Approximate Nearest Neighbor Search in ℓ_p

Huy L. Nguyễn
Princeton

Abstract

We present a new locality sensitive hashing (LSH) algorithm for c-approximate nearest neighbor search in ℓ_p with $1 < p < 2$. For a database of n points in ℓ_p, we achieve $O(dn^\rho)$ query time and $O(dn + n^{1+\rho})$ space, where $\rho \le O((\ln c)^2/c^p)$. This improves upon the previous best upper bound $\rho \le 1/c$ by Datar et al. (SOCG 2004), and is close to the lower bound $\rho \ge 1/c^p$ by O'Donnell, Wu and Zhou (ITCS 2011). The proof is a simple generalization of the LSH scheme for ℓ_2 by Andoni and Indyk (FOCS 2006).

1 Introduction

Approximate nearest neighbor search has been studied extensively in the last few decades. In this problem, a database of n points in $\mathbb{R}^d$ is preprocessed so that given a query point q, if the point closest to q in the database is at distance r from q, then the algorithm will return a point p in the database within distance cr from q. The parameter $c > 1$ is the approximation factor of the algorithm. At the moment, the best approach giving good guarantees both in time and space in high dimensions is locality sensitive hashing (LSH) [HPIM12]. The running time and space of LSH-based algorithms depend on a parameter ρ, which is determined by the metric space, the approximation factor c, and the hash functions: the query time is $dn^{\rho+o(1)}$ and the space is $nd + n^{1+\rho+o(1)}$. For the norm ℓ_p with $1 \le p \le 2$, it is known that there exists a distribution over hash functions such that $\rho \le 1/c$ [DIIM04]. For only the special case of ℓ_2, it is known that we can also get $\rho = 1/c^2 + o(1)$ [AI06]. In [OWZ11], it was shown that $\rho \ge c^{-p}$ for all $p \in [1, 2]$. In this paper we give a new LSH for $1 < p < 2$ achieving $\rho = O((\ln c)^2 c^{-p})$. Independently, there is an algorithm [IK13] close to matching the lower bound from [OWZ11] but we believe it is worthwhile to present the argument here as it is simple and might be applicable elsewhere.

2 Preliminaries

Fist we need the formal definition of LSH.

Definition 1 ([HPIM12]). *A family of hash functions h is (r, cr, p_1, p_2)-sensitive if*

- *If $\|x - y\|_p \le r$ then $\Pr[h(x) = h(y)] \ge p_1$.*
- *If $\|x - y\|_p \ge cr$ then $\Pr[h(x) = h(y)] \le p_2$.*

Define $\rho = \frac{\ln 1/p_1}{\ln 1/p_2}$.

Given such a hash family for every r, one immediately gets an algorithm for approximate nearest neighbor search.

Theorem 2 ([HPIM12]). *If for every r, there exists an (r, cr, p_1, p_2)-sensitive hash family with the parameter ρ uniformly bounded from above by ρ_0, evaluation time $dn^{o(1)}$, and $1/p_1 = n^{o(1)}$, then there is an algorithm for finding c-approximate nearest neighbor with query time $dn^{\rho_0+o(1)}$ and space $O(dn) + n^{1+\rho_0+o(1)}$.*

The rest of the paper focuses on analyzing ρ for a fixed r. Because we can always scale all distances, assume wlog that $r = 1$.

Let $B_p(x, r)$ denote the ℓ_p ball of radius r centered at x. Let V_t be the volume of $B_p(\vec{0}, w)$ in $\mathbb{R}^t$. Let L_t denote the lattice $\{\sum_{i=1}^{t} \Delta a_i w e_i \mid a_i \in \mathbb{Z}\}$, where e_i is the ith standard basis vector, $\Delta = 4$, and $w = O(c \ln c)$.

3 The Hash Function

The hash function works in a similar way to [AI06], with some modifications to the parameters. First it uses the p-stable distribution to reduce the dimension to $t = \Theta((c \ln c)^p)$. Then, it partitions the t-dimensional space using lattices of balls of radius $w = O(c \ln c)$. See Figure 1 for details.

