CS 15-851: Algorithms for Big Data Spring 2025

Lecture 8 Part 2 — 03/13

Prof. David Woodruff *Scribe: Esther Cao*

1 p-Norm estimator for $p > 2$

For $p > 2$, p-stable distributions don't exist. We'll prove that $\Omega(n^{1-2/p})$ bits of space are needed to approximate p-norms for $p > 2$ to a constant factor, with constant probability. This will be achieved using exponential random variables. Define ϵ to be the constant approximation parameter.

Our sketch will be defined as $P \cdot D$:

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 1 & 0 & \cdots \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & -1 & 1 & 0 & -1 & \cdots \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 & \cdots \end{bmatrix} \begin{bmatrix} 1/E_1^{1/p} & 0 & 0 & 0 & \cdots \\ 0 & 1/E_2^{1/p} & 0 & 0 & \cdots \\ 0 & 0 & 1/E_3^{1/p} & 0 & \cdots \\ \cdots & & & & \\ 0 & 0 & 0 & \cdots & 1/E_n^{1/p} \end{bmatrix}$$

Where P is a CountSketch matrix, and D is a diagonal matrix with entries $1/E_i^{1/p}$, where E_i is an exponential random variable. Note that P, D are linear maps which don't depend on x.

When right-multiplied by a vector x, this sketching matrix first scales the entries of x by the entries of D, then applies CountSketch to the output.

1.1 Stability of Exponential Random Variables

An exponential random variable E with parameter λ is defined as:

- PDF: $f(x) = \begin{cases} \lambda e^{-\lambda x} \text{ if } x \geq 0 \\ 0 \text{ otherwise} \end{cases}$
- CDF: $F(x) = 1 - e{-}\lambda x$ if $x \geq 0$
- $t \cdot E$ for scalar $t \geq 0$ has CDF $F(x) = 1 - e^{-\lambda/t \cdot x}$

Lemma 1. *Exponential random variables are min-stable; the minimum of exponential random variables is an exponential r.v.*

Proof. Consider independent exponential random variables $E_1, \cdots, E_n$, and scalars $|y_1|, \cdots, |y_n|$, and let $q = \min(\frac{E_1}{|y_1|p}, \cdots, \frac{E_n}{|y_n|^p})$.

$$\begin{aligned} F_q(x) = Pr[q > x] &= Pr\left[\forall i, \frac{E_i}{|y_i|^p} \geq x\right] \\ &= Prod_i Pr\left[\frac{E_i}{|y_i|^p} \geq x\right] && (E_i\text{'s are independent}) \\ &= Prod_i e^{-x|y_i|^p} = e^{-x|y|_p^p} \end{aligned}$$

So q is also an exponential random variable with parameter $\lambda = |y|_p^p$. Equivalently, $q = (1/|y|_p^p)E$ for a standard exponential random variable E.

Therefore, the p-norm information of y is preserved by taking the minimum of exponentials. ■

1.2 Analyzing $|Dy|_\infty$

Return to our sketch, $P \cdot D$ defined previously.

Theorem 1. *With constant probability, $|Dy|_\infty$ is a constant-factor approximation of $|y|_p$. Specifically with probability $> 4/5$:*

$$|Dy|_\infty \in [|y|_p/10^{1/p}, 10^{1/p}|y|_p]$$

Proof. Consider the value of $|Dy|_\infty = \max_i(|y_i|/E_i^{1/p})$, the maximum value in the vector Dy.

$$\begin{aligned} |Dy|_\infty^p &= \max_i \left(\frac{|y_i|^p}{E_i}\right) \\ &= \frac{1}{\min_i \left(\frac{E_i}{|y_i|^p}\right)} \\ &= \frac{1}{E/|y|_p^p} && (\text{Min-stability of exponentials; } E \sim Exp(1)) \\ &= \frac{|y|_p^p}{E} \end{aligned}$$

Taking the p-root of both sides, $|Dy|_\infty = \frac{|y|_p}{E^{1/p}}$.

