

Outline

- Quick recap of ℓ_1 -regression, and how to speed it up
- Introduction to the Streaming Model
- Estimating Norms in the Streaming Model

Example Problem: Norms

- Suppose you want $\|x\|_p^p = \sum_{i=1}^n |x_i|^p$
- Want Z for which $(1-\epsilon) \|x\|_p^p \leq Z \leq (1+\epsilon) \|x\|_p^p$ with probability $> 9/10$
- $p = 1$ corresponds to total variation distance between distributions
- $p = 2$ useful for geometric and linear algebraic problems
- $p = \infty$ is the value of the maximum entry, useful for anomaly detection, etc.

Example Problem: Euclidean Norm

- Want Z for which $(1-\epsilon) \|x\|_2^2 \leq Z \leq (1+\epsilon) \|x\|_2^2$
- Sample a random CountSketch matrix S with $1/\epsilon^2$ rows
- Can store S efficiently using limited independence
- If $x_i \leftarrow x_i + \Delta_j$ in the stream, then $Sx \leftarrow Sx + \Delta_j S_{*,i}$
- At end of stream, output $\|Sx\|_2^2$
- With probability at least $9/10$, $\|Sx\|_2^2 = (1 \pm \epsilon) \|x\|_2^2$
- Space complexity is $1/\epsilon^2$ words, each word is $O(\log n)$ bits

Example Problem: 1-Norm

- Want Z for which $(1-\epsilon) \|x\|_1 \leq Z \leq (1+\epsilon) \|x\|_1$
- Sample a random Cauchy matrix S ?
- Can store S with $\frac{1}{\epsilon}$ words of space [Kane, Nelson, W]
- If $x_i \leftarrow x_i + \Delta_j$ in the stream, then $Sx \leftarrow Sx + \Delta_j S_{*,i}$
- Space complexity is $1/\epsilon^2$ words, each word is $O(\log n)$ bits
- At end of stream, output $\|Sx\|_1$?
- *Cauchy random variables have no concentration...*

1-Norm Estimator

- Probability density function $f(x)$ of $|C|$ for a Cauchy random variable C is $f(x) = \frac{2}{\pi(1+x^2)}$

- Cumulative distribution function $F(z)$:

$$F(z) = \int_0^z f(x)dx = \frac{2}{\pi} \arctan(z)$$

- Since $\tan(\pi/4) = 1$, $F(1) = 1/2$, so $\text{median}(|C|) = 1$
- If you take $r = \frac{\log(1/\delta)}{\epsilon^2}$ independent samples X_1, \dots, X_r from F , and $X = \text{median}_i X_i$, then $F(X)$ in $[1/2 - \epsilon, 1/2 + \epsilon]$ with probability $1 - \delta$
- $F^{-1}(X) = \tan\left(\frac{X\pi}{2}\right) \in [1 - 4\epsilon, 1 + 4\epsilon]$

p-Norm Estimator

- Can achieve $1/\epsilon^2$ words of space for p-norm estimation for any $0 < p < 2$
- Proof is similar to 1-norm estimation, and uses p-stable distributions, which exist only for $0 < p < 2$
- No closed form expression for their probability density function but they are efficiently sampleable:

- If $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ and $r \in [0,1]$ are uniformly random, then

$$\frac{\sin(p \theta)}{\cos^{\frac{1}{p}} \theta} \left(\frac{\cos(\theta(1-p))}{\ln(\frac{1}{r})} \right)^{\frac{1-p}{p}} \text{ is a sample from a p-stable distribution!}$$

- Can discretize them and store a sketching matrix of samples from the p-stable distribution using limited independence

p-Norm Estimator for $p > 2$

- For $p > 2$, p-stable distributions do not exist!
- We will see later that $\Omega(n^{1-\frac{2}{p}})$ bits of space needed to approximate p-norms, $p > 2$, up to a constant factor with constant probability
- To achieve an $\widetilde{O}(n^{1-2/p})$ bits of space algorithm, we will use exponential random variables. We will focus on constant approximation parameter ϵ
- Our sketch will be $P \cdot D$:

