Faster Binary Embeddings for Preserving Euclidean Distances

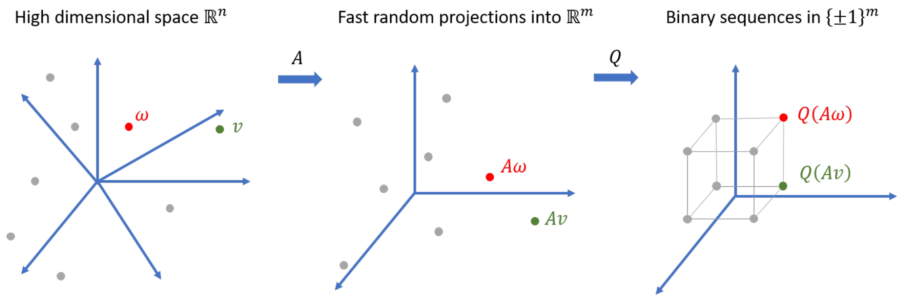
Jinjie Zhang, Rayan Saab

Department of Mathematics, Halıcıoğlu Data Science Institute

University of California San Diego

Wednesday 24th March, 2021

Framework



Euclidean distance recovery: $d(Q(A\omega), Q(Av)) \approx \| \omega - v \|_2$

- Our method is

$$q_x := Q(Ax)$$

where Q is a stable $\Sigma\Delta$ quantization scheme, $A \in \mathbb{R}^{m \times n}$ is a sparse Gaussian matrix and $x \in \mathbb{R}^n$ is well-spread, i.e., those that are not sparse.

Sigma-Delta Quantization Schemes Q

For simplicity, we focus on the quantizer $Q : \mathbb{R}^m \to \{1, -1\}^m$. Specifically, $q = Q(x)$ such that for $i = 1, ..., m$

$$\begin{cases} u_0 = 0, \\ q_i = \text{sign}(x_i + u_{i-1}), \\ u_i = u_{i-1} + x_i - q_i. \end{cases} \tag{1}$$

Algorithm 1: Fast Binary Embedding for Finite $\mathcal{T}$

Input: $\mathcal{T} = \{x^{(j)}\}_{j=1}^k \subseteq B_2^n(\kappa)$ ▷ Data points in ℓ_2 ball
Generate $A \in \mathbb{R}^{m \times n}$ ▷ Sparse Gaussian matrix A
for $j \leftarrow 1$ **to** k **do**
  $z^{(j)} \leftarrow Ax^{(j)}$
  $q^{(j)} = Q(z^{(j)})$ ▷ Stable $\Sigma\Delta$ quantizer Q as in (1)
Output: Binary sequences $\mathcal{B} = \{q^{(j)}\}_{j=1}^k \subseteq \{-1, 1\}^m$

ℓ_2 Distance Recovery

Algorithm 2: ℓ_2 Norm Distance Recovery

Input: $q^{(i)}, q^{(j)} \in \mathcal{B}$ ▷ Binary sequences produced by Algorithm 1
$y^{(i)} \leftarrow \widetilde{V} q^{(i)}$ ▷ Condense the components of q
$y^{(j)} \leftarrow \widetilde{V} q^{(j)}$
Output: $\|y^{(i)} - y^{(j)}\|_1$ ▷ Approximation of $\|x^{(i)} - x^{(j)}\|_2$

Definition (Condensation operator)

Let p, r, λ be fixed positive integers such that $\lambda = r\widetilde{\lambda} - r + 1$ for some integer $\widetilde{\lambda}$. Let $m = \lambda p$ and v be a row vector in $\mathbb{R}^\lambda$ whose entry v_j is the j-th coefficient of the polynomial $(1 + z + \ldots + z^{\widetilde{\lambda}-1})^r$. Define the normalized condensation operator $\widetilde{V} \in \mathbb{R}^{p \times m}$ by

$$\widetilde{V} = \frac{\sqrt{\pi/2}}{p\|v\|_2} \begin{bmatrix} v & & \\ & \ddots & \\ & & v \end{bmatrix}.$$

