Lecture 3

1 Overview

In the last lecture we have seen two streaming algorithms for $L_2$ estimation, each using polylogarithmic space: one by Alon, Matias, and Szegedy [1], and the other based on the Johnson-Lindenstrauss dimensionality reduction lemma [4]. Today we will see another algorithm [3] for estimating the $L_2$ norm using polylogarithmic space. Unlike the previous two algorithms, this algorithm generalizes to estimating the $L_p$ norm for any $p \in (0, 2]$. We will also see an algorithm due to [1] for estimating the $L_p$ norm for any $p \geq 2$. The algorithm uses sampling and not sketches, and it does not work in the case where items can be deleted in the stream. To obtain a $(1 \pm \varepsilon)$-approximation with constant probability the space usage is $O\left(\frac{pm^{1-1/p}}{\varepsilon^2}\right)$. The latter bound is not polylogarithmic, but for a good reason: there is a space lower bound of $\Omega\left(\frac{m^{1-2/p}}{p}\right)$ for estimating $L_p$ norm in a streaming model. We will see that lower bound later in the course.

2 A Median Estimator for $L_2$

The first two algorithms worked by choosing a random $k \times m$ matrix $R$ for $k = O\left(\frac{1}{\varepsilon^2}\right)$, computing $Z = [Z_1, \ldots, Z_k] = Rx$, then using $\sum_i Z_i^2 / k$ as the final estimator for $\|x\|_2^2$. The only way the algorithms differed was in their way of choosing $R$: the AMS algorithm chose each row to have entries in $\{-1, +1\}$ in a way that was 4-wise independent, while the Johnson-Lindenstrauss method required full independence and chose each matrix element from the distribution $\mathcal{N}(0, 1)$ (here $\mathcal{N}(\mu, \sigma^2)$ is the normal distribution with expectation $\mu$ and variance $\sigma^2$). The bonus of using the latter approach is that the distribution of random variables $Z_i$ can be completely characterized: they are independently drawn from $\mathcal{N}(0, \|x\|_2^2)$.

The algorithm we now present uses a linear sketch that is identical to that of the Johnson-Lindenstrauss method. However, it uses a different estimator. Consider a random variable $G$ drawn from $\mathcal{N}(0, 1)$. Let $M$ be the median of $|G|$, i.e., a point such that $\Pr_G[|G| \leq M] = 1/2$ (we will refer to the distribution of $|Z_i|$ as “half-Gaussian”). Note that $M$ is an absolute constant. Our estimator is defined as

$$Y = \frac{\text{median}\{|Z_1|, \ldots, |Z_k|\}}{M}$$

where the median of a sequence $\{Z_i\}$ is an element $Z_j$ such that at most $\lceil k/2 \rceil$ $Z_i$’s are less than or greater than it.

A digression: The problem we are actually trying to solve here is a statistical one: given samples $Z_1, \ldots, Z_k$ that are i.i.d. from some normal distribution, estimate the variance of that distribution. One may expect that such a fundamental statistical problem has been solved by the statisticians, and in fact it has. The absolute best estimator for the variance is the maximum likelihood estimator, $\hat{\sigma}^2 = \frac{1}{k} \sum_{i=1}^{k} (Z_i - \bar{Z})^2$ where $\bar{Z} = \frac{1}{k} \sum_{i=1}^{k} Z_i$. Note that the word “median” has been used here to denote two different notions, one applying to sequences of numbers, and another to random variables. To make the distinction more clear, in lecture slides we use different colors for different notions.
MLE, which gives the value of the variance that maximizes the probability of the samples seen. However, using the MLE would only shave constant factors in our analysis; it would also be quite messy to extend it to $L_p$ estimation with $p < 2$.

Back to our estimator. Intuitively, the reason the AMS algorithm worked well is that $\|x\|_2^2$ in expectation, and averaging $k$ copies of the $Z_i^2$ decreased the variance of our estimator to allow for tight concentration bounds. The intuition behind the new median estimator is that for “nice”-looking distributions (like the Gaussian), as you take many samples from the distribution the median of those samples should converge to the median of the distribution as $k \to \infty$.

For the remainder of these notes we will assume without loss of generality that $\|x\|_2^2 = 1$ (since scaling $x$ by some scalar $s$ multiplies the estimator $Y$ by the same factor). Our goal now is to show that $Y$ is sharply concentrated around 1. Also, we assume $k$ is odd, so that the median of $|Z_1| \ldots |Z_k|$ is well-defined.

The proof will be in two steps. First we show closeness in probability to the median: with high probability, the probability mass of the half Gaussian lying between $Y$ and $M$ is at most $\varepsilon$. We then use this to show closeness in value to the median: $|Y - M| \leq \varepsilon$ with high probability.