Choosing a hash function $h \in \mathcal{H}$

1. For each $u \in \{1, 2, \ldots, U_t\}$, pick a random shift $s_u \in [0, \Delta w]^t$ to specify the shifted lattice $s_u + L_t$.
2. Pick a random matrix $A \in \mathbb{R}^{t \times d}$ whose entries are i.i.d. p-stable random variables with the scale parameter 1. Let $A' = T^{-1/p} A$ (T is the threshold defined in Lemma 5).

Applying the hash function h to a point $x \in \mathbb{R}^d$

1. Let $x' = A'x$.
2. Find the smallest $u \in \{1, 2, \ldots, U_t\}$ such that there exists a point $y \in L_t$ satisfying $x \in B_p(y + s_u, w)$. If u exists then the hash value of x is the pair (u, y). Otherwise, the hash value of x is the pair $(0, \vec{0})$.

Figure 1: The algorithm for computing the hash value of a given point $x \in \mathbb{R}^d$.

4 Analysis

First, we bound the number of lattices of balls needed to cover the entire space $\mathbb{R}^t$. This number determines the running time of the hash function as finding the closest ball in a lattice to a given point is simple: one just needs to find the closest lattice point in each coordinate separately. While this operation uses the floor function, we believe the usage is justified as the coordinates do not encode special information and it is also widely used in the LSH literature. The following lemma is a generalization of [And09, Lemma 3.2.2] with an analogous proof.

Lemma 3 ([And09])**.** *Consider the t-dimensional space $\mathbb{R}^t$ and let δ be a positive constant. Let B_u be the collection of ℓ_p balls centered at lattice points of $s_u + L_t$, where s_u is a uniformly random vector in $[0, \Delta w]^t$. If $U_t = \Delta^t t^{t/p+1} \ln(\Delta t/\delta)$ then the collections $B_1, \ldots, B_{U_t}$ cover all of $\mathbb{R}^t$ with probability at least $1 - \delta$.*

Proof. The proof is a standard covering argument. We present it here for completeness. By the regularity of the lattices, the whole space is covered iff the cube $[0, \Delta w]^t$ is covered. Divide the cube into subcubes of side length $w/t^{1/p}$. If some s_u lies in a subcube then the whole subcube is covered. The probability that some s_u lies in a particular cube is $1/N$, where N is the number of subcubes. We have $N = (\Delta t^{1/p})^t$. If $U_t \geq N \ln(N/\delta)$ then by the union bound, the probability that some subcube is not covered is bounded by

$$N(1 - 1/N)^{U_t} \leq \exp(\ln N - U_t/N) = \exp(-\ln 1/\delta) = \delta$$

□

Corollary 4. *For $\delta = \exp(-\Theta(t))$, the time to evaluate the hash function is $dc^{O((c \ln c)^p)}$.*

To analyze the first step of the hash function, we need a concentration bound for p-stable distribution. The proof is similar to that of a similar bound for $p = 1$ by [Ind06].

Lemma 5. *Let p be a constant in $(1, 2)$. Let $x \in \mathbb{R}^d$ and a random matrix $A \in \mathbb{R}^{t \times d}$ whose entries are i.i.d. p-stable random variables with the scale parameter 1. For $t \to \infty$, there exists a threshold $T = T(t, \epsilon)$ such that*

- $\Pr[\|Ax\|_p^p < T\|x\|_p^p] \leq \exp(-\Theta(t^{1-\epsilon p}(\epsilon \ln t)^2))$
- $\Pr[\|Ax\|_p^p > 2^{(4+p)/2}\epsilon^{-1}T\|x\|_p^p] \leq 1/2$

Proof. First, we need an approximation of the probability density function of the p-stable distribution. We use the following theorem from [Nel11, Theorem 42].

