# Optimal compression of approximate inner products and dimension reduction

Noga Alon Bo'az Klartag

## 1 Main result

Let $X$ be a set of $n$ points of norm at most 1 in the Euclidean space $R^k$, and suppose $\epsilon > 0$. An *$\epsilon$-distance sketch for $X$* is a data structure that, given any two points of $X$ enables one to recover the square of the Euclidean distance between them, and their inner product, up to an additive error of $\epsilon$.
Let $f(n, k, \epsilon)$ denote the minimum possible number of bits of such a sketch.

**Theorem 1.** *For all $n$ and $\frac{1}{n^{0.49}} \leq \epsilon \leq 0.1$ the function $f(n, k, \epsilon)$ satisfies the following*

- *For $\frac{\log n}{\epsilon^2} \leq k \leq n$,*
$$f(n, k, \epsilon) = \Theta\left(\frac{n \log n}{\epsilon^2}\right)$$
- *For $\log n \leq k \leq \frac{\log n}{\epsilon^2}$,*
$$f(n, k, \epsilon) = \Theta\left(nk \log\left(2 + \frac{\log n}{\epsilon^2 k}\right)\right)$$
- *For $1 \leq k \leq \log n$,*
$$f(n, k, \epsilon) = \Theta\left(nk \log\left(1/\epsilon\right)\right).$$

## 2 Definitions

### 2.1 Gram matrices

For $n$ vectors $w_1, \ldots, w_n$ the Gram matrix $G(w_1, \ldots, w_n)$ is the $n$ by $n$ matrix G given by $G(i, j) = \langle w_i, w_j \rangle$. We say that two Gram matrices $G_1$, $G_2$ are *$\epsilon$-separated* if there are two indices $i \neq j$ so that $|G_1(i, j) - G_2(i, j)| > \epsilon$.

Let $\mathcal{G}$ be a maximal (with respect to containment) set of $\epsilon$-separated Gram matrices of ordered sequences of $n$ vectors $w_1, \ldots, w_n$ in $R^m$, where the norm of each vector $w_i$ is at most $k$. Then by maximality of $\mathcal{G}$, for every Gram matrix $M$ of vectors of norm at most $k$ in $R^m$ there is a member of $\mathcal{G}$ in which all inner product of pairs of distinct points are within $\epsilon$ of the corresponding inner products in $M$. Therefore we can use an index of an appropriate member of $\mathcal{G}$ as a sketch for $M$, requiring $\log |\mathcal{G}|$ bits.

### 2.2 $\delta$-nets

For $0 < \delta < 1/4$ and for $k \geq 1$ a $\delta$-net, denoted by $N(k, \delta)$, be the set of all vectors of Euclidean norm at most 1 in which every coordinate is an integral multiple of $\frac{\delta}{\sqrt{k}}$. Given a vector in the unit ball in $R^k$ we can round it to a vector in the net that lies within distance $\delta/2$ from it by simply rounding each coordinate.

Each point of $N(k, \delta)$ can be represented by at most $k \log(1/\delta) + 2k$ bits as the size of $N(k, \delta)$ has size $(1/\delta)^k 2^{O(k)}$.

# 3 Upper bounds

**Lemma 2.** *For* $\frac{\log n}{\epsilon^2} \leq k \leq n,\ f(n,k,5\epsilon) = O\left(\frac{n \log n}{\epsilon^2}\right)$.

Use Johnson-Lindenstrauss Lemma to reduce dimension to $C\frac{\log n}{\epsilon^2}$ $\rightarrow$ encode inner products using maximal set $\mathcal{G}$ of $\epsilon$-separated Gram matrices $\rightarrow$ show that $\mathcal{G}$ is "small".

**Lemma 3.** *For* $\log n \leq k \leq \frac{\log n}{\epsilon^2},\ f(n,k,4\epsilon) = O\left(nk \log\left(2 + \frac{\log n}{\epsilon^2 k}\right)\right)$

Similar to Lemma 2, except the initial usage of Johnson-Lindenstrauss Lemma.

# 4 Algorithmic proof

For $\frac{40 \log n}{\epsilon^2} \leq k \leq n$, apply Johnson-Lindenstrauss Lemma to $m = 40 \log n/\epsilon^2$. Then for $w_i \in X$ round each coordinate to an integral multiple of $1/\sqrt{m}$ $\rightarrow$ random vector $V_i$. Suppose the $j$-th coordinate of $w_i$ is $\frac{s+p}{\sqrt{m}}$ for $s \in \mathbb{Z}$ and $0 \leq p < 1$, then

$$V_i(j) = \begin{cases} \frac{s}{\sqrt{m}} & \text{with probability } 1-p, \\ \frac{s+1}{\sqrt{m}} & \text{with probability } p. \end{cases}$$

For $\log n \geq k \leq \frac{40 \log n}{\epsilon^2}$, let $\delta$ be such that $k = \frac{40\delta^2 \log n}{\epsilon^2}$. Round similarly as before, this time to points of $N(k, \delta)$.

# 5 Lower bounds

**Lemma 4.** *If* $k = \delta^2 \log n/(200\epsilon^2)$ *where* $2\epsilon \leq \delta \leq 1/2$*, then* $f(n,k,\epsilon/2) = \Omega(kn \log(1/\delta))$.

Fix maximal set of point $N$ in the unit ball with pairwise distances at least $\delta$ $\rightarrow$ find set $R$, $|R| = n/2$ such that for any $N_1, N_2 \subset N$ with $|N_1| = |N_2| = n/2$, the matrices $G(R, N_1)$ and $G(R, N_2)$ are $\epsilon$-separated $\rightarrow$ use size of $N$ to bound $f(n,k,\epsilon)$ from below.

# 6 Known results

**Theorem 5** (Johnson-Lindenstrauss Lemma)**.** *Let* $X \subset R^k, |X| = n$ *and* $0 < \epsilon \leq 1/2$*. Then there exists map* $f\colon X \rightarrow R^m$ *for some* $m = O(\frac{\log n}{\epsilon^2})$ *such that*

$$\forall x, y \in X, (1-\epsilon)\|x-y\|^2 \leq \|f(x) - f(y)\|^2 \leq (1+\epsilon)\|x-y\|^2$$

.

*Moreover, there is a probabilistic algorithm that outputs the map in time* $O(\frac{\log^3 n}{\epsilon^2})$.

**Theorem 6** (Hoeffding's Inequality)**.** *If* $X_1, \ldots, X_n$ *are independent and* $a_i \leq X_i \leq b_i$ *for every* $i$*, then for* $t > 0$

$$\Pr\left[\sum_{i=1}^n X - \mu > t\right] \leq e^{-2t^2/\sum(b_i - a_i)^2}.$$