COMS E6998-9: Algorithmic Techniques for Massive Data — Oct 8, 2015

# Lecture 10 – Sketching and Nearest Neighbour Search

Instructor: *Alex Andoni* — Scribes: *Chaoxu Zhou*

## 1 Sketching

In this section, we will look at space complexity in number of bits instead of in number of words. First we define sketch $S(\cdot)$ as a map from $\mathbf{R}^d$ to the space of *"short-bit-strings"* that given $S(x)$ and $S(y)$, we are able to estimate some function of $x$ and $y$ with a constant success probability. The goal in this section is to solve the decision version problem using sketching, that is, given $r$ and $\epsilon$, distinguishing $\|x-y\| \leq r$ and $\|x-y\| > (1+\epsilon)r$ based on $S(x)$ and $S(y)$ with constant probability. We will show that a sketch of size $O(1/\epsilon^2)$ is sufficient to achieve this goal, .

**Lemma 1.** *For every $\epsilon > 0, 1 < r < d$, a sketch $S : \{0,1\}^d \to \{0,1\}^k$ can be constructed such that*

**(1)** *for any $x, y \in \{0,1\}^d$, a decision can be generated based on $S(x)$ and $S(y)$, which can distinguish*

$$\begin{aligned} &\|x-y\|_1 \leq r \qquad \textit{or} \\ &\|x-y\|_1 > (1+\epsilon)r \end{aligned}$$

*with constant success probability.*

**(2)** $k = O(1/\epsilon^2)$.

*Proof.*
**Part 1** $r = d/C$, where $C > 1$ is a known constant.
Let $J = \{j_1, \ldots, j_k\}$ be a random sample chosen from $\{1, 2, \ldots, d\}$ uniformly without replacement. For every $x \in \{0,1\}^d$, define $S(x) = (x_{j_1}, \ldots, x_{j_k})$ where $x_i$ is the $i-$th entry of vector $x$. For every $x, y \in \{0,1\}^d$, the following decision will be made based on $S(x)$ and $S(y)$:

$$\begin{aligned} &\text{if } \ \|S(x)-S(y)\|_1 < \frac{r}{d}k(1+\sqrt{C}\epsilon), \quad \text{report } \|x-y\|_1 < r \\ &\text{if } \ \|S(x)-S(y)\|_1 \geq \frac{r}{d}k(1-\sqrt{C}\epsilon), \quad \text{report } \|x-y\|_1 > (1+\epsilon)r. \end{aligned}$$

Denote $\Delta = \|x-y\|_1$ . and $\hat{\Delta} = \|S(x)-S(y)\|_1$. It's easy to see that $\|x-y\|_1 = \sum_{j=1}^d \mathbb{I}\{x_j \neq y_j\}$. As a consequence, if we pick $i$ from $[d]$ uniformly randomly, then

$$\mathbb{P}(x_i = y_i) = \frac{\sum_{j=1}^d \mathbb{I}(x_j = y_j)}{d} = \frac{d - \sum_{j=1}^d \mathbb{I}(x_j \neq y_j)}{d} = \frac{d-\Delta}{d}.$$

In order establish the desired result, we need calculate the expectation and variance of $\hat{\Delta}$.

$$\mathbb{E}(\hat{\Delta}) = \mathbb{E}(\|S(x) - S(y)\|_1) = \mathbb{E}(\sum_{l=1}^{k} \mathbb{I}\{x_{j_l} \neq y_{j_l}\}) = \sum_{l=1}^{k} \mathbb{P}(x_{j_l} \neq y_{j_l}) = \frac{k}{d}\Delta.$$

And

$$\begin{aligned}
\mathrm{var}(\|\hat{\Delta}\|_1) &= \mathrm{var}(\sum_{l=1}^{k} \mathbb{I}\{x_{j_l} \neq y_{j_l}\}) \\
&= \sum_{l=1}^{k} \mathrm{var}(\mathbb{I}\{x_{j_l} \neq y_{j_l}\}) && \text{(by independence of } j_1, \dots, j_l) \\
&= \sum_{l=1}^{k} (1 - \frac{\Delta}{d})\frac{\Delta}{d} = \frac{k}{d}\Delta(1 - \frac{\Delta}{d}) \leq \frac{k}{d}\Delta.
\end{aligned}$$

