CS698C: Sketching and Sampling for Big Data Analysis

CountSketch Random Matrix and J-L Distribution

The *CountSketch* is a well-known data structure invented in 2001. Here, we describe the same data structure in the form of a random matrix. It is based on using a random hash function h that maps $h : \{1, \ldots, n\} \to \{1, 2, \ldots, s\}$. s is often referred to as the size of the hash table or the number of buckets in the hash table. The item i is mapped to the position $h(i)$. The random hash function is drawn uniformly at random from a family of hash function $\mathcal{H}$. It satisfies a d-wise independence property, which implies, 2-wise through d-wise independence.

Firstly, for any $j \in \{1, 2, \ldots, n\}$ and $i \in \{1, 2, \ldots, s\}$

$$\mathrm{P}\left[h(j) = i\right] = \frac{1}{s}$$

that is, each input element i is mapped with uniform probability to any of the s buckets.

Pair-wise independence. Suppose j and j' are two distinct items $j \neq j'$, $j, j' \in \{1, 2, \ldots, n\}$. Then, the pair-wise independence property states that, for any $i, i' \in \{1, 2, \ldots, s\}$,

$$\mathrm{P}\left[h(j) = i \text{ and } h(j') = i'\right] = \mathrm{P}\left[h(j) = i\right] \cdot \mathrm{P}\left[h(j') = i'\right] = \frac{1}{s^2} \ .$$

It states that for any pair of distinct items j and j' from the domain of items $\{1, 2, \ldots, n\}$, the probability that under the hash function, jointly, h maps j to bucket i and j' to bucket i', for any $i, i' \in \{1, 2, \ldots, s\}$, is the product of their individual probabilities, and hence, $\frac{1}{s^2}$.

d-wise independence. Suppose $j_1, j_2, \ldots, j_d$ are some d distinct items from the set $\{1, 2, \ldots, n\}$. Let $i_1, i_2, \ldots, i_d$ be d positions from the set $\{1, 2, \ldots, s\}$, where, $i_1, \ldots, i_d$ could overlap (arbitrarily). The d-wise independence property states that

$$\mathrm{P}\left[h(j_1) = i_1, h(j_2) = i_2, \ldots, h(j_d) = i_d\right] = \frac{1}{s^d} = \prod_{r=1}^{d} \mathrm{P}\left[h(j_r) = i_r\right] \ .$$

d-wise independence implies $d-1$ wise indpendence. Suppose $\{j_1, \ldots, j_{d-1}\}$ are $d-1$ distinct item set from $\{1, 2, \ldots, n\}$. Let $i_1, i_2, \ldots, i_{d-1}$ be arbitrary d positions from $\{1, 2, \ldots, s\}$. Let the hash family $\mathcal{H}$ from which h is drawn is d-wise independent. Pick an arbitrary item distinct from $j_1, \ldots, j_{d-1}$ and call it j_d. Then, for any $i_d \in \{1, 2, \ldots, m\}$,

$$\mathrm{P}\left[h(j_1) = i_1, \ldots, h(j_{d-1}) = i_{d-1}, h(j_d) = i_d\right] = \frac{1}{s^d}.$$

We wish to show that

$$\mathrm{P}\left[h(j_1) = i_1, \ldots, h(j_{d-1}) = i_{d-1}\right] = \frac{1}{m^{d-1}} \ .$$

Define the event

$$E = \{h(j_1) = i_1, \ldots, h(j_{d-1}) = i_{d-1}\} \ .$$

We wish to show that $\mathrm{P}\left[E\right] = \frac{1}{m^{d-1}}$. Note that j_d can take a value i_d between 1 and m and so,

$$E = \{h(j_1) = i_1, \ldots, h(j_{d-1}) = i_{d-1}\} = \cup_{i_d=1}^{m} \{h(j_1) = i_1, \ldots, h(j_{d-1}) = i_{d-1}, h(j_d) = i_d\}$$

and the events $\{h(j_1) = i_1, \dots, h(j_{d-1}) = i_{d-1}, h(j_d) = i_d\}$ are disjoint. Therefore,

$$
\begin{aligned}
\mathrm{P}[E] &= \sum_{i_d=1}^{m} \mathrm{P}[h(j_1) = i_1, \dots, h(j_{d-1}) = i_{d-1}, h(j_d) = i_d] \\
&= \sum_{i_d=1}^{m} \frac{1}{m^d} = \frac{m}{m^d} = \frac{1}{m^{d-1}} \quad .
\end{aligned}
$$

Likewise, it can be shown that if $d \geq 3$ and h is $d-1$-wise independent, then, h is $d-2$-wise independent, and so on downward.

