From Average Embeddings To Nearest Neighbor Search*

Alexandr Andoni  
CS and DSI  
Columbia University

David Cheikhi  
Columbia University

May 13, 2021

Abstract

In this note, we show that one can use average embeddings, introduced recently in [Nao20], to obtain efficient algorithms for approximate nearest neighbor search. In particular, a metric $X$ embeds into $\ell_2$ on average, with distortion $D$, if, for any distribution $\mu$ on $X$, the embedding is $D$ Lipschitz and the (square of) distance does not decrease on average (wrt $\mu$). In particular existence of such an embedding (assuming it is efficient) implies an $O(D^3)$ approximate nearest neighbor search under $X$. This can be seen as a strengthening of the classic (bi-Lipschitz) embedding approach to nearest neighbor search, and is another application of data-dependent hashing paradigm.

1 Introduction

In the $c$-Approximate Nearest Neighbor Search ($c$-ANN) problem, we preprocess a dataset $P$ living in some metric space $(X, d_X)$ for some threshold $r > 0$, such that, for a given query $q \in X$, we can efficiently find a point in $p \in P$ at distance $d_X(p, q) \leq cr$, as long as there exists a point $p^* \in P$ at distance $d_X(q, p^*) \leq r$. In many applications, the metric space is high-dimensional, with the $d$-dimensional Euclidean space $\mathbb{R}^d$ being the most well-studied. See, e.g., survey [AIR18] for further background on high-dimensional ANN problem.

A common approach for solving the ANN problem is via Locality Sensitive Hashing (LSH) [HIM12], which can be viewed as a random, oblivious space partition. LSH mainly applies to the Euclidean ($\ell_2$) and Hamming ($\ell_1$) spaces. For example, for any fixed $\epsilon > 0$, one can solve ANN under $\ell_2$ with $c = \sqrt{1/\epsilon + o(1)}$ approximation, $O(n'd)$ query time and $O(n^{1+\epsilon}d)$ space (we call such query/space parameters as "efficient" henceforth).

Beyond $\ell_1$ and $\ell_2$ metrics, there has been much less progress and the best bounds are usually far from understood. A long-standing approach is via metric embeddings: a map from a metric $(X, d_X)$ into an "easier" space, say, $(\mathbb{R}^d, \ell_2)$, while approximately preserving all the distances:

**Definition 1.1** (bi-Lipschitz embedding into $\ell_2$). A map $f : X \to \ell_2$ is an embedding into $\ell_2$, with distortion $D \geq 1$, iff for any $x, y \in X$ we have that

$$
\|f(x) - f(y)\|_2 \leq d_X(x, y) \leq D \cdot \|f(x) - f(y)\|_2.
$$

Using such an embedding for a metric $(X, d_X)$, one can immediately extend the ANN under $\ell_2$ algorithms to ANN under $X$, with an extra factor $D$ approximation. Indeed, this approach



---

*Research supported in part by NSF grants (CCF-2008733, CCF-1617955, and CCF-1740833), and Simons Foundation (#491119).
yields best-known ANN algorithms for spaces such as the edit distance [OR07]. However, there are also known lower bounds for the best possible distortion, including in the case of edit distance [KN06, KR06]. A possible extension of the embedding approach is to embed a snowflake of the metric \((X, d_X)\) which means embedding \((X, d_X^\omega)\) for some \(\omega < 1\), though oftentimes there are lower bounds for this approach as well (albeit weaker), see, e.g., [AK10] for our example of edit distance.

A recent qualitative development is that of data-dependent hashing/space partitioning methods. For \(\ell_2\) and \(\ell_1\) spaces, [AINR14, AR15, ALRW17] developed data-dependent LSH that gave a better approximation: e.g., \(c = \sqrt{\frac{1+\epsilon}{2\epsilon}} + o(1)\) for \(\ell_2\); also see [Ind98] for an earlier example of data-dependent technique for \(\ell_\infty\).