We can bound E, with constant probability:

$$\begin{aligned} Pr[E \in [1/10, 10]] &= (1 - e^{-10} - (1 - e^{-1/10}) && (\text{CDF of exponential}) \\ &= e^{-1/10} - e^{-10} > 4/5 \end{aligned}$$

Therefore, the maximum entry $|Dy|_\infty$ is a constant-factor approximation for the target value of $|y|_p^p$, with constant probability; i.e. this embedded the p-norm into the infinity-norm. ■

We have shown that, with probability at least 4/5, $|Dy|_\infty$ is a constant-factor estimate of $|y|_p$.

- It suffices to approximate the maximum entry of $|Dy|$.
- However, Dy is an n-dimensional vector, which is too expensive to store. Thus we need to do dimensionality reduction, through the CountSketch matrix P.
- We hope to approximate $|Dy|_\infty$ with $|PDy|_\infty$.

1.3 Analyzing $|PDy|_\infty$

Recall that P is defined as a CountSketch matrix. Let s be the number of rows of P, which we think of as hash buckets. Intuitively, P hashes the coordinates of Dy into s buckets and takes a signed sum of the entries in each bucket. We expect that the large entries of Dy stand out, while the small values cancel out, yielding $|PDy|_\infty \approx |Dy|_\infty$.

P is fully specified by the following functions:

- Hash function $h : [n] \to [s]$
- Sign function $\sigma : [n] \to \{-1, 1\}$

For simplicity, we assume h, σ are truly random (rather than 2-wise, 4-wise independent, respectively), though they can be derandomized.

Theorem 2. *$|PDy|_\infty \approx |Dy|_\infty$ with good probability.*

This consists of two sub-parts. Let j be the coordinate of the max entry of Dy, namely $|(Dy)_j| = |Dy|_\infty$.

Claim 1. The buckets not containing the maximum entry each have small sum: in each bucket i not containing the element j, we have $|(PDy)_i| \le |y|_p/100$.

Claim 2. The bucket containing the maximum entry is close to the maximum value of $|Dy|$: $||(PDy)_i| - |Dy|_\infty| \le |y|_p/100$.

Let $\delta(E)$ be the indicator random variable for event E: $\begin{cases} \delta(E) = 1 \text{ if E holds} \\ \delta(E) = 0 \text{ otherwise} \end{cases}$

The ith bucket value $(PDy)_i$ sums over all coordinates in the vector which hash to bucket i, multiplied by a random sign. This has the form

$$(PDy)_i = \sum_j \delta(h(j) = i) \cdot \sigma_j (Dy)_j$$

To prove concentration bounds on $(PDy)_i$, we compute its expectation and variance, which depend on both the randomness in P and D.

Begin by considering the expectation and variance over the CountSketch matrix P:

- $E_P[(PDy)_i] = 0$, since σ_j is equally likely to be ± 1.
- $Var_P[(PDy)_i^2] = E_P[(PDy)_i^2] = \sum_{j,j'} E_P[\delta(h(j) = i)\delta(h(j') = i)\sigma_j\sigma_{j'}](Dy)_j(Dy)_{j'} = (1/s)|Dy|_2^2$. The last step follows because when $j \neq j'$, the term is 0 by independence. When $j = j'$, $\sigma_j\sigma_{j'} = 1$, so this simplifies to $\sum_j E_P[\delta(h(j) = i)(Dy)_j^2] = (1/s)|Dy|_2^2$.

Note that D is also a random variable, so we next compute the expectation over D:

$$E_D[|Dy|_2^2] = \sum_i y_i^2 \cdot E[D_{i,i}^2] \qquad (D \text{ is a diagonal matrix})$$

Recall that $D_{i,i}$ is defined to be $E_i^{1/p}$, where $E_i \sim Exp(1)$. To compute $E[D_{i,i}^2]$, we integrate over the PDF of the exponential distribution.

$$\begin{aligned}
E[D_{i,i}^2] = E_i^{-2/p} &= \int_{t\geq 0} t^{-2/p}e^{-t}dt \\
&= \int_{t\in[0,1]} t^{-2/p}e^{-t}dt + \int_{t>1} t^{-2/p}e^{-t}dt \\
&\leq \int_{t\in[0,1]} t^{-2/p}dt + \int_{t>1} e^{-t}dt \\
&= \frac{1}{(1-2/p)t^{1-2/p}}\Big|_0^1 - e^{-t}\Big|_1^\infty \\
&\in O(1)
\end{aligned}$$

So far, we have computed the variance of $(PDy)_i^2$ by first taking the expectation over P, then over D, to obtain $E[(PDy)_i^2] = O(1/s)|y|_2^2$.