CountSketch Matrix. The Countsketch random matrix S is an $m \times n$ random matrix defined as follows. As per the definition above, let h be a random hash function chosen from a family of hash functions $\mathcal{H}$ that map $h : \{1, \dots, n\} \to \{1, 2, \dots, m\}$. For analysis, we will assume 4-wise independence. For each column $j \in \{1, 2, \dots, n\}$,

$$
S_{i,j} = \begin{cases} \epsilon_{i,j} & \text{if } i = h(i) \\ 0 & \text{otherwise.} \end{cases}
$$

That is, in the column j of S, there is exactly one non-zero entry in row index $i = h(i)$, and it is filled by an iid Rademacher variable ϵ_{ij}, all other entries in this column j in S are zeros.

The matrix S is a sparse matrix. Every column has exactly one entry which is a random ± 1 sign. So there are a total of only n non-zero entries in this matrix, one per column.

As per the usual notion of dimensionality reduction, S maps a given n-dimensional vector x to an m-dimensional vector Sx.

Analysis. For simplicity, we will assume that x is a unit vector, so that $\|x\|_2^2 = 1$. Let us proceed with an analysis of the concentration of $\|Sx\|_2^2$. All norms are $\|\cdot\|_2$ norms. Equivalently, we write

$$
S_{ij} = b_{ij}\epsilon_{ij}, \qquad i \in \{1, 2, \dots, m\}, j \in \{1, 2, \dots, n\}
$$

The ϵ_{ij}'s are iid Rademacher variables. The b_{ij}'s are random Bernoulli variables, *but not independent.* They are Bernoulli variables, in that they take values either 0 or 1. By construction, for any column index j,

$$
b_{1j} + b_{2j} + \cdots + b_{mj} = 1 \quad .
$$

So, in a column, the Bernoulli variables add to exactly 1, and hence $\{b_{ij} : i = 1, \dots, n\}$ are clearly not independent. Across the columns, we will only assume 4-wise independence of the b_{ij}'s. This follows from the four-wise independence assumption of the hash family from which h is drawn, namely, for any set of four distinct elements j_1, j_2, j_3, j_4 from $\{1, 2, \dots, n\}$ and any four position in $\{1, 2, \dots, s\}$,

$$
\mathrm{P}[h(j_1) = i_1, h(j_2) = i_2, h(j_3) = i_3, h(j_4) = i_4] = \frac{1}{m^4} = \prod_{r=1}^{r} \mathrm{P}[h(j_r) = i_r] \quad .
$$

It follows that for distinct j_1, j_2, j_3, j_4,

$$
\begin{aligned}
&\mathrm{P}\left[b_{i_1,j_1} = 1, b_{i_2,j_2} = 1, b_{i_3,j_3} = 1, b_{i_4,j_4} = 1\right] \\
&= \mathrm{P}\left[h(j_1) = i_1, h(j_2) = i_2, h(j_3) = i_3, h(j_4) = i_4\right] \\
&= \prod_{r=1}^{4} \mathrm{P}\left[b_{i_r,j_r} = 1\right] \\
&= \frac{1}{m^4} \ .
\end{aligned}
$$

Therefore, for distinct column indices j, j' and the same row index i, the Bernoulli variable product b_{ij} and $b_{ij'}$ are independent; $\mathrm{P}\left[b_{ij} = 1, b_{ij'} = 1\right] = \mathrm{P}\left[b_{ij} = 1\right] \mathrm{P}\left[b_{ij'} = 1\right] = \frac{1}{m^2}$. Therefore,

$$
\mathrm{E}\left[b_{ij} b_{ij'}\right] = \mathrm{E}\left[b_{ij}\right] \mathrm{E}\left[b_{ij'}\right] .
$$

For the same column j, b_{ij}'s are not pair-wise independent. In particular, for $i \neq i'$, $b_{ij} b_{i'j} = 1$ iff $h(j) = i$ and $h(j) = i'$, which clearly cannot happen, since h is a function. Hence, for $i \neq i'$, $b_{ij} b_{i'j} = 0$ and therefore, $\mathrm{P}\left[b_{ij} b_{i'j} = 1\right] = 0$, $i \neq i'$.

