CS261 Winter 2018 - 2018
Lecture 16: Sketching (Part 3)

Instructor: Ashish Goel
Scribe: Kaidi Yan
Edited: Geoff Ramseyer

March 4, 2019

1 Cauchy Distribution

In previous lectures, we introduced the notion of $p-$stable distributions. In most applications, we will use 2-stable distributions, but occasionally, one might want to use a distribution that is stable for some other p. A distribution that is stable for $p = 1$ is the Cauchy distribution.

Consider the following probability density function.

$$f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$$

Suppose Z is a random variable drawn from the Cauchy distribution. The expected value of $|Z|$ is:

$$\mathrm{E}[|Z|] = 2\int_0^\infty \frac{1}{\pi}\frac{x}{1+x^2}dx$$

Using variable substitution $t = 1 + x^2$,

$$\mathrm{E}[|Z|] = \frac{1}{\pi}\int_1^\infty \frac{1}{t}dt = [\ln t]_1^\infty = \infty$$

Thus the mean of $|Z|$ is unbounded.

If $X_1, ..., X_M$ are i.i.d. random variables drawn from the Cauchy distribution and $a_1, ..., a_M \in \mathbb{R}$ are arbitrarily chosen, then

$$\sum_{i=1}^{M} a_i X_i = \sum_{i=1}^{M} |a_i| Z$$

where Z is also drawn from the Cauchy distribution. Hence the Cauchy distribution is 1-stable.

Using the fact that the Cauchy distribution is 1-stable, we can construct a new sketch to estimate $F_1(t)$. Assume $h_j(k)$ is a consistent Cauchy hash function for $j = 1, ..., M$. Define the sketch as following:

$$\sigma(S) = \langle \sigma_1(S), \sigma_2(S), ..., \sigma_M(S) \rangle$$

where $\sigma_j(S)$ $(j = 1, ..., M)$ is defined as:

$$\begin{aligned} \sigma_j(S) &= \sum_{i=1}^{t} v_i h_j(k_i) \\ &= \sum_{k \in \{k_1, ..., k_t\}} h_j(k) \sum_{i \in [1,t]: k_i = k} v_i \end{aligned}$$

Define $V_k(t) = \sum_{i \in [1,t]: k_i = k} v_i$. Since $h_j(k)$ are i.i.d. and drawn from a 1-stable distribution, we have:

$$\begin{aligned} \sigma_j(S) &= \sum_{k \in \{k_1, ..., k_t\}} h_j(k) \sum_{i \in [1,t]: k_i = k} v_i \\ &= \sum_{k \in \{k_1, ..., k_t\}} h_j(k) V_k(t) \\ &= Z_j \sum_{k \in \{k_1, ..., k_t\}} |V_k(t)| \end{aligned}$$

where Z_j follows the Cauchy distribution. Taking the absolute value on both side we get:

$$|\sigma_j(S)| = |Z_j| \sum_{k \in \{k_1, ..., k_t\}} |V_k(t)| = |Z_j| F_1(t)$$

Since E$[|Z_j|]$ is infinity as we have shown just now, we can only apply the median lemma here. Using the median lemma, we know that the median of $|\sigma_1(S)|, ..., |\sigma_M(S)|$, which is equivalent to the median of $F_1(t)|Z_1|, ..., F_1(t)|Z_M|$, is within the range $[F_1(t)G_D^{-1}(1/2 - \delta), F_1(t)G_D^{-1}(1/2 + \delta)]$ with probability at least $1 - \epsilon$ if $M > \frac{c}{\delta^2} \log \frac{1}{\epsilon}$ for some constant c, where G_D is the culmulative density function of the Cauchy distribution. Hence we can define the following estimator of $F_1(t)$:

$$\widehat{F_1(t)} = \frac{\text{Median}(|\sigma_1(S)|, ..., |\sigma_M(S)|)}{G_D^{-1}(1/2)} = \frac{\text{Median}(|\sigma_1(S)|, ..., |\sigma_M(S)|)}{\text{Median}(Z)}$$

where Z follows the Cauchy distribution. The last step is to compute Median(Z). We know that for any k,

$$G_D(k) = 2 \int_0^k \frac{1}{\pi} \frac{1}{1 + x^2} dx = \frac{2}{\pi} [\arctan(x)]_0^k = \frac{2}{\pi} \arctan(k)$$

Hence for $G_D(k) = 1/2$, we need to have $\frac{2}{\pi} \arctan(k) = 1/2$, implying that $k = 1$. Thus Median$(Z) = 1$.