For metrics beyond \(\ell_2, \ell_1\), [ANN+18] developed a new data-dependent approach, based on the following geometric property of the space \((X, d_X)\), termed cutting modulus:

**Definition 1.2 (Cutting modulus).** For fixed \(\epsilon > 0\), the cutting modulus of the metric \(X = (X, d_X)\), denoted \(\Xi(X, \epsilon)\) is the infimum \(D\) such that the following holds for any \(r > 0\) and any integer \(m > 1\). Fix a pointset \(V = \{x_1, \ldots, x_m\} \subset X\), and consider a positively-weighted graph \(G = (V, E, w)\) where \(w_{pq} > 0 \implies d_X(p, q) \leq r\) and \(\sum_{p \in V} \sum_{q \in V} w_{p,q} = 1\) (i.e., it’s a distribution over the ordered pairs \((p, q)\)). Then, one of the following must hold:

- Either there exists some point \(z \in X\) such that \(\sum_{p \in X} (z,p) \leq D\) and \(\sum_{q \in V} w_{p,q} \geq 1/2\) (i.e., there is a ball of radius \(Dr\) containing 1/2 mass of the pointset),
- Or there exists a non-trivial cut \(S \subset V\) with conductance \(\Phi(S) \leq \epsilon\).

The main result of [ANN+18] is that any metric \(X\) with cutting modulus \(D\) admits a \(D\)-ANN, albeit with a couple important caveats. First, without further properties, the query algorithm is efficient only in the cell-probe model, as it involves operating with an exponentially-sized graph \(G\) (think a net in \(X\)). Nonetheless, as [ANN+18] show, one can use a slightly stronger notion of cutting modulus to remove this caveat, for example obtaining efficient ANN for \(\ell_p\) and Schatten-\(p\) norms, with \(O(p/\epsilon)\) approximation, for any \(p \geq 2\). Second, the preprocessing time is exponential and seems unavoidable for this approach.

There are currently few bounds on the cutting modulus, and most of them follow from the theory of non-linear spectral gaps, whose systematic study was started in [MN14]; see also [Nao14, MN15]. Perhaps most striking is that the cutting modulus of any \(d\)-dimensional norm \((\mathbb{R}^d, \| \cdot \|_X)\) is \(O\left(\frac{\log d}{\epsilon}\right)\), following the bound on the non-linear spectral gap from [Nao17].

The work of [Nao20] introduced the notion of average embedding and proved a connection to non-linear spectral gap bounds. An average embedding relaxes the bi-Lipschitz embedding from above in that the contraction holds only on average with respect to a fixed given measure (and hence the embedding is "data dependent"):

**Definition 1.3 (Average embedding).** For a metric \(X = (X, d_X)\), a map \(f : X \to \ell_2\) is an \(q\)-average embedding, for a dataset \(P \subset X\), with distortion \(D\) if the following is satisfied:

\[
\|f(p) - f(q)\|_2 \leq D \cdot d_X(p, q) \quad \text{for all } p, q \in X
\]

\[
\sum_{p, q \in P} |f(p) - f(q)|_2^q \geq \sum_{p, q \in P} d_X(p, q)^q.
\]

When \(q = 2\), we call it simply average embedding.

In particular, [Nao20] showed that the average embedding distortion for 1/2-snowflake of a \(d\)-dimensional normed space is \(O(\sqrt{\log d})\). Similarly, [AINR14, Nao20] show that the average embedding distortion for \(\ell_p\) is \(O(p)\), for \(p > 2\). Both of these embeddings are non-constructive, proven by duality.
Our result. In this note, we show that one can use average embeddings directly to solve ANN.

Theorem 1.4. Suppose a metric $X$ has an average embedding $f$ with distortion $\sqrt{D}$, where time to construct $f$ is $T_P$ and time to compute $f$ on a point $q$ is $T_c$. Then there is a $c$-ANN with query time $O(n^{1+\epsilon} \cdot T_c)$, space $n^{O(1)}$, and preprocessing time $n^{O(1)} \cdot T_P$, and approximation $c = O(D^3/\epsilon)$.

Our approach avoids some of the aforementioned caveats of using cutting modulus, as long as we can construct an efficient average embedding for a given metric $X$. In particular, there is no more need for exponential time preprocessing (again assuming efficient average embedding map).

At high-level, our algorithm is similar to the ANN approach via standard bi-Lipschitz embedding: embed the space $X$ into $\ell_2$ and then use an efficient ANN for $\ell_2$, though the actual algorithm is a little bit more nuanced. The overall algorithm is “data-dependent” because the average embedding $f$ is data-dependent (note that $f$ from above may depend on the dataset $P$). In Appendix A we also propose an explicit, efficient average embedding for $\ell_p$, $p > 2$, albeit we leave it as a conjecture to prove its second property.

Independent work. In independent work, [KNT21] showed a similar result, obtaining a tighter bound of $O(D \log D)$-ANN. Their algorithm requires one more step: to construct a weak average embedding out of a generic average embedding. They also show how to compute the 1-average embedding for $\ell_p$ with distortion $O(p)$ (thus also resolving a related version of our conjecture), and hence obtain an $O(p)$-ANN with polynomial-time preprocessing (removing also the $O(\log p)$ from their generic reduction). Their algorithm thus improves over ANN from [BG19, ANN+18].