$\mathrm{E}\left[\|Sx\|^2\right]$. Let us first calculate $\mathrm{E}\left[\|Sx\|^2\right]$. The rows of Sx are denoted as $(Sx)_i$, $i = 1, 2, \ldots, m$.

$$
(Sx)_i = \sum_{j=1}^{n} b_{ij} \epsilon_{ij} x_j, \qquad i = 1, 2, \ldots, m \ .
$$

Therefore,

$$
\begin{aligned}
(Sx)_i^2 &= \left(\sum_{j=1}^{n} b_{ij} \epsilon_{ij} x_j \right)^2 \\
&= \sum_{j=1}^{n} b_{ij}^2 \epsilon_{ij}^2 x_j^2 + 2 \sum_{j<j'} b_{ij} b_{ij'} \epsilon_{ij} \epsilon ij' x_j x_{j'}
\end{aligned}
$$

Note that Rademacher variable squared $\epsilon_{ij}^2 = 1$, and Bernoulli variable squared $b_{ij}^2 = b_{ij}$, since, Bernoulli variables take values 0 or 1. Using this and simplifying, we have,

$$
(Sx)_i^2 = \sum_{j=1}^{n} b_{ij} x_j^2 + 2 \sum_{j<j'} b_{ij} b_{ij'} \epsilon_{ij} \epsilon_{ij'} x_j x_{j'}
$$

We now sum $(Sx)_i^2$ across the rows.

$$
\begin{aligned}
\|Sx\|^2 &= \sum_{i=1}^{m} (Sx)_i^2 \\
&= \sum_{i=1}^{m} \sum_{j=1}^{n} b_{ij} x_j^2 + 2 \sum_{i=1}^{m} \sum_{1 \leq j < j' \leq n} b_{ij} b_{ij'} \epsilon_{ij} \epsilon_{ij'} x_j x_{j'}
\end{aligned} \tag{1}
$$

Interchanging the summation across rows and columns,

$$
\sum_{i=1}^{m} \sum_{j=1}^{n} b_{ij} x_j^2 = \sum_{j=1}^{n} x_j^2 \left(\sum_{i=1}^{m} b_{ij} \right) = \sum_{j=1}^{n} x_j^2 \cdot 1 = \sum_{j=1}^{n} x_j^2 = \|x\|_2^2 \ .
$$

The second equality uses the fact that exactly one of b_{ij}'s in the jth column is a 1, and all others are 0, and therefore, $\sum_{i=1}^{m} b_{ij} = 1$, for each $j = 1, \ldots, n$.

Since, x is assumed to be a unit vector, and substituting in Equation (1), we have,

$$\begin{aligned}\|Sx\|^2 &= \|x\|_2^2 + 2\sum_{i=1}^{m} \sum_{1\le j<j'\le n} b_{ij}b_{ij'}\epsilon_{ij}\epsilon_{ij'}x_j x_{j'} \\ &= 1 + 2\sum_{i=1}^{m} \sum_{1\le j<j'\le n} b_{ij}b_{ij'}\epsilon_{ij}\epsilon_{ij'}x_j x_{j'}\end{aligned}$$

We therefore have,

$$\|Sx\|^2 - 1 = 2\sum_{i=1}^{m} \sum_{1\le j<j'\le n} b_{ij}b_{ij'}\epsilon_{ij}\epsilon_{ij'}x_j x_{j'} \tag{2}$$

Let us now take the expectation on both sides.

$$\mathrm{E}\left[\|Sx\|^2 - 1\right] = 2\sum_{i=1}^{m} \sum_{1\le j<j'\le n} \mathrm{E}\left[b_{ij}b_{ij'}\right]\mathrm{E}\left[\epsilon_{ij}\epsilon_{ij'}\right] \tag{3}$$