### Closeness in Probability

**Lemma 1.** Let $U_1, \ldots, U_k$ be i.i.d. real random variables chosen from any distribution having continuous c.d.f $F$ and median $M$. That is, $F(t) = \Pr[U_i < t]$ and $F(M) = 1/2$. Defining $U = \text{median}\{U_1, \ldots, U_k\}$, there is an absolute constant $C > 0$ such that

\[
\Pr[F(U) \in (1/2 - \varepsilon, 1/2 + \varepsilon)] \geq 1 - e^{-Ck\varepsilon^2}
\]

*Proof.* Let $E_i$ be an indicator random variable for $F(U_i) < 1/2 - \varepsilon$ so that $\Pr[E_i = 1] = 1/2 - \varepsilon$. Then $F(U) < 1/2 - \varepsilon$ iff at least $\sum_i E_i \geq k/2$, which happens with probability at most $e^{-Ck\varepsilon^2}$ via the Chernoff bound. Bounding the probability of $F(U) > 1/2 + \varepsilon$ occurring can be handled similarly. \hfill $\square$

### Closeness in Value

**Lemma 2.** Let $F$ be a c.d.f. of a random variable $|G|$, $G$ drawn from $\mathcal{N}(0, 1)$. There exists an absolute constant $C' > 0$ such that if for any $z \geq 0$ we have $F(z) \in (1/2 - \varepsilon, 1/2 + \varepsilon)$, then $z = M \pm C'\varepsilon$.

*Proof.* The proof is specific to the half-Gaussian and boils down to using calculus to lower bound the derivative of $F$ around its median. We omit the details. \hfill $\square$

Overall, we have the following theorem.

**Theorem 3.** If $Y = \text{median}\{|Z_1|, \ldots, |Z_k|\}/M$, then there exist absolute constants $C, C', C'' > 0$ such that

\[
Y = \|x\|_2(M \pm C'\varepsilon)/M = \|x\|_2(1 \pm C''\varepsilon)
\]

with probability at least $1 - e^{-Ck\varepsilon^2}$. 

2
3 Estimating $L_p$ for $p < 2$

The key property of the normal distribution $\mathcal{D}_2 = \mathcal{N}(0, 1)$ that made the median estimator of [3] work is that if $U_1, \ldots, U_k$ and $U'$ are i.i.d. random variables drawn from $\mathcal{D}_2$, then $U = \sum_i x_i U_i$ is distributed as $||x||_2 U'$. We may now wonder whether there are other distributions $\mathcal{D}_p$ where $U_i$'s and $U'$ drawn from $\mathcal{D}_p$ causes $U$ to be distributed as $||x||_p U'$ for $p \neq 2$. Such distributions are called $p$-stable, and it is known that $p$-stable distributions in fact do exist for $p \leq 2$ but not for $p > 2$ [7, 6].

Here we briefly consider the case of $p = 1$. In this case, the following “Cauchy distribution” can be shown to be 1-stable. Its density function is $f(x) = \frac{1}{\pi(1+x^2)}$, and its c.d.f. is $F(z) = 1/2 + \arctan(z)/\pi$. From this information one can verify that the analog of Lemma 2 holds for Cauchy distribution. Moreover, one can generate a random variable from Cauchy distribution by taking $V$ drawn uniformly at random for $[0, 1]$, and computing

$$F^{-1}(V) = \tan[\pi(V - 1/2)].$$

Therefore, there exists an $L_1$ norm estimation algorithm with parameters as in Theorem 3.

There is also a closed form for the p.d.f. and c.d.f. of the 1/2-stable “Lévy distribution”. Unfortunately, for $p \notin \{0, 1/2, 1\}$, there are no closed form expressions for the density functions, or medians of $p$-stable distributions. It is also not clear what is the derivative of the c.d.f. around the median. Nonetheless, it is possible to generate random variables drawn from these distributions and push the $L_2$ arguments through for the general $L_p$'s. Details can be found in papers by Indyk [3] (who uses median estimator) and the recent work of Ping Li [5] (who uses moment estimators).

Final remarks: there are at least two issues that we have ignored:

- The $p$-stable random variables are continuous, but we somehow need to represent them using a bounded number of bits in our computer. This issue can be solved by showing that discretized versions of the distributions suffice, where each random variable requires only a logarithmic number of bits to represent. See [3] for details.

- We assumed that all matrix entries were chosen i.i.d., but we do not have space to store that much randomness. This can be resolved by the general hammer of pseudorandom generators against bounded-space computation, which we will cover later in the course.