Then the probability of making a wrong decision can be calculated as

$$\begin{aligned}
\mathbb{P}(\text{A wrong decision is made}) &= \mathbb{P}\{\Delta \leq r, \text{we report } \Delta \geq (1+\epsilon)r\} + \mathbb{P}\{\Delta \geq (1+\epsilon)r, \text{we report } \Delta \leq r\} \\
&= \mathbb{P}\{\Delta \leq r, \hat{\Delta} > \frac{r}{d}k(1 + \sqrt{C}\epsilon)\} + \mathbb{P}\{\Delta \geq (1+\epsilon)r, \hat{\Delta} < \frac{r}{d}k(1 - \sqrt{C}\epsilon)\} \\
&\leq \mathbb{P}\{\hat{\Delta} > \frac{\Delta}{d}k + \frac{rk\sqrt{C}}{d}\epsilon\} + \mathbb{P}\{\hat{\Delta} < \frac{\Delta}{d}k - \frac{rk\sqrt{C}}{d}\epsilon\} \\
&= \mathbb{P}(|\hat{\Delta} - \frac{k}{d}\Delta| > \frac{kr\sqrt{C}}{d}\epsilon) = \mathbb{P}(|\hat{\Delta} - \mathbb{E}\hat{\Delta}| > \frac{kr\sqrt{C}}{d}\epsilon) \\
&\leq \frac{\mathrm{var}(\hat{\Delta})}{k^2 r^2 C \epsilon^2 / d^2} && \text{(by Chebyshev's inequality)} \\
&\leq \frac{d^2}{k^2 r^2 C \epsilon^2} \frac{k\Delta}{d} \\
&\leq \frac{1}{k\epsilon^2}. && \text{(Since } \Delta \leq r \text{ and } r = d/C)
\end{aligned}$$

Thus set $k = O(\epsilon^{-2})$ is sufficient to achieve a constant success probability.

**Part 2** $2 < r \ll d$ which correspond to the case $C \to \infty$ in **Part 1**
Let $u_1, \dots, u_k \in \{0,1\}^d$ be random vectors that $\mathbb{P}(u_{ij} = 1) = 1/r$, where $u_{ij}$ denotes the $j$-th entry of $u_{ij}$. We can choose $u_1, \dots, u_k$ such that $\{u_{ij}\}$'s are independent for all $i = 1, \dots, d$, $j = 1, \dots, k$. We define the sketch $S : \{0,1\}^d \to \{0,1\}^k$ by

$$S(x) = (x^\top u_1, \dots, x^\top u_k) \mod 2$$

for every $x \in \{0,1\}^d$.

For every $x, y \in \{0,1\}^d$, a decision rule based on $S(x)$ and $S(y)$ is constructed as

$$\begin{aligned}
&\text{if } \ \|S(x)-S(y)\|_1 < (1-\epsilon)\frac{k}{2}\{1-(1-\frac{2}{r})^r\} \quad \text{report } \|x-y\|_1 < r \\
&\text{if } \ \|S(x)-S(y)\|_1 > (1+\epsilon)\frac{k}{2}\{1-(1-\frac{2}{r})^r\} \quad \text{report } \|x-y\|_1 \geq r.
\end{aligned}$$

To verify that this decision rule has constant success probability when $k = O(\epsilon^{-2})$, we first need to calculate the expectation and variance of $\hat{\Delta} = \|S(x)-S(y)\|_1$. Note that

$$\begin{aligned}
\mathbb{E}\{\hat{\Delta}\} &= \mathbb{E}\sum_{i=1}^{k} \mathbb{I}(x^\top u_i - y^\top u_i \equiv 1 \mod 2) \\
&= \sum_{i=1}^{k} \mathbb{P}(x^\top u_i - y^\top u_i \equiv 1 \mod 2) \\
&= k\mathbb{P}(x^\top u_1 - y^\top u_1 \equiv 1 \mod 2).
\end{aligned}$$

