Abstract

We show tight upper and lower bounds for time–space trade-offs for the c-approximate Near Neighbor Search problem. For the d-dimensional Euclidean space and n-point datasets, we develop a data structure with space $n^{1+\rho_u+o(1)}+O(dn)$ and query time $n^{\rho_q+o(1)}+dn^{o(1)}$ for every $\rho_u, \rho_q \geq 0$ with:

$$c^2\sqrt{\rho_q} + (c^2-1)\sqrt{\rho_u} = \sqrt{2c^2-1}. \quad (1)$$

In particular, for the approximation $c = 2$ we get:

- Space $n^{1.77\ldots}$ and query time $n^{o(1)}$, significantly improving upon known data structures that support very fast queries [Indyk, Motwani 1998] [Kushilevitz, Ostrovsky, Rabani 2000];
- Space $n^{1.14\ldots}$ and query time $n^{0.14\ldots}$, matching the optimal data-dependent Locality-Sensitive Hashing (LSH) from [Andoni, Razenshteyn 2015];
- Space $n^{1+o(1)}$ and query time $n^{0.43\ldots}$, making significant progress in the regime of near-linear space, which is arguably of the most interest for practice [Lv, Josephson, Wang, Charikar, Li 2007].

This is the first data structure that achieves sublinear query time and near-linear space for *every* approximation factor $c > 1$, improving upon [Kapralov 2015]. The data structure is a culmination of a long line of work on the problem for all space regimes; it builds on Spherical Locality-Sensitive Filtering [Becker, Ducas, Gama, Laarhoven 2016] and data-dependent hashing [Andoni, Indyk, Nguyen, Razenshteyn 2014][Andoni, Razenshteyn 2015]. Our matching lower bounds are of two types: conditional and unconditional. First, we prove tightness of the *whole* trade-off in a restricted model of computation, which captures all known hashing-based approaches. We then show *unconditional* cell-probe lower bounds for one and two probes that match the trade-off for $\rho_q = 0$, improving upon the best known lower bounds from [Panigrahy, Talwar, Wieder 2010]. In particular, this is the first space lower bound (for *any* static data structure) for two probes which is not polynomially smaller than the one-probe bound. To show the result for two probes, we establish and exploit a connection to *locally-decodable codes*.