

# Supplementary Material for “WaveQ: Gradient-Based Deep Quantization of Neural Networks through Sinusoidal Regularization”

## A DETAILED THEORETICAL ANALYSIS

### A.1 MOTIVATION

The results of this section are motivated by the following question.

**Question A.1.** *Suppose that a function  $F : \mathbb{R}^n \rightarrow [0, \infty)$  has many global minima and that  $Q \subset \mathbb{R}^n$  is closed. How do we isolate the global minima of  $F$  that are closest to  $Q$  without actually computing the full set of global minima of  $F$ ?*

Intuitively, we would like to show that if  $\epsilon > 0$  is very small, then the global minima of the function

$$F(x) + \epsilon d(x, Q)$$

are very close to the global minima of  $F$  closest to  $Q$ . To achieve this we will have to introduce first the concept of convergence of sets and then we will show that our intuition is correct by proving that the set of global minima to the above relaxed function converges to a subset of global minima of  $F$  closest to  $Q$ .

### A.2 RELEVANT DEFINITIONS

**Definition A.2.** If  $F : \mathbb{R}^n \rightarrow [0, \infty)$  satisfies  $\lim_{|x| \rightarrow \infty} F(x) = +\infty$ , we will say that  $F$  is coercive.

**Definition A.3.** For a coercive function  $F : \mathbb{R}^n \rightarrow [0, \infty)$  we let  $S_F = \{x \in \mathbb{R}^n : F(x) = \min_{y \in \mathbb{R}^n} F(y)\}$  be coercive.

**Lemma A.4.** *Assume that  $F : \mathbb{R}^n \rightarrow [0, \infty)$  is continuous and coercive. Then  $F$  has at least one global minimum. That is,  $S_F$  is non-empty. Furthermore,  $S_F$  is a compact set.*

**Definition A.5.** Let  $F, G : \mathbb{R}^n \rightarrow [0, \infty)$  are continuous and assume that  $F$  is coercive. Define

$$S_{F,G} = \{x \in S_F : G(x) = \inf_{y \in S_F} G(y)\},$$

the minima of  $F$  which minimize  $G$  among the minima of  $F$ .

**Definition A.6.** Let  $Q \subset \mathbb{R}^n$  be a closed set and assume that  $x \in \mathbb{R}^n$ . Define the distance from  $x$  to the set  $Q$  to be

$$d(x, Q) = \inf_{y \in Q} \|x - y\|.$$

Observe that since  $Q$  is a closed set we have that  $x \in Q$  if and only if  $d(x, Q) = 0$  and otherwise  $d(x, Q) > 0$ .

**Definition A.7.** Let  $A, B \subset \mathbb{R}^n$  be compact sets. We define the Hausdorff distance between  $A$  and  $B$  by

$$d_H(A, B) = \max\{\sup_{x \in B} d(x, A), \sup_{y \in A} d(y, B)\}.$$

Observe that  $d_H(A, B) = 0$  if and only if  $A = B$ .

**Definition A.8.** Let  $\{S_\delta\}_{\delta > 0}$  be a family of compact subsets of  $\mathbb{R}^n$ . We say that  $\lim_{\delta \rightarrow 0} S_\delta = S_*$  if

$$\lim_{\delta \rightarrow 0} d_H(S_\delta, S_*) = 0.$$

**Lemma A.9.** *Let  $S_\delta$  be a family of compact subsets of  $\mathbb{R}^n$ , then  $\lim_{\delta \rightarrow 0} S_\delta = S_*$  if and only if the following two conditions hold.*

1. *If  $x_\delta \in S_\delta$  converges to  $x$ , then  $x \in S_*$*
2. *For every  $x \in S_*$ , there exists a family  $x_\delta \in S_\delta$  with  $x_\delta \rightarrow x$ .*

The lemma is just an exercise in the definition.

### A.3 STATEMENT OF THE THEOREM

**Theorem 2.** *Let  $F, G : \mathbb{R}^n \rightarrow [0, \infty)$  are continuous and assume that  $F$  is coercive. Consider the sets  $S_{F+\delta G}$ , the set of points at which  $F + \delta G$  is globally minimum. The following are true:*

1. *If  $\delta_n \rightarrow 0$  and  $S_{F+\delta_n G} \rightarrow S_*$ , then*

$$S_* \subset S_{F,G}$$

2. *If  $\delta_n \rightarrow 0$  then there is a subsequence  $\delta_{n_k} \rightarrow 0$  and a non-empty set  $S_* \subset S_{F,G}$  so that  $S_{F+\delta_{n_k} G} \rightarrow S_*$ .*