2 ANN Algorithm from Average Embedding

We now describe our main ANN algorithm that uses an average embedding. Our algorithm is solves a special case of the ANN problem, termed bounded ANN: the $\beta$-bounded ANN is the problem where we are guaranteed that all distances within the dataset $P$ verify $r \leq d_X(p, q) \leq \beta \cdot cr$. We note that it is enough to solve $\beta$-bounded ANN problem, for $\beta = 18$, due to the result of [BG19].

Our algorithm uses an LSH scheme for $\ell_2$, formally defined as follows.

Definition 2.1. (LSH) For a metric $X$, a family of hash functions $h : X \to U$ is $(p_1, p_2, r, cr)$-LSH if, for all $x, y \in X$:

- if $d_X(x, y) \leq r$, then $\Pr[h(x) = h(y)] \geq p_1$,
- if $d_X(x, y) > cr$, then $\Pr[h(x) = h(y)] < p_2$.

The exponent of the LSH scheme is defined as $\rho = \frac{\log 1/p_1}{\log 1/p_2}$.

The metric $X = \ell_2$ admits $(p_1, p_2, r, cr)$-LSH, for $\rho = 1/c$ with $p_1$ arbitrarily close to 1 [DIIM04]. We also assume that $r = 1$ henceforth for simplicity.

Fix some $\lambda, w \geq 1$ tbd. We construct the data structure for a $\beta$-bounded pointset $P$ recursively as a collection of randomized trees, constructed iid. A tree recursively partitions the dataset until the size of the dataset becomes constant (below think of $P$ is the dataset assigned to the current node). There are two type of nodes:
For a given query point $q$, we traverse the tree as follows:

- If the node is of the first type: if $d(x_0, q) < \sqrt{D}$, return the associated $p_0$, and otherwise, recurse into the single child.
- If the node is of the second type: branch into the child node $k = h(f(q))$.
- In a leaf node: check the distance to all stored nodes and return a valid answer if there exists one.

**Theorem 2.2.** For any $\epsilon > 0$, there is $\lambda, w$, and $c = O(D^3/\epsilon)$ such that the tree constructed as above has a probability at least $n^{-\epsilon}$ to find a $cr$-near neighbor, assuming there’s an $r$-near neighbor.

Hence, we can build a $\beta$-bounded ANN which constructs $n^\epsilon$ trees as above, and uses space $n^{1+O(\epsilon)}$ for approximation $c = O(D^3/\epsilon)$. The reduction from $[\text{BG19}]$ then yields an algorithm for the vanilla ANN.

**2.1 Analysis: Proof of Theorem 2.2**

First we note that nodes of the first type do not affect the correctness. So in the rest we fix a node of the tree above, and suppose we are in the second case: there is no $x_0 \in X$ such that $|P \cap B_X(x_0, \lambda D)| > |P|/8$. We further assume that $P$ contains a near neighbor point $p^*$, with $d(q, p^*) \leq 1$. We first prove the following lemma bounding the number of points that are "close" to $q$ after the embedding $f$.

**Lemma 2.3.** Consider a set $P$ satisfying:

- $|B_X(x, \lambda D) \cap P| \leq p|P|$ for all $x \in P$;
- For all points $x, y \in P$, $d_X(x, y) \leq \beta c$.

Consider a query $q$ at distance at most 1 from some point in $P$. Then, we have that

$$|B_{\ell_2}(f(q), wD) \cap f(P)| \leq 1 - \frac{(1 - p)\lambda^2}{4\beta^2 c^2},$$

as long as $(1 - p)\lambda^2 = \Omega(w^2)$, and $\beta c \geq \lambda$.

**Proof.** Let $\alpha = \frac{|B_{\ell_2}(f(q), wD)\cap f(P)|}{|P|}$. By the second property of the average embedding for $P$:

$$\sum_{x,y \in P} d(x, y)^2 \leq \sum_{x,y \in P} \|f(x) - f(y)\|^2$$

$$\Rightarrow \sum_{x,y \in P} d(x, y)^2 \leq (1 - \alpha^2)\beta^2 c^2 D^2 n^2 + 4\alpha^2 w^2 D^2 n^2$$
using $D$-Lipschitzness of $f$, and that for an $\alpha$ ratio of $x \in P$, $\|f(x) - f(q)\| \leq wD$.