The Bernoulli family of variables $\{b_{ij}\}$'s are independent of the Rademacher family of variables $\{\epsilon_{ij}\}$. Hence, $\mathrm{E}\left[b_{ij}b_{ij'}\epsilon_{ij}\epsilon_{ij'}\right] = \mathrm{E}\left[b_{ij}b_{ij'}\right]\mathrm{E}\left[\epsilon_{ij}\epsilon ij'\right]$. The Rademacher variables are all iid, and hence,

$$\mathrm{E}\left[\epsilon_{ij}\epsilon_{ij'}\right] = \mathrm{E}\left[\epsilon_{ij}\right]\mathrm{E}\left[\epsilon_{ij'}\right] = 0\cdot 0 = 0 \ .$$

As discussed earlier, the Bernoulli variables b_{ij} and $b_{ij'}$ are independent, for distinct column indices j, j' and so,

$$\mathrm{E}\left[b_{ij}b_{ij'} = 1\right] = \mathrm{P}\left[b_{ij} = 1, b_{ij'} = 1\right] = \mathrm{P}\left[b_{ij} = 1\right]\mathrm{P}\left[b_{ij'} = 1\right] = \frac{1}{m^2} \ .$$

It follows that Equation (3) becomes

$$\mathrm{E}\left[\|Sx\|^2 - 1\right] = 2\sum_{i=1}^{m} \sum_{1\le j<j'\le n} \mathrm{E}\left[b_{ij}b_{ij'}\right]\mathrm{E}\left[\epsilon_{ij}\epsilon_{ij'}\right] = 2\sum_{i=1}^{m} \sum_{1\le j<j'\le n} \frac{1}{m^2}\cdot 0\cdot 0 = 0. \tag{4}$$

So we have for unit vector x,

$$\mathrm{E}\left[\|Sx\|^2 - 1\right] = 0 \quad , \text{ or,} \quad \mathrm{E}\left[\|Sx\|^2\right] = 1 \ .$$

The idea of the proof is to use Markov's inequality for the random variable $\|Sx\|^2 - 1$ using the second moment. (This is same as Chebychev's inequality). That is,

$$\mathrm{P}\left[\left|\|Sx\|^2 - 1\right| \le \epsilon\right] = \mathrm{P}\left[\left(\|Sx\|^2 - 1\right)^2 \le \epsilon^2\right] \le \frac{\mathrm{E}\left[(\|Sx\|^2 - 1)^2\right]}{\epsilon^2}$$

So we first attempt to give an upper bound on $\mathrm{E}\left[(\|Sx\|^2 - 1)^2\right]$.

Calculation of $\mathrm{E}\left[(\|Sx\|^2-1)^2\right]$. Continuing from Equation (2), we have,

$$(\|Sx\|^2-1)^2 = \left(\sum_{i=1}^{m}\sum_{1\le j<j'\le n} 4b_{ij}b_{ij'}\epsilon_{ij}\epsilon_{ij'}x_jx_{j'}\right)^2 \tag{5}$$

To simplify the calculation, let us write

$$\|Sx\|^2-1 = \sum_{i=1}^{m} T_i$$

where the term $T_i = 2\sum_{1\le j<j'\le n} b_{ij}b_{ij'}\epsilon_{ij}\epsilon_{ij'}x_jx_{j'}$ correspond to the sum expression pertaining to indices of row i only. With this notation,

$$\left(\|Sx\|^2-1\right)^2 = \left(\sum_{i=1}^{m} T_i\right)^2 = \sum_{i=1}^{m} T_i^2 + 2\sum_{i<i'} T_iT_{i'}$$

We take the expectation of the terms on the *RHS* step at a time.