2 Median versus Mean

In the previous two lectures, when we discussed min-hash sketches and sketches of $F_2(t)$, we claimed that the guarantee given by the median lemma also holds for the mean. As a recap, the median lemma states that:

"There exists a constant $c > 0$ such that for all distributions D where G_D is continuous, for all $\delta \in (0, 1/2)$ and for all $\epsilon > 0$, if $X_1, X_2, ..., X_M$ are i.i.d random variables with distribution D and $M > \frac{c}{\delta^2} \ln(\frac{1}{\epsilon})$, then $\text{Median}(X_1, ..., X_M)$ lies within $[G_D^{-1}(\frac{1}{2}-\delta), G_D^{-1}(\frac{1}{2}+\delta)]$ with probability at least $1 - \epsilon$."

The statement of a "mean lemma"[1] is simply the above lemma, with "median" replaced by "mean." Unfortunately, the mean lemma only holds for certain types of distributions D, such as the Bernoulli distribution and the square of $N(0, 1)$. This result is typically called Hoeffding's inequality, though it is often also called a Chernoff bound or the Chernoff-Hoeffding inequality.

As an intuition, the "mean" lemma tends to only hold when the "tails" of the distribution, i.e. the probability of producing highly extreme values, are very small. The median as a metric is much more robust against outliers, and as such, holds under weaker conditions on the distribution.

As an example of the proof techniques involved, here is an informal proof of the median lemma, provided that we already proved the "mean lemma" for Bernoulli random variables.. Suppose we have M i.i.d. random variables $X_1, ..., X_M$ following a distribution D. Define new variables Z_i ($i = 1, ..., M$) as following:

$$Z_i = \begin{cases} 1 & \text{if } X_i \geq G_D^{-1}(\frac{1}{2} - \delta) \\ 0 & \text{otherwise} \end{cases}$$

Then $Z_1, ..., Z_M$ are i.i.d. random variables following a Bernoulli distribution, meaning that we can apply the "mean lemma" to these variables. Notice that for each i from 1 to M:

$$\text{E}[Z_i] = \Pr[X_i \geq G_D^{-1}(\frac{1}{2} - \delta)] = \frac{1}{2} + \delta$$

Hence

$$\frac{1}{M}\text{E}[\sum_{i=1}^{M} Z_i] = \frac{1}{M}\sum_{i=1}^{M} \text{E}[Z_i] = \frac{1}{M}\sum_{i=1}^{M}(\frac{1}{2} + \delta) = \frac{1}{2} + \delta \qquad (1)$$

Now let's consider the probability that $\text{Median}(X_1, ..., X_M) < G_D^{-1}(\frac{1}{2} - \delta)$. If this happens then we know at least half of $X_1, ..., X_M$ are less than $G_D^{-1}(\frac{1}{2} - \delta)$, which implies that at least half of $Z_1, ..., Z_M$ are 0. Thus $\frac{1}{M}\sum_{i=1}^{M} Z_i \leq \frac{1}{2}$. Given (1) and applying the "mean lemma" with 2δ as the error bound and $\epsilon/2$ as the probability bound,, we know that $\Pr[\frac{1}{M}\sum_{i=1}^{M} Z_i \leq \frac{1}{2}]$ happens with at most probability $\frac{1}{2}\epsilon$. Thus we prove that $\text{Median}(X_1, ..., X_M) < G_D^{-1}(\frac{1}{2}-\delta)$ happens with probability at most $\frac{1}{2}\epsilon$.

[1] We will revisit the "mean lemma" later in this class.

For the otherside where Median$(X_1, ..., X_M) > G_D^{-1}(\frac{1}{2} + \delta)$, we can use an analogous technique to show that it also happens with probability at most $\frac{1}{2}\epsilon$. Therefore we prove the median lemma.

