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
Problem setup

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

$$\min_{\alpha, \mathbf{z}_1, \ldots, \mathbf{z}_N} \text{LOSS}(\alpha, \mathbf{Z}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk}(w_n - \alpha c_k)^2$$

$$\text{s.t. } \mathbf{z}_n^T \mathbf{1} = 1, \quad \mathbf{z}_n \in \{0, 1\}^K. \tag{1}$$

We give an $O(NK \log K)$ algorithm that finds the optimal solution of this problem.

The problem definition and the algorithm is general, includes many important cases:

1. symmetric INT8 quantization: $C = \{0, \pm 1, \pm 2, \ldots, \pm 127\}$
2. powers-of-two quantization: $C = \{0, \pm 1, \pm 2^2, \pm 2^3, \ldots, 2^s\}$
3. log-scale codebooks: $C = \{0, \pm \log 2, \pm \log 3, \pm \log 4, \ldots, \pm \log s\}$
4. and many others!
The problem we are solving is well-known, and many related work has been published.

- Previously, it was believed that problem (1) is hard to optimize [8] and requires exponential-time algorithm [13].
- Iterative or heuristic search algorithms without any optimality guarantees [1, 2, 8, 11, 13].
- Optimal algorithms for certain special cases are known: for binary codebook of $C = \{-1, 1\}$ [10], for ternary codebook of $C = \{-1, 0, 1\}$ [5, 12], or assuming some analytical distribution on $w$ [3, 4, 6, 7, 9].

We show that we can find a globally optimal solution in polynomial time $\mathcal{O}(NK \log K)$ for any codebook $C$ without any data distribution assumptions.
Overall idea of the algorithm

We build our algorithm by analyzing the properties of the optimal quantizer of (1). To do so, we analyze the loss wrt $\alpha$ and $Z$ separately.

- **Locally optimal scale** for a fixed $Z$:
  \[
  \min_{\alpha} \text{LOSS}(\alpha, Z) \implies \alpha = \text{OPTSCALE}(Z)
  \]

- **Locally optimal assignments** for a fixed $\alpha$:
  \[
  \min_{Z} \text{LOSS}(\alpha, Z) \implies Z = \text{OPTASSIGNMENT}(\alpha)
  \]

Exact forms of OPTSCALE($Z$) and OPTASSIGNMENT($\alpha$) are in the paper.
Clearly, the global solution \((\alpha^*, Z^*)\) must satisfy

\[
\alpha^* = \text{OPTSCALE}(Z^*) \quad \text{and} \quad Z^* = \text{OPTASSIGNMENT}(\alpha^*)
\]  

(2)

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

**Our contribution:** We show that there exists at most \(NK\) pairs of \((\alpha, Z)\) which satisfy the fixedpoint criterion above. We give a simple enumeration algorithm which iterates over these pairs and finds globally optimal solution in \(O(NK \log K)\). See paper for details.
References