$$n^2(1 - p)\lambda^2 D^2 \leq \beta^2 c^2 D^2 n^2 - \alpha^2 (\beta^2 c^2 D^2 - 4w^2 D^2) n^2$$

since $d_X(x, y) \geq \lambda D$ for a $1 - p$ ratio of $x, y \in P$.

$$\implies \alpha^2 \leq \frac{\beta^2 c^2 D^2 - (1 - p)\lambda^2 D^2}{\beta^2 c^2 D^2 - 4w^2 D^2}$$

$$\implies \alpha^2 \leq \frac{\beta^2 c^2 - (1 - p)\lambda^2}{\beta^2 c^2 - 4w^2}$$

$$\implies \alpha^2 \leq \left(1 - \frac{(1 - p)\lambda^2}{\beta^2 c^2}\right)\left(1 + O\left(\frac{w^2}{\beta^2 c^2}\right)\right)$$

$$\implies \alpha^2 \leq 1 - \frac{(1 - p)\lambda^2 - O(\beta^2)}{\beta^2 c^2}$$

$$\implies \alpha \leq 1 - \frac{(1 - p)\lambda^2}{4\beta^2 c^2}$$

for $(1 - p)\lambda^2 \geq \frac{1}{2}O(w^2)$. \hfill \Box

We use the lemma from above with $p = 1/8$, $\beta = 18$, and $c = \lambda D$. After the embedding $f$, it is guaranteed that $|f(P) \cap B(f(q), wD)| > |P| \cdot (1 - \Omega\left(\frac{1}{D^2}\right))$. Using an $(p_1, p_2, D, wD)$-LSH in the embedded space, the fraction of points that collides with the query is:

$$p'_2 \leq 1 - \Omega\left(\frac{1}{D^2}\right) + p_2 \left(1 - (1 - \Omega\left(\frac{1}{D^2}\right))\right) \leq 1 - (1 - p_2)\frac{\Omega(1)}{D^2}.$$ 

Therefore, embedding the query and the dataset and applying an $(p_1, p_2, D, wD)$-LSH is like applying an $(p_1, p'_2, 1, \lambda D)$-LSH in the original space. Therefore, we pick

$$p_1 = 1 - \epsilon(1 - p'_2) = 1 - \epsilon(1 - p_2)\frac{\Theta(1)}{D^2}.$$ 

Such a choice is possible as long as $w \geq \Theta(D^2/\epsilon)$. As, from above, we require $\lambda \geq \Omega(w)$, we have that $\lambda \geq \Omega(D^2/\epsilon)$. Finally, we get that the approximation ratio is $c = \lambda D = O(D^3/\epsilon)$. We obtain $\rho = \epsilon$ and hence $n^{-\epsilon}$ success probability.

References

[AINR14] Alexandr Andoni, Piotr Indyk, Huy L. Nguyen, and Ilya Razenshteyn. Beyond locality-sensitive hashing. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), 2014. Full version at \url{http://arxiv.org/abs/1306.1547}.

[AIR18] Alexandr Andoni, Piotr Indyk, and Ilya Razenshteyn. Approximate nearest neighbor search in high dimensions. 2018. This survey accompanied the ICM 2018 talk of Piotr Indyk. arXiv preprint arXiv:1806.09823.

[AK10] Alexandr Andoni and Robert Krauthgamer. The computational hardness of estimating edit distance. SIAM J. Comput. (SICOMP), 39(6):2398–2429, 2010. Previously in FOCS’07.

[ALRW17] Alexandr Andoni, Thijs Laarhoven, Ilya P. Razenshteyn, and Erik Waingarten. Optimal hashing-based time-space trade-offs for approximate near neighbors. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), 2017.
[ANN+18] Alexandr Andoni, Assaf Naor, Aleksandar Nikolov, Ilya P. Razenshteyn, and Erik Waingarten. Data-dependent hashing via nonlinear spectral gaps. In Proceedings of the Symposium on Theory of Computing (STOC), 2018.

[AR15] Alexandr Andoni and Ilya Razenshteyn. Optimal data-dependent hashing for approximate near neighbors. In Proceedings of the Symposium on Theory of Computing (STOC), 2015. Full version at http://arxiv.org/abs/1501.01062.

[BG19] Yair Bartal and Lee-Ad Gottlieb. Approximate nearest neighbor search for ℓ_p-spaces (2<p<∞). Theor. Comput. Sci., 757:27–35, 2019.

[DIIM04] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In Proceedings of the ACM Symposium on Computational Geometry (SoCG), 2004.

[HIM12] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. Theory of Computing, 1(8):321–350, 2012.

