Technical Report for "Neighbor-Sensitive Hashing"

In this manuscript, we present the proofs to the theorems and lemmas in our main paper "Neighbor-Sensitive Hashing". In the future, this manuscript may be augmented to include more details on our experiment setting and implementation.

Theorem 1. Let q be a query, and v_1 and v_2 two data items. Also, let $\mathbf{h}$ be an LSH-like hash function consisting of b *independent* bit functions $h_1, \ldots, h_b$. Then, the following relationship holds for all v_1 and v_2 satisfying $0.146 < \|q - v_1\| < \|q - v_2\|$: A larger value of $E\|\mathbf{h}(q) - \mathbf{h}(v_2)\|_H - E\|\mathbf{h}(q) - \mathbf{h}(v_1)\|_H$ implies a larger value of $\Pr(\|\mathbf{h}(q) - \mathbf{h}(v_1)\|_H < \|\mathbf{h}(q) - \mathbf{h}(v_2)\|_H)$, i.e., the probability of successful ordering of v_1 and v_2 based on their hashcodes.

Proof. Let us denote the probability that an individual bit function h_i where $i = 1, 2, \ldots, b$ assigns q and v_1 into *different* hash bits is p_1, and the probability that assigns q and v_2 into different hash bits is p_2. A reasonable bit function satisfies $p_1 < p_2 \leq 0.5$. Note that $E\|\mathbf{h}(q) - \mathbf{h}(v_2)\|_H - E\|\mathbf{h}(q) - \mathbf{h}(v_1)\|_H = b \cdot (p_2 - p_1)$. There are two cases in which the above expression increases: first, p_1 becomes smaller, and second, p_2 becomes larger. Let us start with the first case.

To compute the probability distribution of the difference of hamming distances, $\Pr(\|\mathbf{h}(q) - \mathbf{h}(v_2)\|_H - \|\mathbf{h}(q) - \mathbf{h}(v_1)\|_H)$, we take a look at the distribution of $\|\mathbf{h}(q) - \mathbf{h}(v_1)\|_H$. Since b number of bit functions that compose the $\mathbf{h}$ are independent of one another, $\|\mathbf{h}(q) - \mathbf{h}(v_1)\|_H$ follows the binomial distribution with mean bp_1 and variance $bp_1(1 - p_1)$. Similarly, $\|\mathbf{h}(q) - \mathbf{h}(v_2)\|_H$ follows the binomial distribution with mean bp_2 and variance $bp_2(1 - p_2)$. Exploiting the fact that binomial distributions can be closely approximated by the normal distributions with the same mean and the variance, and that the difference between two normal distributions follows another normal distribution, we can state the following:

$$\begin{aligned} &\Pr(\|\mathbf{h}(q) - \mathbf{h}(v_1)\|_H < \|\mathbf{h}(q) - \mathbf{h}(v_2)\|_H) \\ &\approx \int_0^\infty \mathcal{N}\left(bp_2 - bp_1,\, bp_2(1 - p_2) + bp_1(1 - p_1)\right)\, dx \end{aligned}$$

where $\mathcal{N}$ denotes the probability distribution function of a normal distribution. Note that the above quantity is a function of two values: mean and standard

deviation. Due to the shape of a normal distribution, higher mean and smaller standard deviation results in a higher chance of successful ordering of v_2 and v_1 based on their hashcodes. If p_1 decreases, the mean of the above normal distribution increases, and the standard deviation of the distribution decreases. Therefore, the quantity of our interest increases. **(End of the first case)**

Next, let us discuss the case where p_2 increases. This case asks for more careful analysis because the standard deviation of the normal distribution increases. Recall that the area computed by the integration is a function of the mean and the standard deviation of the normal distribution; thus, *if the mean increases faster than the standard deviation, the quantity of our interest still increases.* Let us compute the condition that the mean increases faster. Since we are dealing with the case in which p_2 increases,

$$\frac{\partial(bp_2 - bp_1)}{\partial p_2} = b \tag{1}$$

must be larger than

$$\frac{\partial\sqrt{bp_2(1-p_2) + bp_1(1-p_1)}}{\partial p_2} = b \cdot \frac{1-2p_2}{2\sqrt{p_2 - p_2^2}} \tag{2}$$

The condition for this is $p_2 > 0.146$. **(End of the second case)** □

Theorem 2. Let $\mathbf{h}$ be an LSH-like hash function and f be a q-(η_{min}, η_{max})-sensitive transformation. Then, for all constants t_i and t_j, where $\eta_{min} \leq t_i \leq t_j \leq \eta_{max}$, we have the following:

$$\begin{aligned} &E(\|\mathbf{h}(f(q)) - \mathbf{h}(f(v_j))\|_H - \|\mathbf{h}(f(q)) - \mathbf{h}(f(v_i))\|_H) \\ &> E(\|\mathbf{h}(q) - \mathbf{h}(v_j)\|_H - \|\mathbf{h}(q) - \mathbf{h}(v_i)\|_H) \end{aligned}$$

where the expectations are computed over data items v_i and v_j chosen uniformly at random among data items whose distances to q are t_i and t_j, respectively.

