1 \( p \)-norm Estimation

Recall the sketching matrices \( P \cdot D \), where \( P \) consists of a CountSketch matrix, and \( D \) consists of a diagonal matrix with diagonal elements \( 1/E_i^{1/p} \), with \( E_i \) being independent standard exponential random variables.

For arbitrary \( y \), \(||Dy||_\infty|\) looks like

\[
| |Dy||_\infty = \max_i \frac{||y_i||_p^p}{E_i} = \frac{1}{\min_i E_i^{1/p}} = \frac{||y_i||_p^p}{E_i}
\]

and the probability of a reasonable value of \( E \) is \( \Pr \left[ E \in [1/10, 10] \right] = (1 - e^{-10}) - (1 - e^{-1/10}) > 4/5 \) (this actually evaluates to just over 9/10).

As such, \(||Dy||_\infty|\) is a good estimate for \(||y||_p|\), but \( Dy \in \mathbb{R}^n \) is a large vector, so sketching using matrix \( P \in \mathbb{R}^{s \times n} \) is needed to reduce computation cost.

Intuitively, \( P \) is hashing coordinates of \( Dy \) into buckets and taking a signed sum; most items cancel out and then \(| |PDy||_\infty| \simeq | |Dy||_\infty| \). It is known previously that \( P \) is composed of hash functions \( h : [n] \rightarrow [s] \) and \( \sigma : [n] \rightarrow \{-1, 1\} \) (assuming they are truly random). Given that \(| |Dy||_\infty/||y||_p| \in [1/10^{1/p}, 10^{1/p}] \) with probability > 4/5, to achieve \(| |PDy||_\infty| \simeq | |Dy||_\infty| \) with good probability, it is necessary to have:

1. in each bucket \( i \) not containing the maximum value, \(| (PDy)_i | \leq ||y||_p/100 \)
2. in each bucket \( i \) containing the maximum value, \(| |(PDy)_i| - | |Dy||_\infty| \leq ||y||_p/100 \)

Let \( \delta(\text{event}) = 1 \) if a given event holds and \( \delta(\text{event}) = 0 \) otherwise. It is then possible to define a given element of \( PDy \) as \( (PDy)_i = \sum_j \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j \). Due to \( \sigma \), its expectation is \( \mathbb{E}[(PDy)_i] = 0 \). The evaluation of its variance as follows:
\[ \mathbb{E}_P[(PDy)^2_j] = \sum_{j,j'} \mathbb{E}[\delta(h(j) = i) \cdot \delta(h(j') = i) \cdot \sigma_j \cdot \sigma_{j'} | (Dy)_j(Dy)_{j'} = \frac{1}{s}||Dy||_2^2] \] (2)

\[ \mathbb{E}_D[||Dy||_2^2] = \sum_i y^2_i \mathbb{E}[D^2_{i,i}] \] (3)

\[ \mathbb{E}[D^2_{i,i}] = \int_0^\infty t^\frac{2}{p} e^{-t} dt = \int_0^1 t^\frac{2}{p} e^{-t} dt + \int_1^\infty t^\frac{2}{p} e^{-t} dt \]
\[ \leq \int_0^1 t^\frac{2}{p} dt + \int_1^\infty e^{-t} dt = \left(\frac{1}{1 - \frac{2}{p}}\right) t^\frac{2}{p} |1_0 - e^{-t}|_1 = O(1) \] (5)

\[ \mathbb{E}_P[(PDy)^2_j] = O\left(\frac{1}{s}\right) ||y||_2^2 = O\left(\frac{1}{s}\right) \left(n^{-\frac{2}{p}}||y||_p^2\right). \] (6)

The last line holds due to Hölder’s Inequality (\(||y||_2^2 = \sum_i y^2_i \leq (\sum_i^n y^2_i)^{p/2}/p(\sum_i^n 1^n)^{1/q} = ||y||_p^2 \cdot n^{1-2/p}\)).

**Definition (Bernstein’s Bound).** Suppose independent random variables \(R_1, \ldots, R_n\), and for all \(j\), \(|R_j| \leq K\), and \(\text{Var}[\sum_j R_j] = \sigma^2\). Then, there exists constants \(c, C\) such that for all \(t > 0\),

\[ \Pr\left[ \left| \sum_j R_j - \mathbb{E}\left[\sum_j R_j\right] \right| > t \right] \leq C \left( e^{-\frac{t^2}{2\sigma^2}} + e^{-\frac{ct}{K}} \right) \] (7)

In order to get \(1/poly(n)\) error probability, set \(R_j = \delta(h(j) = i)\sigma_j \cdot (Dy)_j\), \(t = ||y||_p/100\), and \(s = \Theta(n^{-1-2/p} \log n)\) to handle all parameters required for Bernstein’s bound other than \(K\).