Theorem 6 ([Nel11])**.** *Define ϕ_p^+ and ϕ_p^- as follows:*

$$\phi_p^-(x) = \frac{a}{x^{p+1}} - \frac{b}{x^3}$$

and

$$\phi_p^+(x) = \frac{2^{(p+1)/2}a}{x^{p+1}} + \frac{b}{x^3}$$

for certain constants a, b. Then for any $x \geq 1$, $\phi_p^-(x) \leq \phi_p(x) \leq \phi_p^+(x)$.

Since $x \to Ax$ is a linear map, we can assume wlog that $\|x\|_p = 1$. Each coordinate of Ax is an i.i.d. p-stable random variable with the scale parameter 1. Let Y_i denote the absolute value of the ith coordinate of Ax. Define $Z_{i,M} := \min(Y_i, M)$. We first prove some properties of $Z_{i,M}$.

Lemma 7. *For $M \to \infty$, we have*

$$\mathbb{E}[Z_{i,M}^p] = \Theta(\ln M)$$
$$\mathbb{E}[Z_{i,M}^{2p}] = \Theta(M^p)$$

and in particular, $\mathbb{E}[Z_{i,M^{1/\epsilon}}^p] \leq (2^{(1+p)/2} + o(1))\mathbb{E}[Z_{i,M}^p]/\epsilon$.

Proof. Let p_M be the probability that a standard p-stable random variable exceeds M. We can bound p_M by

$$p_M \leq \int_M^\infty \phi_p^+(x)dx = \frac{2^{(p+1)/2}a}{pM^p} + \frac{b}{2M^2}$$

and

$$p_M \geq \int_M^\infty \phi_p^-(x)dx = \frac{a}{pM^p} - \frac{b}{2M^2}$$

First we bound $\mathbb{E}[Z_{i,M}^p]$.

$$\begin{aligned}\mathbb{E}[Z_{i,M}^p] &\leq \int_0^1 dx + \int_1^M x^p \phi_p^+(x)dx + p_M M^p \\ &\leq O(1) + 2^{(p+1)/2} a \ln M\end{aligned}$$

Similarly

$$\begin{aligned}\mathbb{E}[Z_{i,M}^p] &\geq \int_1^M x^p \phi_p^-(x)dx \\ &\geq a \ln M - O(1)\end{aligned}$$

Next we bound $\mathbb{E}[Z_{i,M}^{2p}]$.

$$\begin{aligned}\mathbb{E}[Z_{i,M}^{2p}] &\leq \int_0^1 dx + \int_1^M x^{2p} \phi_p^+(x)dx + p_M M^{2p} \\ &\leq (2^{(p+3)/2} a/p + o(1)) M^p\end{aligned}$$

Similarly

$$\begin{aligned}\mathbb{E}[Z_{i,M}^{2p}] &\geq \int_1^M x^{2p} \phi_p^-(x)dx \\ &\geq (a/p - o(1)) M^p\end{aligned}$$

□

Set $M = t^\epsilon$ and $T = t\mathbb{E}[Z_{i,M}^p]/2 = \Theta(\epsilon t \ln t)$. We have $\Pr[\|Ax\|_p^p < T] \leq \Pr[\sum_i Z_{i,M}^p < T]$. By an inequality by Maurer [Mau03].

$$\Pr[\sum_i Z_{i,M}^p < T] \leq \exp\left(-\frac{T^2}{2\sum_i \mathbb{E}[Z_{i,M}^{2p}]}\right) = \exp(-\Theta(t^{1-\epsilon p}(\epsilon \ln t)^2))$$

On the other hand,

$$\begin{aligned}\Pr[\sum_i Y_i^p > 2^{(4+p)/2}\epsilon^{-1}T] &\leq \Pr[\exists i : Y_i \geq M^{1/\epsilon}] + \Pr[\sum_i Y_i^p > 2^{(4+p)/2}\epsilon^{-1}T | \forall i : Y_i < M^{1/\epsilon}] \\ &\leq t(2^{(p+1)/2}a/(pM^{p/\epsilon}) + b/(2M^{2/\epsilon})) + \frac{\mathbb{E}[\sum_i Y_i^p | \forall i : Y_i < M^{1/\epsilon}]}{2^{(4+p)/2}\epsilon^{-1}T} \\ &\leq O(t^{1-p}) + \frac{\mathbb{E}[\sum_i Z_{i,M^{1/\epsilon}}^p]}{2^{(4+p)/2}\epsilon^{-1}T} \\ &\leq O(t^{1-p}) + \frac{(2^{(1+p)/2} + o(1))\epsilon^{-1}T}{2^{(4+p)/2}\epsilon^{-1}T} < 1/2\end{aligned}$$

