Optimal compression of approximate inner products and dimension reduction

Bo'az Klartag

Weizmann Institute & Tel Aviv University

FOCS 2017, Berkeley, California

Joint work with Noga Alon

Sketching inner products

- We are given a set X of n points in the unit ball of $\mathbb{R}^k$, and an accuracy parameter $\varepsilon > 0$.

Definition

An ε**-sketch** for X is a data structure that given any query of the form $x, y \in X$ outputs a number α with

$$|\alpha - \langle x, y \rangle| < \varepsilon.$$

- Equivalently, we may approximate squares of the distances.

Questions:

1. What is the minimal number of bits used by such a sketch?
2. Can we implement it efficiently?

The Johnson-Lindenstrauss lemma

As a side-effect of their work on Lipschitz extension, J & L have found a sketch based on dimension reduction:

An excellent ε-sketch (1980s)

Pick a random ℓ-dimensional subspace E, and store the (discretized) projections of the points of X onto this subspace, where

$$\ell = \Theta\left(\frac{\log n}{\varepsilon^2}\right).$$

- **Concentration of measure phenomenon:** With high probability of selecting E,

$$\forall x, y \in X, \quad \left| \frac{n}{\ell} \cdot \langle Proj_E x, Proj_E y \rangle - \langle x, y \rangle \right| < \varepsilon$$

- **Larsen and Nelson '16**: Assuming $\varepsilon \geq n^{-0.49}$, the estimate for the dimension ℓ is tight, even if we are only interested in the existence of a subspace E.

Size of the best sketch

- Write $f(n,k,\varepsilon)$ for the number of bits in the optimal ε-sketch. (Recall: A set X of n points in the unit ball of $\mathbb{R}^k$).

Theorem 1

Assume $n^{-0.49} \leq \varepsilon \leq 1/2$. Then,

$$f(n,k,\varepsilon) = \begin{cases} \Theta\left(nk \log\left(1/\varepsilon\right)\right) & 1 \leq k \leq \log n \\ \Theta\left(nk \log\left(2 + \frac{\log n}{\varepsilon^2 k}\right)\right) & \log n \leq k \leq \frac{\log n}{\varepsilon^2} \\ \Theta\left(n\frac{\log n}{\varepsilon^2}\right) & \frac{\log n}{\varepsilon^2} \leq k \leq n \end{cases}$$

- We also provide an algorithm, query time $O(f(n,k,\varepsilon)/n)$.
- In the “Johnson-Lindenstrauss” range $k \geq \varepsilon^{-2} \log n$, our result follows from Kushilevitz, Ostrovsky and Rabani ’98.

An information theoretic point of view on $f(n, k, \varepsilon)$

1. The Gram matrix of $x_1, \ldots, x_n \in B^k = \{x \in \mathbb{R}^k ; \|x\| \le 1\}$ is
$$G(x_1, \ldots, x_n) = \left\{ \langle x_i, x_j \rangle \right\}_{i,j=1,\ldots,n}$$
2. The distance between two matrices $G, H \in \mathbb{R}^{n \times n}$ is
$$d(G, H) = \max_{ij} |G_{ij} - H_{ij}|$$
3. Information bound: $f(n, k, \varepsilon)$ is the logarithm of the size of the minimal ε-net in this space of Gram matrices.

How do we get the lower bound on $f(n, k, \varepsilon)$?

We need an ε-separated set of Gram matrices. Our choice: A fixed set of $n/2$ unit vectors (selected randomly), plus all $n/2$-subsets of an arbitrary δ-separated set in S^{n-1}. Here,
$$\delta^2 = \min\{1, \max\{k/t, \varepsilon^2\}\}, \qquad t = \varepsilon^{-2} \log(\varepsilon^2 n).$$

A comment on the clumsy assumption $\varepsilon \geq n^{-0.49}$

Like Kasper and Nelson, we think that the "log n" should be replaced by "$\log(\varepsilon^2 n)$" in the J-L dimension $\ell = \varepsilon^{-2} \log n$.

Theorem 1'

For any $\varepsilon > 2/\sqrt{n}$, set $t = \varepsilon^{-2} \log(2 + \varepsilon^2 n)$. Then,

$$f(n,k,\varepsilon) = \begin{cases} \Theta\left(nk \log\left(1/\varepsilon\right)\right) & 1 \leq k \leq \log(\varepsilon^2 n) \\ \Theta\left(nk \log\left(2 + \frac{t}{k}\right)\right) & \log(\varepsilon^2 n) \leq k \leq t \\ \Omega(nt) \ \ \& \ \ O\left(\frac{n \log n}{\varepsilon^2}\right) & t \leq k \leq n \end{cases}$$

- Recovers Larsen-Nelson, wider range of the parameters.
- We think that the lower bound is tight for any $\varepsilon > 2/\sqrt{n}$.
- Our upper bound idea of "linear projection followed by random rounding" is non-optimal when decreasing dimensions by a constant factor.

Better constant-factor dimension reduction

Theorem 2 (bipartite version, non-linear embedding)

Let $a_1, \dots, a_n, b_1, \dots, b_n \in B^{2n} \subseteq \mathbb{R}^{2n}$, let $0 < \varepsilon < 1$. Assume

$$t = \Omega\left(\frac{\log(2 + \varepsilon^2 n)}{\varepsilon^2}\right).$$

Then there exist $x_1, \dots, x_n, y_1, \dots, y_n \in \mathbb{R}^t$ such that

$$\left|\langle x_i, y_j\rangle - \langle a_i, b_j\rangle\right| \leq \varepsilon \qquad (i, j = 1, \dots, n).$$

(Moreover, when $t = \Omega(n)$ also $\|x_i\| + \|y_i\| = O(1)$ for all i).