4 Estimating $L_p$ for $p \geq 2$

We now cover an algorithm of [1] for $L_p$ norm estimation for $p \geq 2$. Unlike the previous three algorithms for $L_2$ norm estimation, this algorithm uses sampling and not sketching (we will return to sketching next week though). Furthermore, the algorithm only works in the case where items can only be inserted in the stream (the reason will become clear in the proof of Claim 4). In this case, a stream is just a sequence of elements $i_1, \ldots, i_n$, where each $i$ can be interpreted as the update operation $x_i \leftarrow x_i + 1$. For a $(1 \pm \epsilon)$-approximation, the space usage is $O(m^{1-1/p}/\epsilon^2)$. Throughout the rest of these notes, we will use the notation $F_p = \sum_{i=1}^n x_i^p = ||x||_p^p$, this is a notation often used in streaming literature, in particular in the original paper [1].

First, imagine we had two passes over the data stream. In this case we can pick a stream element $i = i_j$ uniformly at random in the first pass (you can assume we know $n$, but in fact this
assumption is not necessary—try to see why as an exercise), and in the second pass we can compute $x_i$ and return $Y = nx_i^{p-1}$. The expectation of $Y$ is

$$E[Y] = \sum_i (x_i/n)nx_i^{p-1} = \sum_i x_i^p = F_p$$

as desired. The variance can be bounded from above by

$$E[Y^2] = \sum_i (x_i/n)n^2x_i^{2p-2} = n\sum_i x_i^{2p-1} = nF_{2p-1}^2.$$ 

To relate this upper bound to $F_p$, we use the following claim.

**Claim 4.** For any $p \geq 1$, $nF_{2p-1} \leq m^{1-1/p}(F_p)^2$.

**Proof.** Observe that $n = \sum_i x_i$, since we only allow unit updates. Now we have

$$nF_{2p-1} = n||x||_2^{2p-1}$$

$$\leq n||x||_{2p-1}^2$$

$$= ||x||_1||x||_2^{2p-1}$$

$$\leq m^{1-1/p}||x||_p||x||_2^{2p-1}$$

$$= m^{1-1/p}||x||_p^{2p}$$

$$= m^{1-1/p}F_p^2$$

The last inequality follows from Hölder’s inequality, which states that $||xy||_1 \leq ||x||_p||y||_p/(p-1)$ for any $p \geq 1$ and real vectors $x, y$ (here vector multiplication is component-wise). We apply Hölder’s inequality with $y = \vec{1}$ followed by division by $m$ on both sides of the inequality.

To make the algorithm one-pass we cannot pick a random stream element and then use a second pass to calculate $x_i$. Instead, we collapse both operations into one pass. As before, we pick $i = i_j$ uniformly at random from the stream (again, this can be done without knowing $n$). Then we compute $r$, which is the number of occurrences of $i$ in the tail of the stream after $i_j$, i.e., in $i_j, \ldots, i_n$ and use $r$ instead of $x_i$. Note that it is always true that $r \leq x_i$. Also, note that the value of $r$ is distributed uniformly over $\{1, \ldots, x_i\}$.

In order to cancel out the discrepancy between $r$ and $x_i$, we use a modified estimator $Y' = n(r^p - (r - 1)^p)$. Then we have

$$E[Y'] = nE[(r^p - (r - 1)^p)]$$

$$= n \cdot 1/n \sum_{i=1}^m \sum_{j=1}^x j^p - (j - 1)^p$$

$$= \sum_{i=1}^m x_i^p$$

The last equality follows by telescoping. To bound the second moment we first relate $Y$ to $Y'$ by observing that

$$Y' = n(r^p - (r - 1)^p) \leq npr^{p-1} \leq pY.$$
Thus $\text{Var}[Y'] \leq E[Y']^2 \leq p^2 E[Y]^2 \leq p^2 m^{1-1/p} F_p^2$. The original paper [1] shows a slightly better bound of $pm^{1-1/p} F_p^2$, but the analysis is more complicated. The simplified proof presented here is due to David Woodruff.

Therefore, if we take $k = Cpm^{1-1/p} F_p^2 / \epsilon^2$ independent estimators $Y'_1 \ldots Y'_k$, for large enough constant $C > 0$, and return $Y'' = 1/k \sum_j Y'_j$, then by Chebyshev bound (as in Lecture 2) we get

$$\Pr[|Y'' - F_p| \geq \epsilon F_p] \leq 1/4$$

Thus, we have a one-pass algorithm for $F_p$ with positive updates using space $O(pm^{1-1/p} / \epsilon^2)$ for $(1 \pm \epsilon)$-approximation.

Final notes:

- The analysis in [1] as is only works only for integer values of $p$; however it can be easily adapted to work for any $p > 1$. The analysis presented here works for non-integral $p$’s as well.

- Sampling approach presented here is a general technique, which has led to streaming algorithms for several other problems, e.g. for estimating entropy of a stream in polylogarithmic space (the entropy of a stream is defined as $\sum_i (x_i/n) \log(x_i/n)$).

References