Next, we need to relate the 2-norm to the p-norm. We can apply Holder's inequality, which is a generalization of Cauchy-Schwarz.

Fact 1. Holder's inequality. If $1/p + 1/q = 1$, then $\langle x, y\rangle \leq |x|_p|y|_q$.

Note that norms generally get smaller as the dimension increases: $|y|_1 \geq |y|_2 \geq |y|_\infty$.

Lemma 2. $|y|_2^2 = O(n^{1-2/p}|y|_p^2)$.

Proof. The second step below follows from applying Holder's inequality with $p/2$-norm and q-norm, subject to $2/p + 1/q = 1$.

$$\begin{aligned}
|y|_2^2 &= \sum_{i=1}^n y_i^2 \cdot 1 \\
&\leq \left(\sum_{i=1}^n (y_i^2)^{p/2}\right)^{2/p} \cdot \left(\sum_{i=1}^n 1^q\right)^{1/q} \\
&= \left(\sum_{i=1}^n y_i^p\right)^{2/p} \cdot \left(\sum_{i=1}^n 1^q\right)^{1/q} \\
&\leq |y|_p^2 \cdot n^{1/q} \\
&\leq |y|_p^2 \cdot n^{1-2/p}
\end{aligned}$$

■

Plugging back into the original expression:

$$\begin{aligned}
E[(PDy)_i^2] &= O(1/s)|y|_2^2 && \text{(Expectation over } P) \\
&= O(1/s)(n^{1-2/p}|y|_p^2) && \text{(Expectation over } D)
\end{aligned}$$

To recap, we've now shown $E[(PDy)_i] = 0$ for each hash bucket i, and $E[(PDy)_i^2] = O(1/s)(n^{1-2/p}|y|_p^2)$.

The $n^{1-2/p}$ term in the streaming algorithm bound arises from the norm used in Holder's theorem. The number of buckets we choose should cancel out this term. We have s buckets, $(PDy)_1, \cdots, (PDy)_s$, and hope the bucket containing the maximum entry stands out compared to the other buckets.

Now that we have obtained the expectation and variance, a strong tail bound can be applied.

Fact 2. Bernstein's bound. Suppose $R_1, \cdots, R_n$ are independent, and forall j, $|R_j| \leq K$, and $Var[\sum_j R_j] = \sigma^2$. There are constants C, c such that for all $t > 0$,

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

Note that this error bound drops off exponentially with as t increases.

Recall that $(PDy)_i = \sum_j \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$, the sum of entries which hash to the ith bucket.

As a first attempt, define the following towards applying Bernstein's bound.

- $R_j = \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$. Since we assumed for simplicity that h, σ are truly random, it follows that the summands $R_i's$ are independent from each other.
- $t = |y|_p/100$
- $s = \Theta(n^{1-2/p} \log n)$.
- Note that $\sigma^2 = Var[(PDy)_i] = E[(PDy)_i^2] = O(1/sn^{1-2/p|y|_p^2})$. The above definition of $s = n^{1-2/p} \log n$ yields $\sigma^2 = O(|y|_p^2/\log n)$.

Attempt 1. Applying Bernstein's bound:

$$Pr\left[|(PDy)_i - 0| > |y|_p/100\right] \leq C(e^{-\frac{c(|y|_p/100)^2}{|y|_p^2/\log n}} + e^{-ct/K})$$

The second term still relies on K, an upper bound on the values of R_i. Since $R_i \leq |Dy|_\infty \in \Theta(|y|_p)$, the second term simplifies to a constant. However, a constant success probability doesn't suffice, a union bound would subsequently be necessary over the n buckets.