First, consider distinct row indices $i \ne i'$ and consider $T_iT_{i'}$. Each T_i (resp. $T_{i'}$) are summations of terms of the form $b_{ij}b_{ij'}\epsilon_{ij}\epsilon_{ij'}x_jx'_j$, for $j \ne j'$. Likewise, a typical term in $T_{i'}$ is of the form $b_{i'l}b_{i'l'}\epsilon_{i'l}\epsilon_{i'l'}x_lx_{l'}$. Momentarily, we are denoting distinct columns as l, l', where, $\{l, l'\}\cap\{j, j'\}$ may have no intersection, or an intersection of one column index, or may be identical. However note that by independence of iid Rademacher variables, all the four variables in the respective terms of T_i and $T_{i'}$, namely,

$$\epsilon_{ij}, \epsilon_{ij'}, \epsilon_{i'l}, \epsilon_{i'l'}$$

are all independent, as they belong to different rows i, i', or within the same row, belong to respectively different cols, j, j' or l, l'. Since Rademacher variables have 0 expectation,

$$\mathrm{E}\left[\epsilon_{ij}\epsilon_{ij'}\epsilon_{i'l}\epsilon_{i'l'}\right] = \mathrm{E}\left[\epsilon_{ij}\right]\mathrm{E}\left[\epsilon_{ij'}\right]\mathrm{E}\left[\epsilon_{i'l}\right]\mathrm{E}\left[\epsilon_{i'l'}\right] = 0\cdot 0\cdot 0\cdot 0 = 0, \quad i\ne i', j\ne j', l\ne l' \ .$$

Thus, without taking into further consideration the Bernoulli variables, we have that for $i \ne i'$,

$$\begin{aligned}\mathrm{E}\left[T_iT_{i'}\right] &= \mathrm{E}\left[b_{ij}b_{ij'}b_{i'l}b_{i'l'}\right]\mathrm{E}\left[\epsilon_{ij}\epsilon_{ij'}\epsilon_{i'l}\epsilon_{i'l'}\right]x_jx_{j'}x_lx_{l'}\\ &= \mathrm{E}\left[b_{ij}b_{ij'}b_{i'l}b_{i'l'}\right]\cdot 0\cdot x_jx_{j'}x_lx_{l'}\\ &= 0 \ .\end{aligned}$$

We have the following simplification now:

$$\mathrm{E}\left[\left(\|Sx\|^2-1\right)^2\right] = \sum_{i=1}^{m}\mathrm{E}\left[T_i^2\right] . \tag{6}$$

We now calculate $\mathrm{E}\left[T_i^2\right]$, for some fixed row index i. For column index pairs (j, l), where, $j < l$, denote by $t_{j,l}$ the correspnding term in the sum for T_i, namely,

$$t_{j,l} = 2b_{ij}b_{il}\epsilon_{ij}\epsilon_{il}x_jx_l.$$

Then,

$$T_i = \sum_{\text{ordered pairs over }\{1,\ldots,\text{n}\}\ (j,l):j<l} t_{jl} \ .$$

We would like to calculate $\mathrm{E}\left[T_i^2\right]$. Consider T_i^2. In the following, we always assume that the ordered pair (j,l) or (j',l') have $j<l$, $j'<l'$, etc.

$$T_i^2 = \sum_{(j,l)} t_{j,l}^2 + \sum_{(j,l)\neq(j',l')} t_{j,l}t_{j',l'} .$$

Using linearity of expectation,

$$\mathrm{E}\left[T_i^2\right] = \sum_{(j,l)} \mathrm{E}\left[t_{j,l}^2\right] + \sum_{(j,l)\neq(j',l')} \mathrm{E}\left[t_{j,l}t_{j',l'}\right] . \tag{7}$$

Note that if pairs (j,l) and (j',l') are not equal, then, the corresponding terms t_{jl} and $t_{j'l'}$ must have at least **one unpaired** Rademacher variables among these four Rademacher variables

$$\epsilon_{ij}, \epsilon_{il}, \epsilon_{ij'}, \epsilon_{il'} \ .$$

This is because since $(j,l)\neq(j',l')$, both equalities $j=j'$ and $l=l'$ do not hold together. At least one of $j\neq j'$ or $l\neq l'$ holds. If $j\neq j'$, then, ϵ_{ij} and $\epsilon_{ij'}$ are different random variables and independent. For both to be paired the set $\{j,j'\}$ must equal the set $\{l,l'\}$. Since, $j<l$, the only possibility for such a pairing is for $l=j'$. In this case, $l'>j'>j$, and so the variables ϵ_{ij} and $\epsilon_{il'}$ each remain unpaired. This shows that among the four Rademacher variables $\epsilon_{ij}, \epsilon_{il}, \epsilon_{ij'}, \epsilon_{il'}$, two remain unpaired, and at most one is paired. In the above example, with $j'=l$, the product

$$\epsilon_{ij}, \epsilon_{il}, \epsilon_{ij'}, \epsilon_{il'} = \epsilon_{ij}\epsilon_{ij'}^2\epsilon_{il'} .$$