- Proof relies on an improved "**low M^*-estimate**" (following Gluskin, Gordon, Milman, Pajor, Tomczak, '80s).
- An efficient algorithm using linear programming.
- **Conjecture:** We can find x_i's and y_i's such that additionally

$$\|x_i\| + \|y_i\| \leq O(1) \qquad (i = 1, \dots, n)$$

The upper bound for $f(n, k, \varepsilon)$

- If correct, this conjecture implies the correct asymptotics for $f(n, k, \varepsilon)$ for all values of $\varepsilon > 1/\sqrt{n}$.

In the range $\varepsilon \geq n^{-0.49}$, our tight upper bounds for $f(n, k, \varepsilon)$ are based on the idea of "projection and randomized rounding".

- Given $w_1, \ldots, w_n \in B^k$ and $\varepsilon \geq n^{-0.49}$. How to sketch?

Step 1. Set $m = \min\{k, 40\varepsilon^{-2} \log n\}$. If $k \geq m$, then apply the Johnson-Lindenstrauss lemma, and project the data to $\mathbb{R}^m$.

- May use the fast J-L algorithm of Ailon and Chazelle '09.
- All scalar products are preserved within an additive error of at most ε.
- Next step: If we just round each (projected) point to a closest neighbor in an ε-net, we lose a factor of $\log(1/\varepsilon)$.

Randomized rounding

Balanced random rounding to a multiple of λ

Given $x \in \mathbb{R}$ and a resolution parameter $\lambda > 0$. Define

$$R_\lambda(x) = \begin{cases} i \cdot \lambda & \text{probability } 1-p \\ (i+1)\cdot\lambda & \text{probability } p \end{cases}$$

where $x = (i+p)\cdot\lambda$ and $0 \le p \le 1$. Thus $\mathbb{E}R_\lambda(x) = x$.

- Denote the (projected) points by $w_1, \ldots, w_m \in 2B^m$.

Step 2. Set $\lambda = 1/\sqrt{m}$. Apply balanced random rounding to each coordinate of each w_i, to obtain $V_i \in \frac{1}{\sqrt{m}} \cdot \mathbb{Z}^m$.

- For each i, store $\sqrt{m} \cdot V_i$ (full binary representations), additionally store $|w_i|^2$ to an accuracy ε.

Recovering a scalar product

Memory usage as advertised, since for $v \in 2B^m \cap \frac{1}{\sqrt{m}} \cdot \mathbb{Z}^m$, total length of binary representation of all coordinates is $O(m)$.

- Is it true that with high probability, for all i and j,

$$|\langle V_i, V_j \rangle - \langle w_i, w_j \rangle| < \varepsilon?$$

Answer

Yes, but only if $i \neq j$. (This is why we stored $|w_i|^2$ separately).

- Indeed,

$$|\langle V_i, V_j \rangle - \langle w_i, w_j \rangle| \leq |\langle V_i - w_i, w_j \rangle| + |\langle V_i, V_j - w_j \rangle|$$

and $\langle V_i - w_i, \theta \rangle$ has mean zero, variance at most $|\theta|^2$ and a subgaussian tail (by Hoeffding's inequality) ...

- ... But only if θ is constant or independent of V_i.

Avoiding the union bound

- When estimating probabilities, we apply the union bound, following the footsteps of J & L.
- Harmless if $\varepsilon \geq n^{-0.49}$, but otherwise it seems non-optimal.
- Perhaps we prefer to replace the discrete "randomized rounding" by Gaussians, to make analysis easier.

Theorem 3 (bipartite, constant-factor dimension reduction)

Let $a_1, \ldots, a_n, b_1, \ldots, b_n \in B^{5k}$ and let $\varepsilon > 1/\sqrt{n}$.
Let $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ be i.i.d standard Gaussians in $\mathbb{R}^k$.

Assume $k \geq C\varepsilon^{-2}\log(2 + \varepsilon^2 n)$. Then with prob. of at least $\exp(-ckn)$, setting $\bar{X}_i = X_i/\sqrt{k}$ and $\bar{Y}_j = Y_j/\sqrt{k}$,

$$\forall i, j \qquad \left|\left\langle \bar{X}_i, \bar{Y}_j \right\rangle - \left\langle a_i, b_j \right\rangle\right| < \varepsilon$$

and moreover $\|\bar{X}_i\| + \|\bar{Y}_i\| = O(1)$.

- Probability is tiny, but positive. Recovers size of ε-net.

Deeper mathematical tools

- Our accurate results, where "log *n*" is replaced by "$\log(\varepsilon^2 n)$", use some math tools, and avoid union bounds.

Theorem (Gaussian correlation inequality, Royen '14)

Let $A_1, \ldots, A_N \subseteq \mathbb{R}^n$ be centrally-symmetric convex sets, let Z be Gaussian random vector in $\mathbb{R}^n$ with $\mathbb{E}Z = 0$. Then

$$\mathbb{P}(\forall i, Z \in A_i) \geq \prod_{i=1}^{N} \mathbb{P}(Z \in A_i).$$

- In our case, we only need the case of slabs (Khatri-Sidak '60s), and the case of ellipsoids (Hargé '99).
- For the proof of Theorem 3, we also use the "**finite volume-ratio theorem**" of Szarek and Tomczak '80.

The end

Thank you!