Proof. Since $\mathbf{h}$ is LSH-like,

$$\begin{aligned} E(\|\mathbf{h}(f(q)) - \mathbf{h}(f(v_j))\|_H) &= E_v(E_h(\|\mathbf{h}(f(q))) - \mathbf{h}(f(v_j)\|_H)) \\ &= E_v(c \cdot b \cdot \|f(q) - f(v_j)\|) \\ &= c \cdot b \cdot E_v(\|f(q) - f(v_j)\|) \end{aligned}$$

where E_h is an expectation over $\mathbf{h}$ and E_v is an expectation over v_j. Similarly,

$$\begin{aligned} E(\|\mathbf{h}(f(q)) - \mathbf{h}(f(v_i))\|_H) &= c \cdot b \cdot E_v\|f(q) - f(v_i)\| \\ E(\|\mathbf{h}(q) - \mathbf{h}(v_j)\|_H) &= c \cdot b \cdot E_v\|q - v_j\| \\ E(\|\mathbf{h}(q) - \mathbf{h}(v_i)\|_H) &= c \cdot b \cdot E_v\|q - v_i\| \end{aligned}$$

where E_v is either an expectation over v_i or an expectation over v_j depending on the random variable involved. Due to the third property of NST,

$$E_v\|f(q) - f(v_j)\| - E_v\|f(q) - f(v_i)\| > \|q - v_j\| - \|q - v_i\|$$

Therefore, the relationship we want to show also holds. □

Lemma 1. A pivoted transformation f_p satisfies the second property of NST, i.e., monotonicity.

Proof. Let q be a query, p be a pivot, and t be an arbitrary positive constant. Also, v is a data item chosen uniformly at random among items whose distance to q is t. In addition, let α denote an angle between $\overrightarrow{qv}$ and $\overrightarrow{qp}$. Since v is chosen uniformly at random, α is a random variable whose probability distribution function is a uniform between 0 and 2π.

To show the monotonicity, it is enough to show the following:

$$\frac{E(|f(q) - f(v)|)}{\partial t} \geq 0 \implies E\left(\frac{\partial|f(q) - f(v)|}{\partial t}\right) \geq 0$$

The interchange of E and the partial derivative is valid since the random variable inside the expectation (v) only depends on α.

To simplify the notations, let $t_q = \|p - q\|$ and $t_v = \|p - v\|$. Then, $t_v^2 = t^2 + t_q^2 - 2t_q t\cos\alpha$ from the law of cosines. Note that t_q is constant while t_v varies depending on t and α. Therefore,

$$2t_v\frac{\partial t_v}{\partial t} = 2t - 2t_q\cos\alpha, \quad \frac{\partial t_v}{\partial t} = \frac{t - t_q\cos\alpha}{t_v}$$

We divide this proof into two cases: (1) $t \geq 2t_q$ and (2) $t < 2t_q$. For the first case when $t \geq t_q$, we get $t_v \geq t_q$ using the triangular inequality. As a result, $|f(q) - f(v)| = \exp(-t_q^2/\eta^2) - \exp(-t_v^2/\eta^2)$. Therefore,

$$\frac{\partial|f(q) - f(v)|}{\partial t} = \frac{2t_v}{\eta^2}\frac{t - t_q\cos\alpha}{t_v}\exp\left(-\frac{t_v^2}{\eta^2}\right)$$

From $t_v \geq t_q$, we know $2t - 2t_q\cos\alpha \geq t \geq 0$, so

$$E\left(\frac{\partial|f(q) - f(v)|}{\partial t}\right) \geq 0$$

For the second case, when $t < 2t_q$, the sign of $f(q) - f(v)$ depends on the sign of $t_q - t_v$. In other words,

$$\begin{aligned}\frac{\partial|f(q) - f(v)|}{\partial t} &= \frac{f(q) - f(v)}{|f(q) - f(v)|}\frac{2t_v}{\eta^2}\frac{t - t_q\cos\alpha}{t_v}\exp\left(-\frac{t_v^2}{\eta^2}\right) \\ &= \frac{f(q) - f(v)}{|f(q) - f(v)|}\frac{2(t - t_q\cos\alpha)}{\eta^2}\exp\left(-\frac{t^2 + t_q^2 - 2t_q t\cos\alpha}{\eta^2}\right)\end{aligned}$$

We further simplify the above expression by substituting $l_1\eta$ and $l_2\eta$ for t and t_q, respectively, and we treat the expression as a function of l_1 and l_2. Then, we obtain

$$g(l_1, l_2) = \frac{1}{2\pi} \int_0^{2\pi} \frac{f(q) - f(v)}{|f(q) - f(v)|} \frac{2(l_1 - l_2 \cos\alpha)}{\eta} \exp(-l_1^2 + l_2^2 - 2l_1 l_2 \cos\alpha)\, d\alpha$$

Unfortunately, showing $g(l_1, l_2) > 0$ for all l_1 and l_2 such that $0 < l_2 < 0.5$ and $0 < l_1 < 2l_2 < 1$ analytically is difficult because a closed-form solution of the above integration does not exist. However, we can obtain high confidence from numerical analysis due to the following reasons:

1. $g(l_1, l_2)$ is a continuous function of l_1 and l_2.
2. The function of the form $x\exp(x)$ does not fluctuate fast and is can be approximated well by piece-wise linear functions.

