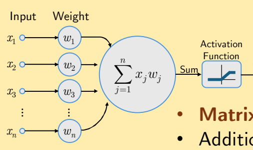
On the Robustness of Quantization Algorithms during the Training Phase of Deep Neural Networks

Researcher: Hok Fong WONG (`hfwong2@cse.cuhk.edu.hk`, Department of Computer Science and Engineering)
Supervisor: Prof. Hoi-To WAI (`htwai@se.cuhk.edu.hk`, Department of Systems Engineering and Engineering Management)

1. Introduction to Quantization in DNNs

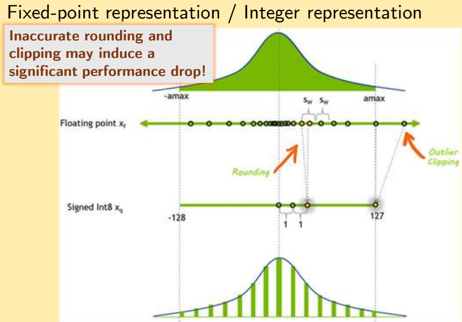
Goal:
- Model deployment on **low-memory** devices
- Lowering the **inference** time

Core Arithmetic Operations in DNNs:



- **Matrix Multiplication**
- Addition Operations

Motivation:
Manipulating number representation –
Fixed-point representation / Integer representation



2. Previous Works

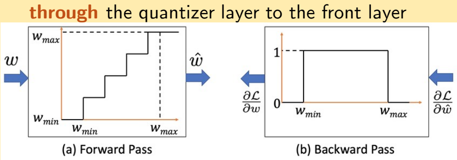
I. BinaryConnect (BC) (Courbariaux et al., 2015)
- Stochastically binarized weights:

$$w_b = \begin{cases} +1, & \text{with probability } p = \sigma(w), \\ -1, & \text{with probability } 1-p. \end{cases} \quad \text{(Simple addition)}$$

$$\sigma(x) = \text{clip}\left(\frac{x+1}{2}, 0, 1\right) = \max\left(0, \min\left(1, \frac{x+1}{2}\right)\right)$$

  - Only binarized on the forward and backward path
  - Full precision for parameter update
- Empirically as a **regularizer** (noisy weights unbiased in expectation)
- **Save 2/3 of multiplications** with specialized hardware design

II. Straight-Through Estimators (STEs) (Bengio et al., 2013)
- Forward propagation: weights are quantized
- Backward propagation: gradients **directly pass through** the quantizer layer to the front layer



(a) Forward Pass (b) Backward Pass

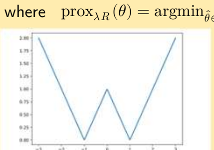
- **Limited understanding** despite the empirical success
- Oscillation of the generated gradient from "quantized" parameters
- Coarse gradient must be chosen with proper STEs, required to correlate positively with the population gradient, e.g. **clipped ReLU** (Yin et al., 2019)

III. ProxQuant (Bai et al., 2019)
- **Proximal Operator**

$$R(\theta) = \sum_{i=1}^{n} \min\{|\theta_i - 1|, |\theta_i + 1|\} \quad \text{(Regularization Function)}$$

$$w_{t+1} = \text{prox}_{\gamma_t \lambda_t R}\left(\theta_t - \gamma_t \widehat{\nabla} \ell(\theta_t)\right) \quad \text{(Soft Projection Function)}$$

where $\text{prox}_{\lambda R}(\theta) = \text{argmin}_{\hat{\theta} \in \mathbb{R}^d} \left\{\frac{1}{2}\left|\hat{\theta} - \theta\right|_2^2 + \lambda R(\hat{\theta})\right\}$



$R(\theta) = 0$ when $\theta \in Q$ and $R(\theta) > 0$ when $\theta \notin Q$
- Best iterate is guaranteed to converge, with smoothed regularizers and loss function, step size is constant $(1/\beta)$