The task is converted to calculate $\mathbb{P}(x^\top u_1 - y^\top u_1 \equiv 1 \mod 2)$. Denote $D = \{j \,|x_j \neq y_j\}$, where $x_j, y_j$ are the $j$-th entry of $x$ and $y$. Write $|D|$ to be the cardinality of $D$ and note that $|D| = \|x-y\|_1 = \Delta$, then we have

$$\begin{aligned}
\mathbb{P}(x^\top u_1 - y^\top u_1 \equiv 1 \mod 2) &= \mathbb{P}\{\sum_{j\in D}(x_j - y_j)u_j \equiv 1 \mod 2\} \\
&= \mathbb{P}(\sum_{j\in D} u_j \equiv 1 \mod 2).
\end{aligned}$$

The above probability equals to the probability of the value of $X$ is even, where $X \sim \text{binomial}(\Delta, 1/r)$. Note that when $\Delta$ is odd

$$\begin{aligned}
1 = (1-\frac{1}{r}+\frac{1}{r})^\Delta &= \sum_{i=0}^{\Delta}\binom{\Delta}{i}(\frac{1}{r})^i(1-\frac{1}{r})^{\Delta-i} \\
&= \sum_{i=0}^{\lfloor\Delta/2\rfloor}\binom{\Delta}{2i}(\frac{1}{r})^{2i}(1-\frac{1}{r})^{\Delta-2i} + \sum_{i=0}^{\lfloor\Delta/2\rfloor}\binom{\Delta}{2i+1}(\frac{1}{r})^{2i+1}(1-\frac{1}{r})^{\Delta-2i-1} \\
&= \mathbb{P}(X \text{ is even}) + \mathbb{P}(X \text{ is odd})
\end{aligned}$$

and

$$\begin{aligned}
(1-\frac{2}{r})^\Delta = (1-\frac{1}{r}-\frac{1}{r})^\Delta &= \sum_{i=0}^{\Delta}\binom{\Delta}{i}(-\frac{1}{r})^i(1-\frac{1}{r})^{\Delta-i} \\
&= \sum_{i=0}^{\lfloor\Delta/2\rfloor}\binom{\Delta}{2i}(\frac{1}{r})^{2i}(1-\frac{1}{r})^{\Delta-2i} - \sum_{i=0}^{\lfloor\Delta/2\rfloor}\binom{\Delta}{2i+1}(\frac{1}{r})^{2i+1}(1-\frac{1}{r})^{\Delta-2i-1} \\
&= \mathbb{P}(X \text{ is even}) - \mathbb{P}(X \text{ is odd}).
\end{aligned}$$

Therefore

$$\begin{aligned}
\mathbb{P}(X \text{ is even}) &= \frac{1}{2} + \frac{1}{2}(1 - \frac{2}{r})^{\Delta} \\
\mathbb{P}(X \text{ is odd}) &= \frac{1}{2} - \frac{1}{2}(1 - \frac{2}{r})^{\Delta},
\end{aligned}$$

when $\Delta$ is odd.
We can use exactly the same method to derive exactly the same result for $\Delta$ is even. Thus we have

$$\begin{aligned}
\mathbb{E}(\hat{\Delta}) &= k\mathbb{P}(x^\top u_1 - y^\top u_1 \equiv 1 \mod 2) \\
&= k\mathbb{P}(\sum_{j\in D} u_j \equiv 1 \mod 2) \\
&= \frac{k}{2}\{1 - (1 - \frac{2}{r})^{\Delta}\}.
\end{aligned}$$

To calculate the variance of $\hat{\Delta}$, we know that $\hat{\Delta}$ is a summation of $k$ independent Bernoulli random variables with parameter $\{1 - (1 - 2/r)^{\Delta}\}/2$. Thus

$$\begin{aligned}
\text{var}(\hat{\Delta}) &= k\frac{1}{2}\{1 - (1 - \frac{2}{r})^{\Delta}\}\frac{1}{2}\{1 + (1 - \frac{2}{r})^{\Delta}\} \\
&\leq k\frac{1}{2}\{1 - (1 - \frac{2}{r})^{\Delta}\}
\end{aligned}$$