3 Johnson-Lindenstrauss Lemma

Suppose that we have some number of data points in some high dimensional space. Many statistical analysis algorithms suffer from what is sometimes called the “curse of dimensionality,” in that their runtimes contain some very strong (i.e. exponential) dependence on the dimension. As such, it would be very useful if one could embed a set of high-dimensional points into some lower dimensional space, while still, for example, preserving the (relativized) distances between each pair of points. The surprising fact is that multiplying each point by a random matrix suffices for this purpose.

More formally, given N points $X_1, ..., X_N \in \mathbb{R}^J$, the problem is to estimate $||X_i - X_k||_2$ for any two points X_i and X_k. The interesting case of this problem is when J is huge, and we are reducing the dimension of our data points to M for some $M << J$.

To do so, we generate M vectors $R_1, ..., R_M \in \mathbb{R}^J$. For each $j = 1, ..., M$, each component of R_j is drawn independently from $N(0, 1)$. Let the ith component of R_j be R_{ji}. Then we define $R_{ji} = h(\langle j, i\rangle)$, where $h(\langle j, i\rangle)$ is a consistent Gaussian hash function. Lastly, let R be a $M \times J$ matrix where the j-th row of R is exactly R_j^T.

Now for $i = 1, ..., N$, define $\hat{X}_i = RX_i$. Clearly $\hat{X}_i \in \mathbb{R}^M$. The Johnson-Lindenstrauss Lemma states the following result:

Theorem 1 (Johnson-Lindenstrauss Lemma) *If $M > \frac{c}{\delta^2} \log \frac{N}{\epsilon}$ for some constant c, then for all X_i, X_k where $i, k = 1, ..., N$, $||\hat{X}_i - \hat{X}_k||_2 \in [\sqrt{M}||X_i - X_k||_2(1 - \delta), \sqrt{M}||X_i - X_k||_2(1 + \delta)]$ with probability at least $1 - \epsilon$.*

Here's an proof sketch of Johnson-Lindenstrauss Lemma. We know that $\hat{X}_i = RX_i$. Therefore:

$$\begin{aligned} ||\hat{X}_i - \hat{X}_k||_2^2 &= ||RX_i - RX_k||_2^2 && (2) \\ &= ||R(X_i - X_k)||_2^2 && (3) \\ &= \sum_{j=1}^{M} (R_j \cdot (X_i - X_k))^2 && (4) \end{aligned}$$

For each j, we can write $R_j \cdot (X_i - X_k)$ as:

$$R_j \cdot (X_i - X_k) = \sum_{l=1}^{J} R_{jl}(X_{il} - X_{kl})$$

But remember R_{jl} are drawn independently from $N(0,1)$, which is a 2-stable distribution. Hence we have:

$$
\begin{aligned}
R_j \cdot (X_i - X_k) &= \sum_{l=1}^{J} R_{jl}(X_{il} - X_{kl}) && (5)\\
&= \sqrt{\sum_{l=1}^{J}(X_{il} - X_{kl})^2} Z_j && (6)\\
&= ||X_i - X_k||_2 Z_j && (7)
\end{aligned}
$$

where $Z_j \sim N(0,1)$. Substituting (7) back to (4) we get:

$$
\begin{aligned}
||\hat{X}_i - \hat{X}_k||_2^2 &= ||X_i - X_k||_2^2 \sum_{j=1}^{M} Z_j^2\\
||\hat{X}_i - \hat{X}_k||_2 &= ||X_i - X_k||_2 \sqrt{\sum_{j=1}^{M} Z_j^2}
\end{aligned}
$$

From this we can easily deduce Johnson-Lindenstrauss Lemma by applying the mean lemma.

In this algorithm, every entry in the matrix is (probably) nonzero. The problem of sparsifying the matrix and simplifying the computation of the JL transformation has been a successful area of research. Likewise, in this problem, we tried to preserve the L_2 norm on $\mathbb{R}^J$, but we could have also asked about preserving some other norm on this space. In fact, questions of this type continue to give rise to active research areas.