OPTIMAL QUANTIZATION USING SCALED CODEBOOK

Yerlan Idelbayev[1], **Pavlo Molchanov**[2], **Maying Shen**[2], **Hongxu Yin**[2], **Miguel Á. Carreira-Perpiñán**[1] and **Jose M. Alvarez**[2]

[1]Dept. CSE, University of California, Merced [2]NVIDIA Corporation

CVPR VIRTUAL JUNE 19-25

1 Abstract

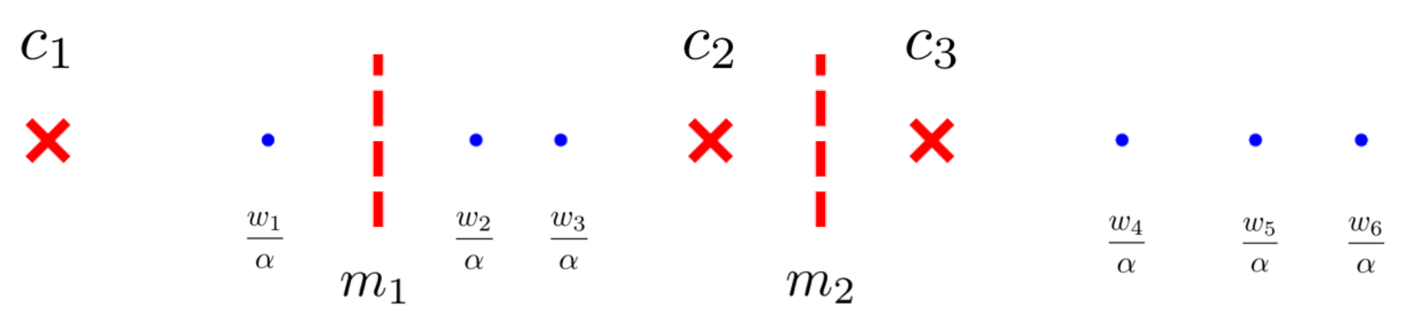
We study the problem of quantizing N sorted, scalar datapoints with a fixed codebook containing K entries that are allowed to be rescaled. By studying the properties of the optimal quantizer, we derive an $\mathcal{O}(NK \log K)$ algorithm that is guaranteed to find the optimal quantization parameters for any fixed codebook regardless of data distribution. We apply our algorithm to synthetic and real-world neural network quantization problems and demonstrate the effectiveness of our approach.

2 Problem formulation

Given a sorted data vector $\mathbf{w}$ with elements $w_1 \leq w_2 \leq \dots \leq w_N$ and a fixed codebook $\mathcal{C} = \{c_1, c_2, \dots, c_K\}$ such that $c_1 < c_2 < \dots < c_K$ we would like to learn the optimal α-rescaling of the codebook and the assignments ($\mathbf{Z}$) of datapoints into the rescaled codebook defined by the following MSE problem:

$$\min_{\alpha, \mathbf{z}_1, \dots, \mathbf{z}_N} \quad \textsc{Loss}(\alpha, \mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} (w_n - \alpha\, c_k)^2 \qquad (1)$$
$$\text{s.t.} \quad \mathbf{z}_n^T \mathbf{1} = 1, \quad \mathbf{z}_n \in \{0, 1\}^K.$$

Below is an illustration of the problem setup:



The problem includes many important cases:

1. symmetric INT8 quantization: $\mathcal{C} = \{0, \pm 1, \pm 2, \dots, \pm 127\}$
2. powers-of-two quantization: $\mathcal{C} = \{0, \pm 1, \pm 2^2, \pm 2^3, \dots, 2^s\}$
3. log-scale codebooks: $\mathcal{C} = \{0, \pm \log 2, \pm \log 3, \pm \log 4, \dots, \pm \log s\}$
4. and many others!

3 Locally optimal solutions

Let us analyze the loss wrt α and $\mathbf{Z}$ separately keeping other variable fixed.

- **Locally optimal scale** for a fixed $\mathbf{Z}$:

$$\min_{\alpha} \textsc{Loss}(\alpha, \mathbf{Z}) \implies \alpha = \textsc{OptScale}(\mathbf{Z})$$

- **Locally optimal assignments** for a fixed α:

$$\min_{\mathbf{Z}} \textsc{Loss}(\alpha, \mathbf{Z}) \implies \mathbf{Z} = \textsc{OptAssignment}(\alpha)$$

Exact forms of $\textsc{OptScale}(\mathbf{Z})$ and $\textsc{OptAssignment}(\alpha)$ are in the paper

4 Characterization of global solution

The global solution $(\alpha^*, \mathbf{Z}^*)$ must satisfy

$$\alpha^* = \textsc{OptScale}(\mathbf{Z}^*) \text{ and } \mathbf{Z}^* = \textsc{OptAssignment}(\alpha^*) \qquad (2)$$

There are many pairs that satisfy the equation above, but the one obtaining the minimal loss is the $(\alpha^*, \mathbf{Z}^*)$-pair we are looking for.

5 Important result and an initial algorithm

Lemma: The number of fixed points satisfying (2) is at most $NK + 1$.