Main Result

Theorem (Main result)

Let $\mathcal{T} \subseteq \mathbb{R}^n$ be a finite, appropriately scaled set with elements satisfying $\|x\|_\infty = O(n^{-1/2}\|x\|_2)$. If $m \gtrsim p := \Omega(\epsilon^{-2}\log(|\mathcal{T}|^2/\delta))$ and $r \geq 1$ is the integer order of Q, then with probability $1 - 2\delta$ on the draw of the sparse Gaussian matrix A, the following holds uniformly over all x, y in $\mathcal{T}$: Embedding x, y into $\{-1, 1\}^m$ using Algorithm 1, and estimating the associated distance between them using Algorithm 2 yields the error bound

$$\left| \|\widetilde{V}(q_x - q_y)\|_1 - \|x - y\|_2 \right| \leq c\left(\frac{m}{p}\right)^{-r+1/2} + \epsilon\|x - y\|_2.$$

- The assumption that $\|x\|_\infty = O(n^{-1/2}\|x\|_2)$ is reasonable.
- The latter part in error bound is essentially proportional to $p^{-1/2}$.

Comparisons with Baselines

Method	Time	Space	Storage	Query Time
Toeplitz [1]	$O(n \log n)$	$O(n)$	$O(m)$	$O(m)$
Bilinear [2]	$O(n\sqrt{m})$	$O(\sqrt{mn})$	$O(m)$	$O(m)$
Circulant [3]	$O(n \log n)$	$O(n)$	$O(m)$	$O(m)$
BOE or PCE [4]	$O(n \log n)$	$O(n)$	$O(p \log_2 \lambda)$	$O(p \log_2^2 \lambda)$
Our Algorithm [5]	$O(m)$	$O(m)$	$O(p \log_2 \lambda)$	$O(p \log_2 \lambda)$

- "Time" is the time needed to embed a data point;
- "Space" is the space needed to store the embedding matrix;
- "Storage" contains the memory usage to store each encoded sequence;
- "Query time" is the time complexity of pairwise distance estimation.

Experiments on different Image Datasets

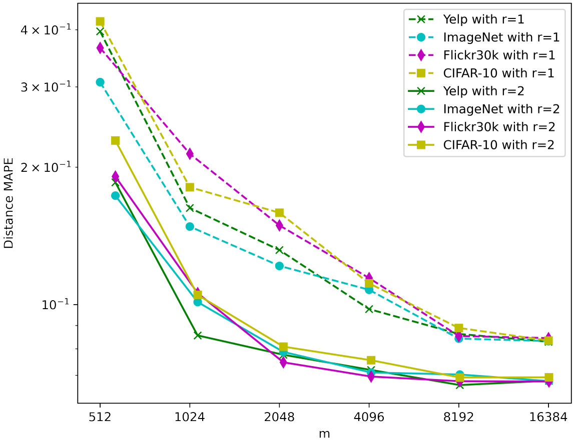


Figure: Plot of ℓ_2 distance reconstruction error on four datasets

References

X. Yi, C. Caramanis, and E. Price.
Binary embedding: Fundamental limits and fast algorithm.
In International Conference on Machine Learning, 2162 – 2170, 2015.

Y. Gong, S. Kumar, H. A Rowley, and S. Lazebnik.
Learning binary codes for high-dimensional data using bilinear projections.
In Proceedings of the IEEE conference on computer vision and pattern recognition, 484–491, 2013.

F. Yu, S. Kumar, Y. Gong, and S. Chang.
Circulant binary embedding.
In International conference on machine learning, 946–954, 2014.

T. Huynh and R. Saab.
Fast binary embeddings and quantized compressed sensing with structured matrices.
Communications on Pure and Applied Mathematics, 73(1):110–149, 2020.

Thank You