Therefore, if we pick four closely located points in the space of l_1 and l_2 and the values of $g(l_1, l_2)$ at all those four points are positive, we obtain high confidence that the function values of the area enclosed by those four points will be positive as well.

For this, we generated 1,000,000 pairs of (l_1, l_2) where l_1 and l_2 are evenly spaced between 0 and 1 and between 0 and 0.5, respectively. Next, we computed the value of the function $g(l_1, l_2)$ for all 1,000,000 pairs. In the end, we confirmed that every pair we generated are positive. This result implies that $g(l_1, l_2) > 0$ for all l_1 and l_2 such that $0 < l_2 < 0.5$ and $0 < l_1 < 2l_2 < 1$. □

Lemma 2. A pivoted transformation f_p with $\|p-q\| < \eta/2$ and $\eta < 0.2$ satisfies the third property of NST, i.e., Larger Gap, for $(\eta_{min}, \eta_{max}) = (0.13\eta, 1.6\eta)$. That is, f_p is a q-$(0.13\eta, 1.6\eta)$-sensitive transformation.[1]

Proof. Let q be a query, p be a pivot, and t be an arbitrary positive constant. Also, v is a data item chosen uniformly at random among items whose distance to q is t. In addition, let α denote an angle between $\overrightarrow{qv}$ and $\overrightarrow{qp}$. Since v is chosen uniformly at random, α is a random variable whose probability distribution function is a uniform between 0 and 2π.

To show the monotonicity, it is enough to show the following:

$$\frac{E(|f(q) - f(v)|)}{\partial t} \geq 1 \implies E\left(\frac{\partial |f(q) - f(v)|}{\partial t}\right) \geq 1$$

for $t \in (0.13\eta, 1.6\eta)$. The interchange of E and the partial derivative is valid since the random variable inside the expectation (v) only depends on α.

[1] When working with *non-normalized distances*, η should be smaller than $0.2 \cdot t_{max}$, where t_{max} is the maximum distance between data items.

To simplify the notations, let $t_q = \|p - q\|$ and $t_v = \|p - v\|$. Then, $t_v^2 = t^2 + t_q^2 - 2t_q t\cos\alpha$ from the law of cosines. Note that t_q is constant while t_v varies depending on t and α. Therefore,

$$2t_v\frac{\partial t_v}{\partial t} = 2t - 2t_q\cos\alpha, \quad \frac{\partial t_v}{\partial t} = \frac{t - t_q\cos\alpha}{t_v}$$

and

$$\begin{aligned}\frac{\partial|f(q)-f(v)|}{\partial t} &= \frac{f(q)-f(v)}{|f(q)-f(v)|}\frac{2t_v}{\eta^2}\frac{t-t_q\cos\alpha}{t_v}\exp\left(-\frac{t_v^2}{\eta^2}\right)\\ &= \frac{f(q)-f(v)}{|f(q)-f(v)|}\frac{2(t-t_q\cos\alpha)}{\eta^2}\exp\left(-\frac{t^2+t_q^2-2t_q t\cos\alpha}{\eta^2}\right)\end{aligned}$$

We substitute $l_1\eta$ and $l_2\eta$ for t and t_q, respectively, and consider the above expression as a function of l_1 and l_2. Then, we should show that

$$g(l_1,l_2) = \frac{1}{2\pi}\int_0^{2\pi}\frac{f(q)-f(v)}{|f(q)-f(v)|}\frac{2(l_1-l_2\cos\alpha)}{\eta}\exp(-l_1^2+l_2^2-2l_1l_2\cos\alpha)\,d\alpha \geq 1$$

Since $\eta < 0.2$, it is enough to show that

$$g'(l_1,l_2) = \frac{1}{2\pi}\int_0^{2\pi}\frac{f(q)-f(v)}{|f(q)-f(v)|}2(l_1-l_2\cos\alpha)\exp(-l_1^2+l_2^2-2l_1l_2\cos\alpha)\,d\alpha \geq 0.2$$

for all l_1 and l_2 such that $0 < l_2 < 0.5$ and $0.13 < l_1 < 1.6$. Similar to Lemma 1, analytically computing the above integration is not easy because a closed-form solution of the above integration does not exist. Thus, to numerically verify this lemma, we generated 1,000,000 pairs of (l_1, l_2) where l_1 and l_2 are evenly spaced between 0.13 and 1.6 and between 0 and 0.5, respectively. Next, we computed the value of $g'(l_1, l_2)$ for all pairs. In the end, we confirmed that every pair we generated is not smaller than 0.204. □