Now we can bound the success probability of the decision rule we defined previously.

$$\begin{aligned}
\mathbb{P}(\text{A wrong decision is made}) &= \mathbb{P}\{\Delta < r, \text{we report } \Delta \geq (1+\epsilon)r\} + \mathbb{P}\{\Delta \geq (1+\epsilon)r, \text{we report } \Delta \leq r\} \\
&= \mathbb{P}(\Delta < r, \hat{\Delta} > (1+\epsilon)\frac{k}{2}\{1 - (1 - \frac{2}{r})^r\}) \\
&\quad + \mathbb{P}(\Delta \geq (1+\epsilon)r, \hat{\Delta} < (1-\epsilon)\frac{k}{2}\{1 - (1 - \frac{2}{r})^r\}) \\
&\leq \mathbb{P}(\hat{\Delta} > \frac{k}{2}\{1 - (1 - \frac{2}{r})^{\Delta}\} + \epsilon\frac{k}{2}\{1 - (1 - \frac{2}{r})^r\}) \\
&+ \mathbb{P}(\hat{\Delta} < \frac{k}{2}\{1 - (1 - \frac{2}{r})^{\Delta}\} - \epsilon\frac{k}{2}\{1 - (1 - \frac{2}{r})^r\}) \\
&= \mathbb{P}\{|\hat{\Delta} - \mathbb{E}(\hat{\Delta})| > \epsilon\frac{k}{2}\{1 - (1 - \frac{2}{r})^r\}\} \\
&\leq \frac{4\text{var}(\hat{\Delta})}{\epsilon^2 k^2}\{1 - (1 - \frac{2}{r})^r\}^{-2} \qquad \text{(by Chebyshev's inequality)} \\
&\leq \frac{2}{\epsilon^2 k}\{1 - (1 - \frac{2}{r})^{\Delta}\}\{1 - (1 - \frac{2}{r})^r\}^{-2}
\end{aligned}$$

Since $\log(1 - 2/r) \leq -2/r$, we have $(1 - 2/r)^r = \exp\{r\log(1 - 2/r)\} \leq \exp(-2)$ and noting that $1 - (1 - \frac{2}{r})^{\Delta} < 1$, we have

$$\mathbb{P}(\text{A wrong decision is made}) \leq \frac{2}{\epsilon^2 k}\frac{1}{(1 - \exp(-2))^2}.$$

Thus setting $k = O(\epsilon^{-2})$ is sufficient to achieve a constant success probability. □

# 2 Approach 1: Nearest Neighbour Search

**Definition 2.** *A c- approximate r- nearest neighbour search procedure is a procedure that given a query point q which returns a point $p'$ in the domain such that* $normp' - q \leq cr$ *given that* $c > 1$ *and there exists* $p^*$ *that* $\|p^* - q\| \leq r$.

## 2.1 Boosted Sketch

Let $S$ be the sketch in the decision version probability that based on which a decision rule with constant success probability could be generated. Define a new sketch $W$ by keeping $k = O(\log n)$ copies of S and the decision is the majority answer of the $k$ decisions. Therefore the sketch size is $O(\epsilon^{-2} \log n)$ and the success probability is $1 - 1/n^2$. To use this method we need to compute all points $p$ in the domain in advance and when we make a query we need to compute $W(q)$, and compute distance to all points using sketch. There is an improvement of computation time from $O(nd)$ to $O(n\epsilon^{-2} \log n)$.

## 2.2 Approach 2

The goal of this approach is to improve computation time from $O(n\epsilon^{-2} \log n)$ to $O(n)$. The result is given by the following theorem.

**Theorem 3** (KOR98)**.** *To achieve* $(1+\epsilon)$*- approximation of nearest neighbour search procedure,* $O(d\epsilon^{-2} \log n)$ *query time and* $n^{O(\epsilon^{-2})}$ *is sufficient.*

The idea of the construction and the proof is by noting that $W(q)$, which is defined in "Approach 1",has $w = O(\epsilon^{-2} \log n)$ bits and there are only $2^w$ possible sketches. Thus we can store an answer for each of $2^w = n^{O(\epsilon^{-2})}$ possible inputs.