Proof idea: Optimal assignments will remain constant for certain region of α-values (see paper)

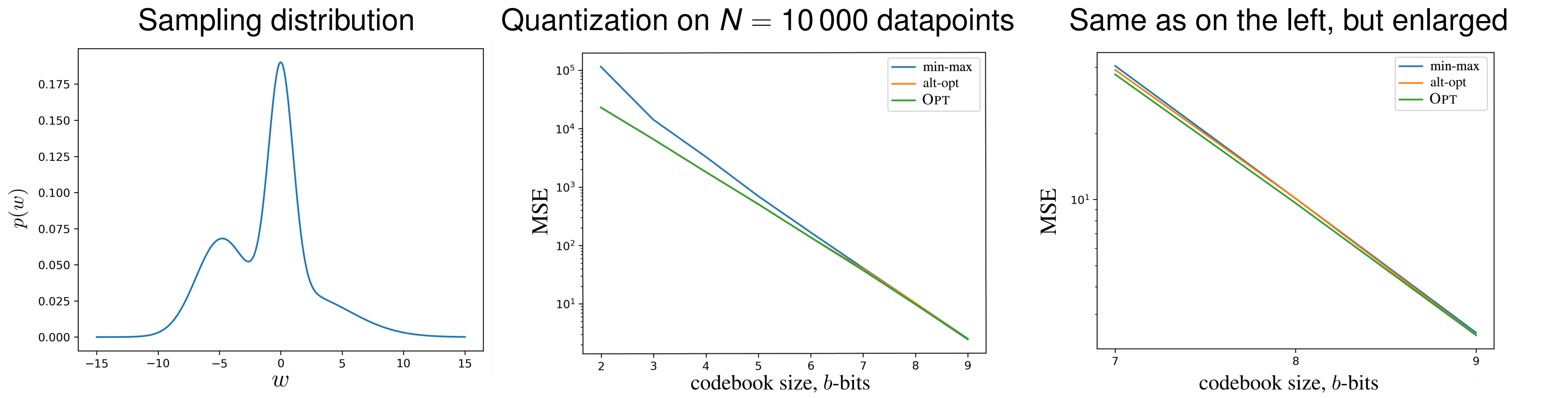
Corollary: There are only $NK + 1$ values of α we need to check!

Initial $\mathcal{O}(NK^2 \log N)$ algorithm enumerating possible α-values.

function $\textsc{OptQuant}(\mathbf{w}, \mathcal{C})$
  $\mathbf{Z}^c_{\min} = (\mathbf{z}^c_1, \dots, \mathbf{z}^c_K) \leftarrow (\mathbf{0}, \dots, \mathbf{0})$
  $L_{\min} \leftarrow \infty$, $\alpha_{\min} \leftarrow \infty$
  for $k = 1 \dots K - 1$ **do**
    $m_k = \frac{c_k + c_{k+1}}{2}$
  for $n = 1 \dots N$ **do**
    for $k = 1 \dots K - 1$ **do**
      $\alpha = \frac{w_n}{m_k}$
      $\mathbf{Z}^c \leftarrow \textsc{OptAssignment}(\alpha)$
      $\alpha \leftarrow \textsc{OptScale}(\mathbf{Z}^c)$
      $L \leftarrow \textsc{Loss}(\alpha, \mathbf{Z}^c)$
      if $L \leq L_{\min}$ **then**
        $L_{\min} \leftarrow L$, $\mathbf{Z}^c_{\min} \leftarrow \mathbf{Z}^c$, $\alpha_{\min} \leftarrow \alpha$
  return $\alpha_{\min}, \mathbf{Z}_{\min}$

6 Experiments on synthetic data

Quantization of $N = 10\,000$ datapoints sampled from the Gaussian mixture distribution. The probability density function of the distribution is plotted on the left. For this particular distribution our algorithm achieves optimal MSE error which outperforms other baselines (min-max and alt-opt). The gap between our optimal MSE and other methods is more visible when enlarged (on the right).



7 Experiments on neural network compression

INT8 quantization results, top-1 accuracy

		Min-max weight with activation calibration							our		
Model	FP32	Max	Entropy	99.9%	99.99%	99.999%	99.9999%	Q	Q+B+S	Q+B+Sv2
MobileNet-v1	71.88	69.84	70.99	70.76	70.89	70.67	70.44	70.78	**71.08**	71.12
MobileNet-v2	71.88	69.41	70.21	70.70	**71.02**	70.47	69.78	70.79	70.72	70.59
ResNet18	69.74	69.42	69.58	68.36	69.55	69.62	69.58	69.61	69.55	**69.63**
ResNet50	76.16	75.87	76.06	75.39	76.04	76.09	75.99	76.00	**76.10**	76.00

INT4 quantization results, top-1 accuracy

		Min-max weight with activation calibration						our		
Model	FP32	Max	Entropy	99.9%	99.99%	99.999%	99.9999%	Q	Q+B+S	Q+B+Sv2
MobileNet-v1	71.88	0.09	0.25	0.19	0.14	0.12	0.1	0.18	**6.38**	4.47
MobileNet-v2	71.88	0.11	0.47	0.38	0.13	0.11	0.08	3.49	**5.46**	4.24
ResNet18	69.74	0.84	42.81	41.07	31.30	19.19	7.15	44.74	57.85	**58.05**
ResNet50	76.16	0.21	52.44	53.45	37.03	7.40	0.93	53.81	63.74	**63.73**