*Proof.* The second statement follows from the standard theory of Hausdorff distance on compact metric spaces and the first statement. For the first statement, assume that  $S_{F+\delta_n G} \rightarrow S_*$ . We wish to show that  $S_* \subset S_{F,G}$ . Assume that  $x_n$  is a sequence of global minima of  $F + \delta_n G$  converging to  $x_*$ . It suffices to show that  $x_* \in S_{F,G}$ . First let us observe that  $x_* \in S_F$ . Indeed, let

$$\lambda = \inf_{x \in \mathbb{R}^n} F(x)$$

and assume that  $x \in S_F$ . Then,

$$\lambda \leq F(x_n) \leq (F + \delta_n G)(x_n) \leq (F + \delta_n G)(x) = \lambda + \delta_n G(x) \rightarrow \lambda.$$

Thus, since  $F$  is continuous and  $x_n \rightarrow x_*$  we have that  $F(x_*) = \lambda$  which implies  $x_* \in S_F$ . Next, define

$$\mu = \inf_{x \in S_F} G(x).$$

Let  $\hat{x} \in S_{F,G}$  so that  $G(\hat{x}) = \mu$ . Now observe that, by the minimality of  $x_n$  we have that

$$\lambda + \delta_n \mu = (F + \delta_n G)(\hat{x}) \geq (F + \delta_n G)(x_n) \geq \lambda + \delta_n G(x_n)$$

Thus,

$$G(x_n) \leq \mu$$

for all  $n$ . Since  $G$  is continuous and  $x_n \rightarrow x_*$  we have that  $G(x_*) \leq \mu$  which implies that  $G(x_*) = \mu$  since  $x_* \in S_F$ . Thus,  $x_* \in S_{F,G}$ .  $\square$

## B QUANTIZER

Here, we give an overview about the used quantization method. Consider a floating-point variable  $w_f$  to be mapped into a quantized domain using  $(b + 1)$  bits. Let  $\mathcal{Q}$  be a set of  $(2k + 1)$  quantized values, where  $k = 2^b - 1$ . Considering linear quantization,  $\mathcal{Q}$  can be represented as  $\{-1, -\frac{k-1}{k}, \dots, -\frac{1}{k}, 0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1\}$ , where  $\frac{1}{k}$  is the size of the quantization bin. Now,  $w_f$  can be mapped to the  $b$ -bit quantization (Zhou et al., 2016) space as follows:

$$w_{qo} = 2 \times \text{quantize}_b \left( \frac{\tanh(w_f)}{2 \max(|\tanh(W_f)|)} + \frac{1}{2} \right) - 1 \quad (\text{B.1})$$

where  $\text{quantize}_b(x) = \frac{1}{2^b-1} \text{round}((2^b - 1)x)$ ,  $w_f$  is a scalar,  $W_f$  is a vector, and  $w_{qo}$  is a scalar in the range  $[-1, 1]$ . Then, a scaling factor  $c$  is determined per layer to map the final quantized weight  $w_q$  into the range  $[-c, +c]$ . As such,  $w_q$  takes the form  $cw_{qo}$ , where  $c > 0$ , and  $w_{qo} \in \mathcal{Q}$ .

These learned parameters  $(b, \alpha)$ , as explained in Section 2.2, can be mapped to the quantizer parameters explained in Equation equation B.1. For  $(b + 1)$  bits quantization (the extra bit is the sign bit):

$$k = 2^b - 1, \quad \text{and} \quad c = \alpha = 2^b / 2^\beta \quad (\text{B.2})$$

Table 4: Hyperparameters settings.

Network	# Epochs			Batch Size	Learning Rate	Weight Decay	Momentum
	Phase 1	Phase 2	Phase 3				
SimpleNet on CIFAR10	59	74	44	128	0.001	0.0001	0.9
VGG11 on CIFAR10	83	104	62	128	0.0001	0.0001	0.9
SVHN-8 on SVHN	36	45	27	128	0.01	0.0001	0.9
ResNet-20 on CIFAR10	32	41	24	128	0.01	0.0001	0.9
AlexNet on ImageNet	28	35	21	256	0.01	0.0001	0.9
ResNet-18 on ImageNet	32	40	24	256	0.01	0.0001	0.9
MobileNet on ImageNet	41	52	31	256	0.01	0.0001	0.9

## C CONVERGENCE ANALYSIS

Figure 7 (a), (b) show the convergence behavior of WaveQ by visualizing both accuracy and regularization loss over finetuning epochs for two networks: CIFAR10 and SVHN. As can be seen, the regularization loss (WaveQ Loss) is minimized across the finetuning epochs while the accuracy is maximized. This demonstrates a validity for the proposed regularization being able to optimize the two objectives simultaneously. Figure 7 (c), (d) contrasts the convergence behavior with and without WaveQ for the case of training from scratch for VGG-11. As can be seen, at the onset of training, the accuracy in the presence of WaveQ is behind that without WaveQ. This can be explained as a result of optimizing for an extra objective in case of with WaveQ as compared to without. Shortly thereafter, the regularization effect kicks in and eventually achieves  $\sim 6\%$  accuracy improvement.