□

To analyze the second step of the hash function, we use the uniform convexity and smoothness properties of ℓ_p, see e.g. [BCL94].

Fact 8. *For any $1 < p \le 2$,*

- *ℓ_p is p-uniformly smooth:*

$$\forall x, y \in \ell_p,\ \ \frac{\|x\|_p^p + \|y\|_p^p}{2} \le \left\|\frac{x+y}{2}\right\|_p^p + \left\|\frac{x-y}{2}\right\|_p^p \tag{1}$$

- *ℓ_p is 2-uniformly convex:*

$$\forall x, y \in \ell_p,\ \ \frac{\|x\|_p^2 + \|y\|_p^2}{2} \ge \left\|\frac{x+y}{2}\right\|_p^2 + (p-1)\left\|\frac{x-y}{2}\right\|_p^2 \tag{2}$$

Finally we are ready to prove the main technical lemma determining the parameter ρ. It can be viewed as a generalization of [And09, Lemma 3.2.3]. Before proceeding to the lemma, we want to note that conditioned on the fact that the whole space $\mathbb{R}^t$ is covered by the lattices (which happens with high probability by Lemma 3), for any two point $x, y \in \mathbb{R}^t$, the probability that they are contained in the same ball in the partition of $\mathbb{R}^t$ defined by the shifted lattices of h is exactly $\frac{\mathrm{Vol}(B_p(x,w)\cap B_p(y,w))}{\mathrm{Vol}(B_p(x,w)\cup B_p(y,w))}$. Removing the conditioning only changes the collision probabilities by at most $\delta = \exp(-\Theta(t))$, which is negligible. In a nutshell, the proof combines two observations. First, by Lemma 5, the mapping $x \to Ax$ does not distort distances by a large amount. Second, for points in $\mathbb{R}^t$, the volumes involved in collision probabilities can be approximated by volumes of balls of different radii. Because the ratio of volumes of balls of different radii can easily be computed from the ratio of the radii, we can approximate the collision probabilities.

Lemma 9. *Let p be a constant in $(1, 2)$. Let x, y be two points in $\mathbb{R}^d$. Let p_1 be the collision probability when $\|x - y\|_p \le 1$ and p_2 be the collision probability when $\|x - y\|_p \ge c$. Then, for $w = \Theta(c \ln c), t = \Theta(w^p)$, we have $\rho = \frac{\ln p_1}{\ln p_2} = O((\ln c)^2 c^{-p})$ as $c \to \infty$.*

Proof. Let $x' = A'x, y' = A'y$. We first analyze the volume of $A_1 = B_p(x', w) \cap B_p(y', w)$ when $\|x' - y'\|_p \le r$. We will show

$$\begin{aligned} B_p((x'+y')/2, w(1-(2+\gamma)r^p/(2w)^p)^{1/p}) \subset A_3 = A_1 \cup B_p(x', w(1-(2+2\gamma)r^p/(2w)^p)^{1/p}) \\ \cup B_p(y', w(1-(2+2\gamma)r^p/(2w)^p)^{1/p}) \end{aligned} \tag{3}$$

for arbitrary $\gamma > 0$. Setting γ close to 0 results in a better constant in the final bound of ρ but for ease of understanding, we can set $\gamma = 1$. Consider a point $z \in B_p((x'+y')/2, w(1-(2+\gamma)r^p/(2w)^p)^{1/p}) \setminus A_3$. Wlog, we assume $\|x' - z\|_p \ge \|y' - z\|_p$. By the assumptions, we have

$$\begin{aligned} \|(x'+y')/2 - z\|_p &\le w\left(1 - \frac{(2+\gamma)r^p}{(2w)^p}\right)^{1/p} \\ \|x' - z\|_p &> w \\ \|y' - z\|_p &> w\left(1 - \frac{(2+2\gamma)r^p}{(2w)^p}\right)^{1/p} \end{aligned}$$