Note that the setup is not correct for all buckets. One of the buckets will contain the largest entry, for which it's not true that $|(PDy)_i| \leq |y|_p/100$ with good probability. So we need to separately consider the large buckets.

1.4 Understanding the large elements

We will separately handle R_j values which are large (for which large is defined by $|R_j| > \alpha|y|_p/\log n$, and α is a sufficiently small constant parameter). This cutoff value is defined to restrict K in the Bernstein bound to be small. Importantly, whether an element is large or not depends only on D, not on the CountSketch matrix P.

Recall that $R_j = \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$. Thus if $|R_j| > \alpha|y|_p/\log n$, then necessarily $|(Dy)_j| > \alpha|y|_p/\log n$.

Next, we show that there are not too many large buckets.

Theorem 3. *With constant probability the number of large elements is $O(\log^p n)$.*

Proof. Recall that $(Dy)_j = |y_j|/E_j^{1/p}$.

$$\begin{aligned} Pr_D[|(Dy)_j| \text{ is large}] &= Pr_D[|y_j|/E_j^{1/p} \geq \alpha|y|_p/\log n] \\ &= Pr_D[E_j \leq |y_j|^p(\log^p n)/(\alpha^p|y|_p^p)] \\ &= 1 - e^{|y_j|^p(\log^p n)/(\alpha^p|y|_p^p)} && \text{(CDF of exponential distribution)} \\ &\leq |y_j|^p(\log^p n)/(\alpha^p|y|_p^p) && (1 - x \leq e^{-x}) \end{aligned}$$

By linearity of expectation, the expected number of large buckets j is $\sum_j |y_j|^p(\log^p n)/(\alpha^p|y|_p^p) = |y|_p^p(\log^p n)/(\alpha^p|y|_p^p) = O(\log^p(n))$. By a Markov bound, with constant probability the number of large elements is $O(\log^p n)$. ■

Condition on the following properties of D:

- $|Dy|_\infty$ is close to the true value. By Theorem 1, $|Dy|_\infty \in [|y|_p/10^{1/p}, 10^{1/p}|y|_p]$ with probability $> 4/5$.
- There are $O(\log^p n)$ large elements. Note that the R_i's remain independent when conditioning on this event (once again, because the definition of large element depends only on D and not on P).

Condition on the following properties of D:

- All large items are perfectly hashed. This occurs with constant probability.

  Perform a balls and bins analysis: we are throwing $O(\log^p n)$ balls into $s \geq n^{1-2/p}$ bins, so

$$Pr[\exists \text{ large } j, j', j \neq j' \text{ which hash to same bucket}] \leq \binom{\lg^p n}{2}\frac{1}{s} \ll 1/100$$

Theorem 4. *With good probability, the sum of the small terms in each bucket is $|y|_p/100$.*

Finally, apply Bernstein's inequality on the small indices j within each hash bucket, so we can assume $K = \max_j |R_j| \leq \alpha|y|_p/\log n$. Plugging this into the bound obtained in Attempt 1:

$$Pr\left[|(PDy)_i| > |y|_p/100\right] \leq C(e^{-\frac{c(|y|_p/100)^2}{|y|_p^2/\log n}} + e^{-c\log n/(100\alpha)}) \leq 1/n^2$$

Therefore, by a union bound over all s buckets, the signed sum of small j in every bucket is at most $|y|_p/100$.

1.5 Wrapping up

For all i:

- $|(PDy)_i| \leq |y|_p/100$ if there are no large indices in the ith bucket.

- $|(PDy)_i| \le \sigma(Dy)_j| \pm |y|_p/100$ if there is exactly one large index j in the ith bucket.
- No bucket contains more than 1 large index j.

We conditioned on $|Dy|_\infty \in \left[\frac{|y|_p}{10^{1/p}}, 10^{1/p}|y|_p\right]$.

Also, $|PDy|_\infty$ is close to $|Dy|_\infty$, since the noise (total contribution of small terms) in each bucket is at most $|y|_p/100$ with good probability.

So we can output $|PDy|_\infty$ as your estimate for $|y|_p$.

The total space complexity is $s = O(n^{1-2/p} \log n)$ words, which is $O(n^{1-2/p} \log^2 n)$ bits.