3. Optimization Formulation and Notions

Minimization of Training Loss with Quantization Constraints on the Weights

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

Inherent difficulty
- Multi-layer DNNs - can be **non-convex, non-differentiable**
- **Combinatorial**: discrete quantization levels
- **NP-Hard** in general for smooth functions
- **MINLP** could fail due to the **scale of the number of parameters**

Smoothed Interval Constraint Relaxation

$$\min_{w \in C} \ell(w), C = \{w \in \mathbb{R}^n : g(w) \geq 0\}$$

$$\psi_\epsilon^i(w^i) := \begin{cases} \epsilon - (q_1^i - w^i)^2, & w^i < q_1^i, \\ \epsilon - (w^i - q_{j-1}^i)^2 (w^i - q_j^i)^2, & q_{j-1}^i \leq w^i < q_j^i, j = 2, \dots, K, \\ \epsilon - (w^i - q_K^i)^2, & w^i \geq q_{K^i}^i. \end{cases}$$

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

Tangent Cone and Normal Cone Induced by MFCQ

$T_C(w) = \{v \mid \nabla g_i(x)^\top v \geq 0, \forall i \in I(w)\}$, (Directions to mend **all** violated constraints)

$N_C(w) = \left\{-\sum_{i \in I(x)} \lambda_i \nabla g_i(w) \middle| \lambda \in \mathbb{R}_+^d\right\}$ (Descent directions of violated constraints)

Strong Duality and Optimality Conditions

$Z = \{w \in C : 0 \in -\nabla \ell(w) - N_C(w)\}$ (Stationary points)

4. Muehlebach-Jordan's Algorithm (2022)

Assumption 1. ℓ, g are continuously differentiable and have a Lipschitz continuous gradient. ℓ is lower-bounded and C is non-empty and bounded.
Assumption 2. MFCQ is satisfied for all x.
Assumption 3. C is convex and ℓ is strongly convex.
Update rule:

$$\begin{cases} w_{k+1} = w_k + \gamma_k v_k \\ v_k = \text{argmin}_{v \in V_\alpha(w_k)} (1/2)|v + \nabla \ell(w_k)|^2 \end{cases}$$

Theorem 1. The iterates are guaranteed to converge to the minimizer of ℓ at nearly a linear rate, under Assumptions 1-3.

5. Extension: ASkewSGD Algorithm

Techniques by Leconte et al. (2023): Construction of a regularization function with MFCQ + Simulated annealing for discovery + Gradient flow characterization
- **No projection; Stochastic gradients for large-scale ML**
- **"Simple is the best" dictum**

Assumption 4. The step sizes γ_k are non-increasing, non-summable, and square-summable.
Assumption 5. $\ell(\cdot; \xi_i)$ is **d-times continuously differentiable** and has M_{ℓ_i} Lipschitz continuous gradients.
Explicit solution for v_k:

$$[s_{\epsilon,\alpha}(\widehat{\nabla}\ell(w_k), w_k)]^i = \begin{cases} -\widehat{\nabla}\ell(w_k^i), & \text{if } \psi_\epsilon(w^i) \geq 0 \text{ or } -\psi'_\epsilon(w^i)\widehat{\nabla}\ell(w_k^i) \geq -\alpha\psi_\epsilon(w^i) > 0, \\ \text{clip}(-\alpha\psi_\epsilon(w^i)/\psi'_\epsilon(w^i), M_\epsilon), & \text{otherwise.} \end{cases}$$

Theorem 2. Under Assumption 1, 4, 5, and $0 < \epsilon \leq \inf_{1 \leq i \leq d} \inf_{1 \leq j \leq K^i} |c_j^i - c_{j+1}^i|^4/16$, where $\{c_j^i\}$ are the quantization levels. Then, $\ell(w_k)$ converges and $\lim_{k \to \infty} \mathrm{d}(w_k, Z_\epsilon) = 0$ almost surely.