It is possible to treat large \(R_j\) separately, where \(R_j > \frac{\alpha||y||_p}{\log n}\) for a sufficiently small \(\alpha > 0\). If \(|R_j| > \frac{\alpha||y||_p}{\log n}\), then necessarily \((Dy)_j \geq \frac{\alpha||y||_p}{\log n}\) (define \(j\) as “large” if this is the case, “small” otherwise). Then, any \(j\) may be large with probability and expectation

\[ \Pr[j \text{ is large}] = \Pr\left[ \frac{|y_j|}{E_j^{1/p}} \leq \frac{\alpha||y||_p}{\log n} \right] = \Pr\left[ \frac{|y_j|^p}{\alpha^p||y||_p^p \log n} \leq E_j \right] \] (8)
\[ = 1 - e^{-\frac{|y_j|^p \log n}{\alpha^p||y||_p^p}} \leq \frac{|y_j|^p \log n}{\alpha^p||y||_p^p} \] (9)

\[ \mathbb{E}[R_j \text{ for large } j] \leq \sum_j \frac{|y_j|^p \log n}{\alpha^p||y||_p^p} = \frac{\log n}{\alpha^p} \] (10)

There are \(s = O(n^{-1-2/p} \log n)\) buckets and \(\log n\) items. By Markov bound, there are \(O(\log n)\) large \(j\) with constant probability. \(D\) is conditioned on the above as well as \(||Dy||_\infty \in [||y||_p/10^1/p, ||y||_p/10^1/p]\) (which happens with probability > 4/5). All the large \(j\) should then be perfectly hashed into separate buckets by \(P\). (If there are \(b\) balls and \(Cb\) bins, \(\Pr[\text{collision}] \leq \left(\frac{b}{2}\right)^{1/Cb} \leq 1/2C\) )

Bernstein’s bound can then be applied separately for the small indices \(j\) for each hash bucket. \(\mathbb{E}[(PDy)_i] = 0\) for each hash bucket \(i\), and \(\mathbb{E}[(PDy)^2_i] = O(1/s)\left(n^{-1-2/p}||y||_p^2\right)\). Assuming \(K = \max_j |R_j| \leq \alpha||y||_p/\log n\) for small \(j\) in a bucket (it can be shown that \(\text{Var}[R_j] = O(1/s)\left(n^{-1-2/p}||y||_p^2\right)\)
even if no $j$ is large. Setting $t = \|y\|_p/100$ and $s = \Theta(n^{1-2/p}\log p)$ in Bernstein’s bound, for a bucket $(PDy)_i$

$$\Pr \left[ \left\| \sum_{\text{small } j} \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_i \right\|_p > \frac{\|y\|_p}{100} \right] \leq C \left( e^{-\Theta(\log n)} + e^{-\frac{\log n}{100}} \right) \leq \frac{1}{n^2} \quad (11)$$

By union bound over all $s$ buckets, the signed sum of all small $j$ in every bucket will be at most $\|y\|_p/100$. Therefore, for all $i$,

1. in each bucket $i$ without large indices $j$, $|(PDy)_i| \leq \|y\|_p/100$
2. in each bucket $i$ with one large index $j$, $|(PDy)_i| = |\sigma_j(Dy)_j| \pm \|y\|_p/100$

and no bucket has more than one large $j$ as shown in the perfect hashing assumption above. Conditioning on $\|Dy\|_\infty \in [\|y\|_p/10^{1/p}, \|y\|_p \cdot 10^{1/p}]$,

$$\frac{\|y\|_p}{10^p} - \frac{\|y\|_p}{100} \leq \|PDy\|_\infty \leq 10^{\frac{1}{2}} \cdot \|y\|_p + \frac{\|y\|_p}{100} \quad (12)$$

Therefore, it is reasonable to use $\|PDy\|_\infty$ as an estimate for $\|y\|_p$. The total space used is $s = O(n^{1-2/p}\log n)$, i.e. $O(n^{1-2/p}\log^2 n)$ bits. This space complexity still holds even when considering the pseudorandom generation of matrix $P$, see [1].

2 Heavy Hitters

$l_1$ guarantee: output a set containing all items $j$ for which $|x_j| \geq \phi \|x\|_1$, and the set should not contain any $j$ with $|x_j| \leq (\phi - \varepsilon)\|x\|_1$.

$l_2$ guarantee: output a set containing all items $j$ for which $x_j^2 \geq \phi \|x\|_2^2$, and the set should not contain any $j$ with $x_j^2 \leq (\phi - \varepsilon)\|x\|_2^2$. This guarantee is much stronger: suppose $x = [\sqrt{n}, 1, \ldots, 1]$, $\sqrt{n}$ is an $l_2$-heavy hitter for constant $\phi$ and $\varepsilon$, but not an $l_1$-heavy hitter. Also, if $|x_j| \geq \phi \|x\|_1$, it means that $x_j^2 \geq \phi^2 \|x\|_1^2 \geq \phi^2 \|x\|_2^2$ as well.

References