Taking expectations, unpaired Rademacher variables have exponent 1 and their expectation is 0, and hence, the expectation of the product is 0. This shows that

$$\mathrm{E}\left[t_{j,l}t_{j',l'}\right] = 0, \quad (j,l)\neq(j'l') \ .$$

Substituting in Equation (7), we have,

$$\mathrm{E}\left[T_i^2\right] = \sum_{(j,l)} \mathrm{E}\left[t_{j,l}\right]^2 \ . \tag{8}$$

Now,

$$t_{j,l}^2 = 4b_{ij}b_{il}\epsilon_{ij}^2\epsilon_{il}^2x_j^2x_l^2 = 4b_{ij}b_{il}x_j^2x_l^2 \ .$$

Therefore, for ordered pairs (j,l) with $1\leq j<l\leq n$

$$\sum_{(j,l)} \mathrm{E}\left[t_{j,l}^2\right] = 4\sum_{(j,l)} \mathrm{E}\left[b_{ij}b_{il}\right]x_j^2x_l^2 = \sum_{(j,l)} \mathrm{E}\left[b_{ij}\right]\mathrm{E}\left[b_{il}\right]x_j^2x_l^2 = 4\sum_{(j,l)} \frac{1}{m}\cdot\frac{1}{m}x_j^2x_l^2$$

where, the last step uses pair-wise independence of $b_{ij}, b_{i'l}$, where, $j\neq l$. Continuing, this gives,

$$\begin{aligned}\sum_{(j,l)} \mathrm{E}\left[t_{j,l}^2\right] &= \frac{4}{m^2}\sum_{(j,l)} x_j^2x_l^2 \\ &= \frac{1}{m^2}\cdot\frac{1}{2}\left[\left(\sum_{j=1}^n x_j^2\right)\left(\sum_{l=1}^n x_l^2\right) - \sum_{k=1}^n x_k^4\right]\end{aligned}$$

where, the last step is a rewriting of

$$\sum_{j=1}^{n} x_j^2 \sum_{l=1}^{n} x_l^2 = (\sum_{k=1}^{n} x_k^2)^2 = \sum_{k=1}^{n} x_k^4 + 2 \sum_{(j,l)} x_j^2 x_l^2 .$$

Since, x is a unit vector, $\sum_j x_j^2 = 1$, and so,

$$\mathrm{E}\left[T_i^2\right] = \sum_{(j,l)} \mathrm{E}\left[t_{j,l}^2\right] = \frac{4}{2m^2}\left[1 - \sum_{k=1}^{n} x_k^4\right]$$

Therefore,

$$\mathrm{E}\left[(\|Sx\|^2 - 1)^2\right] = \sum_{i=1}^{m} \mathrm{E}\left[T_i^2\right] = \frac{2m}{m^2}\left[1 - \|x\|_4^4\right] \le \frac{2}{m} \ .$$

Therefore, applying Chebychev's inequality, we have,

$$\mathrm{P}\left[\left|\|Sx\|^2 - 1\right| \ge \epsilon\right] \le \frac{\mathrm{E}\left[\left(\|Sx\|^2 - 1\right)^2\right]}{\epsilon^2} \le \frac{2}{\epsilon^2 m} \ . \tag{9}$$

Therefore, $\mathrm{P}\left[\left|\|Sx\|^2 - 1\right| \ge \epsilon\right] < \delta$, provided, m is chosen so that

$$\frac{2}{\epsilon^2 m} \le \delta$$

or, that

$$m \ge \frac{2}{\epsilon^2 \delta} \ .$$

We have thus shown the following lemma.

Lemma 1. *Let S be a random Countsketch matrix of dimension m by n. If $m \ge \frac{2}{\epsilon^2 \delta}$, and for any fixed unit vector x,*

$$P\left[\left|\|Sx\|^2 - 1\right| \le \epsilon\right] \ge 1 - \delta \ .$$