhile to note that the accuracy of the quantized model skyrocketed to 50% only after running for 40 epochs, meaning that quantized models are sensitive to the accumulation of positional errors and the uncertain landscape of the loss function.

LSTM Language Model.

7 Conclusion

Previous optimization algorithms show empirical success for quantization-aware training of deep neural networks. While there are advancements in different quantization schemes, we should keep on investigating the theoretical guarantees provided by these rules in order to keep the error at a justifiable level. As a clear goal, we have managed to show fruitful convergence guarantees of the modified variant of `ASkewSGD` under weaker assumptions that is dedicated to covering a wider family of neural network architectures.

Acknowledgements

I am truly thankful for Professor Wai's insightful advice and suggestions during the research period. His expertise in optimization theory and signal processing has always pinpointed the essence of the great ideas in these fields. I would also like to thank Mr. Yau Chung Yiu for his patience and knowledge in quantization when experimental results did not go as expected. Without his help, I might not be able to finish the implementations of all the algorithms presented.

References

[1] L. Leconte, S. Schechtman and E. Moulines. (2023). ASkewSGD: An Annealed Interval-Constrained Optimisation Method to Train Quantized Neural Networks. In *Artificial Intelligence and Statistics 2023*, **206**:3644-3663.

[2] T. Dockhorn, Y. Yu, E. Sari, M. Zolnouri, V. P. Nia. (2021). Demystifying and Generalizing BinaryConnect. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021).* `arxiv.org/pdf/2110.13220`.

[3] M. Courbariaux, Y. Bengio, J.-P. David. (2015). BinaryConnect: Training Deep Neural Networks with Binary Weights During Propagations. In *Advances in Neural Information Processing Systems (NeurIPS 2015).* `arxiv.org/pdf/1511.00363`.

[4] S. Zhou, Y. Wu, Z. Ni (2016). DoReFa-net: Training Low Bitwidth Convolutional Neural Networks with Low Bitwidth Gradients. `arxiv.org/pdf/1606.06160`.

[5] Y. Bai, Y.-X. Wang, E. Liberty. (2019). Proxquant: Quantized Neural Networks via Proximal Operators. In *The 7th International Conference on Learning Representations 2019.* `arxiv.org/pdf/1810.00861`.

[6] M. Muehlebach, M. I. Jordan. (2022). On Constraints in First-Order Optimization: A View from Non-Smooth Dynamical Systems. In *Journal of Machine Learning Research*, **23**(256):1-47. `https://jmlr.org/papers/v23/21-0798.html`.

[7] S. Ghadimi, G. Lan. (2013). Stochastic First- and Zeroth-Order Methods for Nonconvex Stochastic Programming. In *SIAM Journal on Optimization*, **23**(4):2341-2368.