Optimal and Approximate Adaptive Stochastic Quantization

Ran Ben Basat[1], Yaniv Ben-Itzhak[2], Michael Mitzenmacher[3], Shay Vargaftik[2]

[1] UCL [2] VMware Research [3] Harvard University

1 - Summary

We revisit the Adaptive Stochastic Quantization (ASQ) problem and introduce QUIVER. QUIVER improves ASQ's computational complexity from $O(2^b \cdot d^2)$ to $O(2^b \cdot d)$ and memory usage from $O(d^2)$ to $O(2^b \cdot d)$. This efficiency is achieved by showing that the dynamic programming equations for the problem have special properties that allow for faster solutions, as well as specialized preprocessing and additional implementation optimizations. We also propose an approximation variant of QUIVER that strikes a balance between speed and precision, making it useful for quantizing large vectors on-the-fly.

2 - Introduction

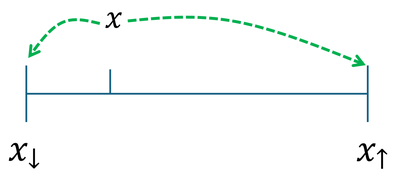
The vector quantization problem

Given a vector $X \in \mathbb{R}^d$, compress it to b bits per coordinate representation $\widehat{X}$ while minimizing

$$\mathbb{E}\left[\|X - \widehat{X}\|^2\right].$$

Essential for a wide range of ML applications, including gradient and model compression.

Stochastic quantization



Given $x \in \mathbb{R}$ and two quantization values $x_\downarrow \le x$ and $x_\uparrow \ge x$, SQ rounds x to $\hat{x} = x_\uparrow$ w.p. $\frac{x - x_\downarrow}{x_\uparrow - x_\downarrow}$ and to $\hat{x} = x_\downarrow$ otherwise. Importantly, it is unbiased, i.e., $\mathbb{E}[\hat{x}] = x$ and satisfies $\mathrm{Var}[\hat{x}] = (x_\uparrow - x)(x - x_\downarrow)$.

Stochastic quantization of a vector

Given a set $Q \subset \mathbb{R}$ such that $\max Q \ge \max X$ and $\min Q \le \min X$, denote for each $x \in X$:

$$x_\downarrow = \max\{q \in Q \mid q \le x\}.$$
$$x_\uparrow = \min\{q \in Q \mid q \ge x\}.$$

Apply stochastic quantization for each $x \in X$ independently.

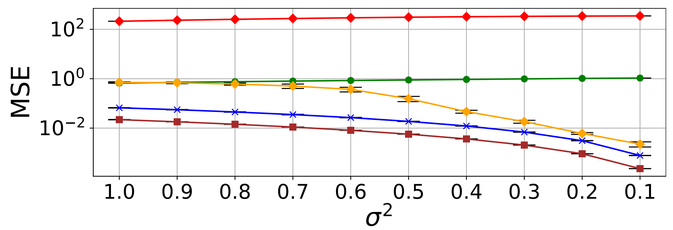
Adaptive stochastic quantization

Given a **sorted** vector $X \in \mathbb{R}^d$, **pick** $Q \subset \mathbb{R}$ of size $|Q| = 2^b$ that minimizes $\mathbb{E}\left[\|X - \widehat{X}\|^2\right]$ when applying stochastic quantization.

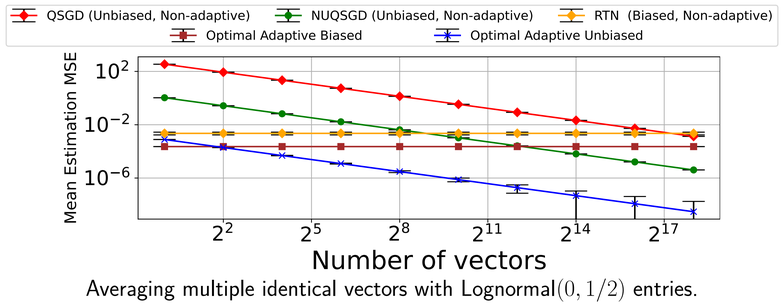
This improves over input-agnostic selection of quantization values at the cost of additional computation.

Our goal is to compute the set Q efficiently.

3 - Why adaptive and why unbiased?



A single vector with i.i.d Lognormal$(0, \sigma^2)$ entries.



Averaging multiple identical vectors with Lognormal$(0, 1/2)$ entries.

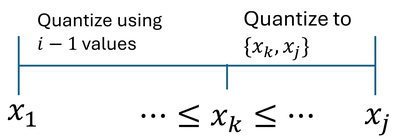
- Adaptive methods have lower error.
- Unbiased solutions' error diminishes with the # of averaged vectors.

