Randomized Algorithms 2015A
Lecture 8 – l_p-norm, $p > 2$, of data streams*

Robert Krauthgamer

1 Data streams ℓ_p-norm, $p > 2$

Input: a vector $x \in \mathbb{R}^n$, given as a stream (sequence) of m updates of the form (i, a), meaning $x_i \leftarrow x_i + a$.

Goal: Estimate $\|x\|_p$ for $p > 2$.

We shall employ the approach of a randomized linear sketch $L : \mathbb{R}^n \to \mathbb{R}^s$, hence updates will be easy to implement, and we shall focus on accuracy and space (which is s plus random bits, modulo bit representation).

Theorem 1: For every dimension n, one can estimate the ℓ_p norm, $p > 2$, within constant factor using a linear sketch of $s = cn^{1-2/p} \log n$ memory words. [with high constant probability]

We will see a rather simple algorithm due to Andoni [blog post]. It simplified previous work, including [Andoni, Krauthgamer and Onak, 2011], which achieves a stronger approximation $1 + \varepsilon$. This space requirement is known to be almost optimal (up to ε and logs).

For simplicity, we shall ignore the issue of storing the randomness.

The Algorithm: It is convenient to break it into two steps, the first one scales each entry by a random scalar, the second one reduces the dimension (folds the vector by hashing coordinates).

1. Compute $y \in \mathbb{R}^n$ by

$$y_i = x_i / u_i^{1/p},$$

where each u_i is drawn independently from an exponential distribution (PDF is e^{-u}).

2. Compute $z \in \mathbb{R}^s$ by

$$z_j = \sum_{i:h(i)=j} r_i y_i,$$

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

where $h : [n] \to [s]$ is a random hash function and $r_i \in \{\pm 1\}$ are random signs.

Observe this is indeed a linear sketch.

Estimator: report $\|z\|_\infty$.

Lemma 2: $\Pr\left[\|y\|_\infty \in (0.5\|x\|_p, 2\|x\|_p)\right] \geq 0.75$.

The proof seen in class uses the following stability property: If u_i have exponential distribution and $\lambda_i > 0$ are scalars, then $\min_i\{u_i/\lambda_i\}$ is distributed like u/λ where u has exponential distributed and $\lambda = \sum_i \lambda_i$.

Exer: Prove this property.

The main idea in analyzing z (i.e., the hashing) is that "big" coordinates will fall into distinct buckets, and the rest ("small" coordinates) will become lower order terms.

Lemma 3: Let $M = \|x\|_p$, and let $l \geq 1$. The expected number of "big" coordinates in y is at most l^p, where a coordinate is called big if $|y_i| \geq M/l$.

The proof was seen in class.

We set $l = c \log n$. Then By Markov's inequality, with probability at least 95%, the number of big coordinates is at most $O(\log n)^p$. Moreover, this number is smaller than $\sqrt{s} = O(n^{1/2-1/p})$, and thus this will go to distinct buckets (birthday paradox).

Lemma 4: Let S denote the small coordinates in y. For (bucket) $j \in [s]$, define

$$z'_j = \sum_{i \in S : h(i) = j} r_i y_i .$$

Then $\mathbb{E}[{z'_j}^2] = O(M^2/(c \log n))$.

The proof was seen in class, using the following inequality. For every $x \in \mathbb{R}^n$ and $p > q \geq 1$,

$$\|x\|_q \leq n^{1/q-1/p}\|x\|_p.$$

Exer: prove it using Holder's inequality between $|x_i|^q$ and the all-ones vector, with norm $r = p/q > 1$.

By Markov's inequality, it follows that with high constant probability, $|z'_j| \leq M/\sqrt{\log n} = o(M)$ (and hence would not affect $|y_i| > M/2$ if that lands in this bucket).

But since we have many buckets j, hence we need higher success probability (not just constant for every bucket), and indeed we use the following generalization of the Chernoff-Hoeffding bound.

Observe that by Markov's inequality, with at least 95% probability, $\|y\|_2^2 \leq O(n^{1-2/p}\|x\|_p^2)$. Let us condition on the u_i's (i.e., the value of y is determined); we focus now on the case where both this event and the one in Lemma 2 occur.

Bernstein's inequality: Let $X_1, \ldots, X_n$ be indepedent random variables, where each $X_i \in [-B, B]$ and has expectation $\mathbb{E}X_i = 0$. Then

$$\forall t > 0, \qquad \Pr[\sum_i X_i > t] \leq e^{-\frac{t^2/2}{V + tB/3}},$$

where $V = \mathrm{Var}(\sum_i X_i) = \sum_i \mathbb{E}[X_i^2]$ is the variance of their sum.

Lemma 5: For each bucket $j \in [s]$, with at least $1 - 2/n^2$ probability, $|z'_j| \leq M/4$.

The proof was seen in class, using Bernstein's inequality.

Theorem 1 follows by a union bound over all the above events, which yields overall success probability $\geq 0.75 - 0.05 - o(1) - 0.05 - 2s/n \geq 0.6$.

Remark about randomness: This analysis assumes full independence, because of Bernstein's inequality. It is possible to avoid it, but it requires some workaround.

Remark about approximation: It is possible to achieve $1 + \varepsilon$ approximation by repeating the estimator $1/\varepsilon^{O(1)}$ times and taking the median of the results.

2 Dimension Reduction in ℓ_2

The Johnson-Lindenstrauss (JL) Lemma: Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and fix $\varepsilon > 0$. Then there exist $y_1, \ldots, y_n \in \mathbb{R}^k$, $k = O(\varepsilon^{-2} \log n)$, such that

$$\forall i, j \in [n], \qquad \|y_i - y_j\| \in (1 \pm \varepsilon)\|x_i - x_j\|.$$

Moreover, there is a randomized linear mapping $L : \mathbb{R}^d \to \mathbb{R}^k$ (oblivious to the given points), such that if we define $y_i = Lx_i$, then with probability at least $1 - 1/n$ all the above inequalities hold.

We started seeing the proof in class, and will finish it next week.