Applying 1 to $x' - z$ and $y' - z$, we have:

$$\begin{aligned}\left\|\frac{x'-y'}{2}\right\|_p^p &\geq \frac{\|x'-z\|_p^p + \|y'-z\|_p^p}{2} - \left\|\frac{x'+y'}{2} - z\right\|_p^p \\ &> \frac{w^p + w^p(1-(2+2\gamma)r^p/(2w)^p)}{2} - w^p(1-(2+\gamma)r^p/(2w)^p) \\ &\geq (r/2)^p\end{aligned}$$

This contradicts the assumption that $\|x'-y'\|_p \leq r$. In other words, there is no such point z and $B_p((x+y)/2, w(1-(2+\gamma)r^p/(2w)^p)^{1/p}) \subset A_3$. Note that for any $\alpha \in \mathbb{R}$ and $u \in \mathbb{R}^t$, we have $\text{Vol}(B_p(u, \alpha w)) = \alpha^t V_t$. Applying this fact to (3), we have

$$\begin{aligned}\text{Vol}(A_1) &\geq V_t\left(\left(1 - \frac{(2+\gamma)r^p}{(2w)^p}\right)^{t/p} - 2\left(1 - \frac{(2+2\gamma)r^p}{(2w)^p}\right)^{t/p}\right) \\ &\geq V_t\left(1 - \frac{(2+\gamma)r^p}{(2w)^p}\right)^{t/p}\left(1 - 2\left(1 - \frac{\gamma r^p/2}{(2w)^p}\right)^{t/p}\right) \\ &\geq V_t\left(1 - \frac{(2+\gamma)r^p}{(2w)^p}\right)^{t/p}(1 - \exp(-\Omega(\gamma t r^p w^{-p})))\end{aligned}$$

By Lemma 5, when $\|x-y\|_p \leq 1$, with probability at least 1/2, we have $\|x'-y'\|_p = O(\epsilon^{-1})$. Therefore we get an upper bound for $\ln(1/p_1)$,

$$\begin{aligned}\ln(1/p_1) &\leq \ln\left(2 \cdot \frac{2V_t - \text{Vol}(A_1)}{\text{Vol}(A_1)}\right) \\ &\leq \ln 4 - \ln(\text{Vol}(A_1)) \\ &\leq \ln 4 + \frac{O(2+\gamma)\cdot t/(p(2\epsilon w)^p)}{1 - O(2+\gamma)\cdot 1/(2\epsilon w)^p} + \exp(-\Omega(\gamma t \epsilon^{-p} w^{-p})) \\ &\leq O(2+\gamma)\cdot t/(p(2\epsilon w)^p)\end{aligned}$$

The second to last inequality follows from the inequality $\ln(1-x) \geq -x/(1-x)\ \forall x \in [0,1)$.

Next, we analyze the volume of $A_2 = B_p(x', w) \cap B_p(y', w)$ when $\|x'-y'\|_p \geq c$. We will show $A_2 \subset B_p((x'+y')/2, w\sqrt{1-(p-1)(c/w)^2/4})$. Let z be an arbitrary point in A_2. Applying 2 to $x'-z$ and $y'-z$, we have:

$$\left\|\frac{x'+y'}{2} - z\right\|_p^2 \leq \frac{\|x'-z\|_p^2 + \|y'-z\|_p^2}{2} - (p-1)\left\|\frac{x'-y'}{2}\right\|_p^2 \leq w^2 - (p-1)(c/2)^2$$