[Ind98] Piotr Indyk. Faster algorithms for string matching problems: matching the convolution bound. Proceedings of the Symposium on Foundations of Computer Science (FOCS), 1998.

[KN06] Subhash Khot and Assaf Naor. Nonembeddability theorems via Fourier analysis. Math. Ann., 334(4):821–852, 2006. Preliminary version appeared in FOCS’05.

[KNT21] Deepanshu Kush, Aleksandar Nikolov, and Haohua Tang. Near neighbor search via efficient average distortion embeddings. In Proceedings of the ACM Symposium on Computational Geometry (SoCG), 2021. To appear.

[KR06] Robert Krauthgamer and Yuval Rabani. Improved lower bounds for embeddings into L_1. In Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), pages 1010–1017, 2006.

[MN14] Manor Mendel and Assaf Naor. Nonlinear spectral calculus and super-expanders. Publications mathématiques de l’IHÉS, 119(1):1–95, 2014.

[MN+15] Manor Mendel, Assaf Naor, et al. Expanders with respect to hadamard spaces and random graphs. Duke Mathematical Journal, 164(4):1471–1548, 2015.

[Nao14] Assaf Naor. Comparison of metric spectral gaps. Analysis and Geometry in Metric Spaces, 2(1), 2014.

[Nao17] Assaf Naor. A spectral gap precludes low-dimensional embeddings. In Proceedings of the ACM Symposium on Computational Geometry (SoCG), 2017. Also at https://arxiv.org/abs/1611.08861.

[Nao20] Assaf Naor. An average John theorem. arXiv preprint 1905.01280, 2020. To appear in Geometry & Topology.

[OR07] Rafail Ostrovsky and Yuval Rabani. Low distortion embedding for edit distance. J. ACM, 54(5), 2007. Preliminary version appeared in STOC’05.
A Average Embedding for $\ell_p$

We consider the following average embedding from $\ell_p$ to $\ell_2$. Let

$$h(x)_i = \text{sign}(x_i)|x_i|^{p/2}$$

and

$$f(x) = \frac{h(x)}{\|h(x)\|_2}\|x\|_p = h(x)\frac{\|x\|_p}{\|x\|^{p/2}} = h\left(\frac{x}{\|x\|_p}\right)\|x\|_2$$

**Theorem A.1.** The embedding $f : \ell_p \to \ell_2$ is $p + 1$ Lipschitz.

**Proof.** We notice that $h$ is the Mazur map from $\ell_p$ to $\ell_2$. The embedding $f$ first normalizes its input, applies the Mazur map and the rescale it. We will therefore make use of the fact that the Mazur map is $\frac{7}{2}$-Lipschitz on the unit sphere.

$$\|f(x) - f(y)\|_2 = \left\| h\left(\frac{x}{\|x\|_p}\right) - h\left(\frac{y}{\|y\|_p}\right) \right\|_2$$

$$= \left\| \|x\|_p \left( \frac{h(x)}{\|h(x)\|_2} - \frac{h(y)}{\|h(y)\|_2} \right) + \frac{h(y)}{\|h(y)\|_2} \left( \|x\|_p - \|y\|_p \right) \right\|_2$$

$$\leq \|x\|_p \frac{h(x)}{\|h(x)\|_2} - \frac{h(y)}{\|h(y)\|_2} + \frac{h(y)}{\|h(y)\|_2} \left( \|x\|_p - \|y\|_p \right)$$

$$\leq \frac{p}{2} \|x\|_p \left( \frac{x}{\|x\|_p} - \frac{y}{\|y\|_p} \right) + \|x - y\|_p$$

$$= \frac{p}{2} \left( \|x\|_p - \|y\|_p \right) + \|x - y\|_p$$

$$= \frac{p}{2} \|x - y\|_p + \frac{p}{2} \|x\|_p - 1 \|y\|_p + \|x - y\|_p$$

$$\leq (p + 1) \|x - y\|_p$$

Where (a) uses the fact that $h(x)/\|h(x)\|_2 = h(x/\|x\|_p)$ and (b) uses the fact that $h$ is $p/2$-Lipschitz on the unit sphere. \hfill \Box

To complete the proof that the map $f$ is average embedding we need the following inequality, left as a conjecture.

**Conjecture A.2.** For any $x_1, \ldots, x_n$, we can find $z$ such that $x \to f(x - z)$ verifies, for some universal constant $C > 0$:

$$\sum_{i,j} \|f(x_i - z) - f(x_j - z)\|_2^2 \geq C \sum_{i,j} \|x_i - x_j\|_p^2.$$