4 - The QUIVER algorithm

Denote $MSE[i, j]$ the optimal MSE of quantizing the prefix vector $X_j = x_1, \ldots, x_j$ using i quantization values *that include* x_j, that is:

$$MSE[i, j] = \min_{Q : |Q| \le i, x_j \in Q} \sum_{x \in X_j} (x_\uparrow - x)(x - x_\downarrow).$$

ZipML (ICML 2017) observed that there exists an optimal Q for which $Q \subseteq X$, allowing them to define the following dynamic program:



$$MSE[i, j] = \min_{k \in \{i, \ldots, j\}} MSE[i-1, k] + C[k, j] \ ,$$

where $C[k, j] = \sum_{x \in \{x_k, \ldots, x_j\}} (x_j - x)(x - x_k)$, and solve it in $O(2^b \cdot d^2)$ time and $O(d^2)$ space complexity.

To accelerate the computation, QUIVER defines a square matrix $A \in \mathbb{R}^{d \times d}$ for which $A[k, j] = MSE[i-1, k] + C[k, j]$.

A is ***implicit*** and is never computed or stored in memory, but our methods allow evaluating each $A[k, j]$ in constant time.

We prove that A is a totally monotone matrix and we can thus find its row minimas using the SMAWK algorithm, yielding $MSE[i, \cdot]$ from $MSE[i-1, \cdot]$ in $O(d)$ time.

5 - Approximate QUIVER

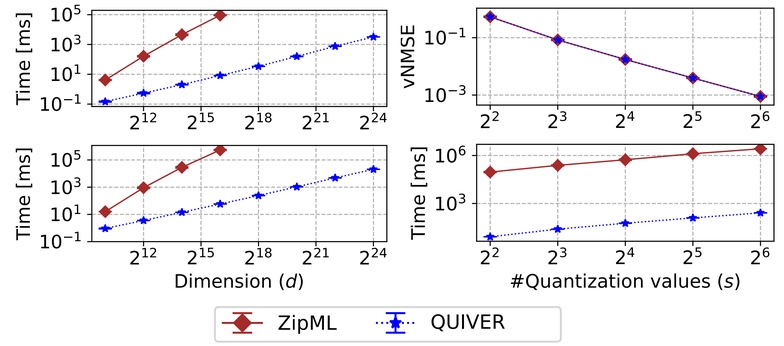
We split the range $[x_1, x_d]$ into m equal-sized intervals to define the set

$$S = \{x_1 + \ell \cdot \frac{x_d - x_1}{m} \mid \ell \in \{0, \ldots, m\}\} \ .$$

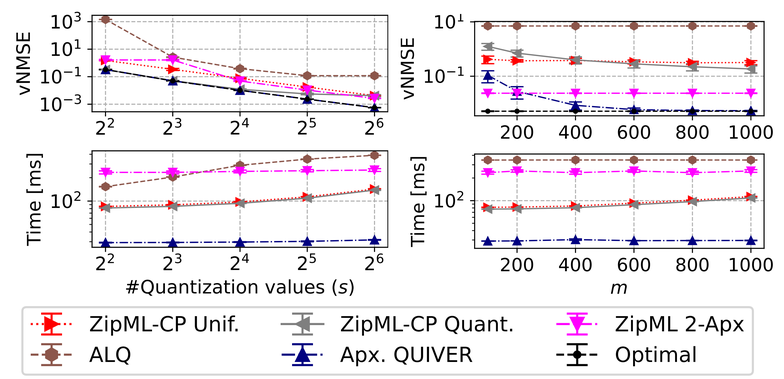
Approximate QUIVER finds the set $Q \subseteq S$ that minimizes $\mathbb{E}\left[\|X - \widehat{X}\|^2\right]$ when applying stochastic quantization in just $O(d + m \cdot 2^b)$ time and space complexity, while providing rigorous accuracy guarantees.

6 - Empirical results

We evaluate vectors with i.i.d. LogNormal(0,1) entries.
Exact QUIVER is optimal and is orders of magnitude faster than ZipML.
Here, upper left is $s = 4$, bottom left is $s = 16$, and the figures on the right have $d = 2^{16}$.



Approximate QUIVER is both faster and more accurate than previous approximate solutions. With small m values, it already has accuracy comparable with an optimal solution. (Shown for $d = 2^{22}$. The figures on the left have $m = 1000$ and those on the right have $s = 32$.)



7 - Summary

Fast optimal and very fast approximate adaptive stochastic vector quantization is possible, and there are many applications (e.g., gradient compression, quantization for faster training, LLM KV cache compression). Plenty more results in the paper.