Our work: Eliminating the need to introduce the highly-differentiable loss function
Observation: Three cases for an iterate:
(a)Taking the descent direction
(b)Descent direction matches with the pushing force
(c)Gradient mismatches with the pushing force
In classical stochastic smooth analysis, we mainly rely on the bounded gradient for local minimization guarantees.
Difficulty: Quantifying the motions of iterates depends on the loss function! Without clipping, v_k can be very large which leads the iterate to infinity!

$$\mathbb{E}[\ell(w_{k+1})|w_k] \leq \ell(w_k) + \gamma_k \mathbb{E}\left[\boxed{v_k^\top \nabla \ell(w_k)}|w_k\right] + \frac{\gamma_k^2 L}{2} \mathbb{E}\left[\|v_k\|^2 | w_k\right]$$

Idea (Coordinate-wise): Given **sufficient** time, iterates stay within a distance ϵ from feasible set. (a) Small gradients on the edge of feasible set: ignorable; (b) Large gradients on the edge of feasible set: iterate converges as a KKT point / takes a small pushing force O(ϵ) / by smoothness gradually leaving the boundary stripe in **finite time**.

6. Stochastic Gradient Descent Ascent

Lagrangian-Primal Problem Formulation

$$\min_{w \in \mathbb{R}^n} \max_{\lambda \geq 0} \mathcal{L}(w, \lambda) \qquad \mathcal{L}(w, \lambda) := \ell(w) - \lambda^\top g(w)$$

- **Stochastic Gradient Descent Ascent (SGDA) for Nonconvex-Concave Minimax Problem** (Lin et al., 2024)
- **Very small step sizes** for w, λ, and **smoothness** of $\mathcal{L}$ is enough to guarantee convergence (ϵ-stationary point)

$$w_k \leftarrow w_{k-1} - \eta_x \widehat{\nabla}_{w_{k-1}} \mathcal{L}(w_{k-1}, \lambda_{k-1})$$

$$\lambda_k \leftarrow \mathcal{P}\left(w_{k-1} + \eta_\lambda \widehat{\nabla}_{\lambda_{k-1}} \mathcal{L}(w_{k-1}, \lambda_{k-1})\right)$$

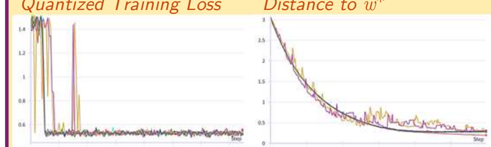
7. Experiment Setup and Results

We used full-precision *SGD* for comparison, and tested *BinaryConnect, Straight-Through Estimator, ASkewSGD, modified ASkewSGD and SGDA*. BinaryConnect and STE both suffered from strong oscillations and exhibited a larger loss. ASkewSGD and SGDA are close to the full precision method for task I and II.

I. Convex Logistic Regression (Single Layer)

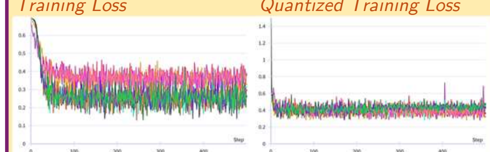
Training Loss *Batch Gradient*

Quantized Training Loss *Distance to w^**

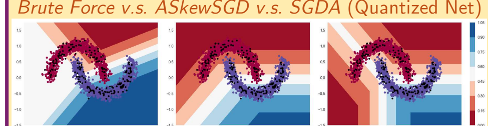


II. Two Moons Classification (Shallow NN)

Training Loss *Quantized Training Loss*



Brute Force v.s. ASkewSGD v.s. SGDA (Quantized Net)



III. Computer Vision Task (ResNet-18 on CIFAR-10)

Method	[W1/A32]	[W2/A4]
BinaryConnect		
Straight-Through Estimator		
ASkewSGD		
SGDA		
Full-precision [W32/A32]	**88.30** (20 epochs)	

8. Future Works

- Does ASkewSGD escape from saddle points?
- Step sizes for Lagrangian-type minimax problems
- Distributed optimization for block-structured constraint formulations
- Possibility of solving combinatorial optimization tasks

9. Major Text

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.