The convergence behavior, however, is primarily controlled by the regularization strengths  $(\lambda_w, \lambda_\beta)$ . As briefly mentioned in Section 2.2,  $(\lambda_w, \lambda_\beta) \in [0, \infty)$  is a hyperparameter that weights the relative contribution of the proposed regularization objective to the standard accuracy objective.

We reckon that careful setting of  $\lambda_w, \lambda_\beta$  across the layers and during the training epochs is essential for optimum results (Choi et al., 2018b).

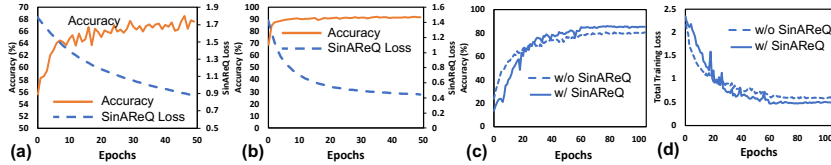


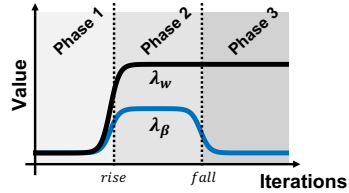
Figure 7: Convergence behavior: accuracy and WaveQ regularization loss over fine-tuning epochs for (a) CIFAR10, (b) SVHN. Comparing convergence behavior with and without WaveQ during training from scratch (c) accuracy, (d) training loss. Network: VGG-11, 2-bit DoReFa quantization

```

1 for i in range(0, iterations):
2     r = 200 # rise edge starting iteration
3     f = 500 # fall edge starting iteration
4     s = 10 # determines the smoothness of the transitions (rise/fall)
5     f1 = 0.5 * (1+tanh((i-r)/s));
6     f2 = 0.5 * (1+tanh((i-d)/s));
7     lambda_w_value = f1
8     lambda_beta_value = 0.05*(f1-f2)

```

(a)



(b)

Figure 8: Math formula for setting  $\lambda_w$  and  $\lambda_\beta$  during training iterations.

Table 5: Performance of WaveQ on BERT.

MODEL: CamemBERT BITWIDTH: W4.5/A8	SPOKEN		PARTUT	
	UPOS	LAS	UPOS	LAS
Baseline (FP)	96.99	81.37	97.65	93.43
Quantized w/ Unregularized Finetuning	89.91	72.32	90.89	84.25
<b>Quantized w/ WaveQ Regularized Finetuning</b>	<b>93.41</b>	<b>79.34</b>	<b>95.76</b>	<b>91.55</b>

Table 6: Validation top-1 accuracy for training from scratch w/ WaveQ vs w/o WaveQ.

	CIFAR10 (FP Accuracy = 74.53 %)			SVHN (FP Accuracy = 96.4 %)		
	3 bits	4 bits	5 bits	3 bits	4 bits	5 bits
Training w/o WaveQ	9.6	31.8	70.3	61.7	79.1	90.6
Training w/ WaveQ	44.8	66.6	73.2	79.3	85.1	94.8
Improvement (%)	(+35.2) ↑	(+34.8) ↑	(+2.9) ↑	(+17.6) ↑	(+6.0) ↑	(+4.2) ↑

## D WAVEQ PERFORMANCE ON BERT

Additionally, Table 5 provides layer-wise quantization with a heterogeneous mix of 4 and 5 bits for the BERT model. In all cases, WaveQ improves UPOS and LAS metrics for two French treebanks (SPOKEN, PARTUT).

## E TRAINING FROM SCRATCH

Table 6 shows a comparison between training from scratch with WaveQ vs without. It can be seen that incorporating WaveQ into the training process achieves strictly better accuracy than the baseline training without WaveQ across all cases. Moreover, higher improvements are obtained at lower bitwidths reaching to 35%

## F REGULARIZATION STRENGTHS

Having a regularization strength is a normal setting associated with any regularization method. The criterion for choosing  $\lambda_w$  and  $\lambda_\beta$  is to balance the magnitude of regularization loss to be smaller than the magnitude of accuracy loss. We then perform a grid search over a few points and chose the ones with the best convergence.

From the theoretical perspective, while the theorem is stated in terms of a limit as the regularization parameter vanishes, the proof in fact gives a corresponding stability result. Namely, if the regularization parameter is sufficiently small relative to the main loss then the minimizers will be “almost” quantized.