Thus,

$$\text{Vol}(A_2) \leq V_t(1-(p-1)(c/w)^2/4)^{t/2}$$

By Lemma 5, when $\|x-y\|_p \geq c$, with probability at least $1 - P = 1 - \exp(-\Theta(t^{1-\epsilon p}(\epsilon \ln t)^2))$, we have $\|x'-y'\|_p \geq c$. Therefore, we get a lower bound for $\ln(1/p_2)$,

$$\begin{aligned}\ln(1/p_2) &\geq \ln\frac{2V_t - \text{Vol}(A_2)}{\text{Vol}(A_2) + (2V_t - \text{Vol}(A_2))P} \\ &\geq \ln\frac{1 - \text{Vol}(A_2)/(2V_t)}{\max(\text{Vol}(A_2)/V_t, 2P)} \\ &\geq \ln(1/2) + \min((p-1)tc^2/(8w^2), -\ln 2P) \\ &\geq (1-o(1))(p-1)tc^2/(8w^2)\end{aligned}$$

when $(p-1)tc^2/(8w^2) < -\ln 2P = \Theta(t^{1-\epsilon p}(\epsilon \ln t)^2) - \ln 2$. In other words, we need $w/c = \Omega(t^{\epsilon p/2}/(\epsilon \ln t))$. Combining the bounds for p_1 and p_2, we have

$$\rho = \frac{\ln(1/p_1)}{\ln(1/p_2)} \leq \frac{O(2+\gamma) \cdot (w/c)^{2-p}}{\epsilon^p 2^p p(p-1)c^p}$$

We get the stated bound by choosing $w = \Theta(c \ln c), t = \Theta(w^p), \epsilon = \Theta(\ln \ln t / \ln t)$. □

Remark 10. *It is possible to slightly tighten the bound by setting* $w = \Theta(c), t = \Theta(w^p), \epsilon = \Theta(1/\ln t)$*. The constant* $2^{(4+p)/2}$ *in Lemma 5 becomes a larger constant depending on the constants in Theorem 6, but the rest of the proof remains the same. This setting gives* $\rho = O((\ln c/c)^p)$*, where the* O *hides a constant depending on the constants in Theorem 6.*

5 Discussion

The second half of the argument uses only the uniform smoothness and convexity properties of the norm while the first half is tailored to ℓ_p. This leads to the question of whether one can generalize the argument here to get an algorithm for approximate nearest neighbor search for a more general class of norms.

6 Acknowledgments

We thank Jelani Nelson and Ilya Razenshteyn for helpful comments.

References

[AI06] Alexandr Andoni and Piotr Indyk. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*, pages 459–468, 2006.

[And09] Alexandr Andoni. *Nearest Neighbor Search: the Old, the New, and the Impossible.* PhD thesis, MIT, 2009.

[BCL94] Keith Ball, Eric A. Carlen, and Elliott H. Lieb. Sharp uniform convexity and smoothness inequalities for trace norms. *Inventiones mathematicae*, 115(1):463–482, 1994.

[DIIM04] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S. Mirrokni. Locality-sensitive hashing scheme based on p-stable distributions. In *Symposium on Computational Geometry*, pages 253–262, 2004.

[HPIM12] Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani. Approximate nearest neighbor: Towards removing the curse of dimensionality. *Theory Of Computing*, 8:321–350, 2012.

[IK13] Piotr Indyk and Michael Kapralov. Personal communication. May 2013.

[Ind06] Piotr Indyk. Stable distributions, pseudorandom generators, embeddings, and data stream computation. *J. ACM*, 53(3):307–323, 2006.

[Mau03] Andreas Maurer. A bound on the deviation probability for sums of non-negative random variables. *J. Inequalities in Pure and Applied Mathematics*, 4(1):15, 2003.

[Nel11] Jelani Nelson. *Sketching and Streaming High-Dimensional Vectors.* PhD thesis, MIT, 2011.

[OWZ11] Ryan O'Donnell, Yi Wu, and Yuan Zhou. Optimal lower bounds for locality sensitive hashing (except when q is tiny). In *ICS*, pages 275–283, 2011.