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

Stability of Exponential Random Variables

- Exponential random variable E with parameter λ
 - (PDF) probability density function: $f(x) = \lambda e^{-\lambda x}$ if $x \geq 0$, and 0 otherwise
 - (CDF) cumulative density function: $F(x) = 1 - e^{-\lambda x}$ for $x \geq 0$
 - $t \cdot E$ for scalar $t \geq 0$ has CDF $F(x) = 1 - e^{-\frac{\lambda}{t}x}$
- Stability: consider independent exponential random variables E_1, \dots, E_n and scalars $|y_1|, \dots, |y_n|$, let $q = \min\left(\frac{E_1}{|y_1|^p}, \dots, \frac{E_n}{|y_n|^p}\right)$
- $\Pr[q > x] = \Pr\left[\forall i, \frac{E_i}{|y_i|^p} \geq x\right] = \prod_i e^{-x|y_i|^p} = e^{-x|y|^p}$
- So q is an exponential random variable with $\lambda = |y|^p$, that is,
 $q \equiv \left(\frac{1}{|y|^p}\right) E$ for a standard exponential random variable E

Stability of Exponential Random Variables

- Recall our sketch $P^*D =$

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

- What does $|Dy|_\infty$ look like for an arbitrary y ?

- $|Dy|_\infty^p = \max_i \left(\frac{|y_i|^p}{E_i} \right) = \frac{1}{\min_i \frac{E_i}{|y_i|^p}} \equiv \frac{1}{E \cdot \frac{1}{|y|^p}} = \frac{|y|^p}{E}$

- $\Pr[E \in \left[\frac{1}{10}, 10 \right]] = (1 - e^{-1}) - (1 - e^{-\frac{1}{10}}) = e^{-\frac{1}{10}} - e^{-10} > \frac{4}{5}$

Stability of Exponential Random Variables

- We know $|Dy|_\infty \in [\frac{|y|_p}{10^{1/p}}, 10^{1/p}|y|_p]$ with probability at least $\frac{4}{5}$
- So $|Dy|_\infty$ is a good estimate of $|y|_p$, but Dy is an n -dimensional vector!

- Recall our sketch $P^*D =$

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

- What can we say about $|PDy|_\infty$ if P has s rows?
- Intuitively P is hashing coordinates of Dy into buckets and taking a signed sum of the entries. Expect everything to cancel out and $|PDy|_\infty \approx |Dy|_\infty$ 😊

Understanding $|PDy|_\infty$

- Let s be the number of rows of P , which we can think of as hash buckets
- P is a CountSketch matrix with hash functions h and σ
 - $h: [n] \rightarrow [s]$
 - $\sigma: [n] \rightarrow \{-1,1\}$
 - Let's assume h and σ are truly random (can be derandomized)
- We know $|Dy|_\infty \in [\frac{|y|_p}{10^{1/p}}, 10^{1/p}|y|_p]$ with probability at least $4/5$
- To achieve $|PDy|_\infty \approx |Dy|_\infty$ with good probability, we want
 - (1) in each bucket i not containing the coordinate j for which $|(Dy)_j| = |Dy|_\infty$, we have $|(PDy)_i| \leq \frac{|y|_p}{100}$
 - (2) in the bucket i containing the coordinate j for which $|(Dy)_j| = |Dy|_\infty$, we have $|| (PDy)_i | - |Dy|_\infty | \leq |y|_p/100$

Analyzing $|\text{PDy}|_\infty$

- Let $\delta(E) = 1$ if event E holds, and $\delta(E) = 0$ otherwise
- What does the i -th bucket value $(\text{PDy})_i$ look like?
- $(\text{PDy})_i = \sum_j \delta(h(j) = i) \sigma_j(\text{Dy})_j$
- $E[(\text{PDy})_i] = 0$
- What about the variance of $(\text{PDy})_i$?

Understanding $|\text{PDy}|_\infty$

- $E_P[(\text{PDy})_i^2] = \sum_{j,j'} E[\delta(h(j) = i)\delta(h(j') = i)\sigma_j\sigma_{j'}](\text{Dy})_j(\text{Dy})_{j'} = \left(\frac{1}{s}\right) |\text{Dy}|_2^2$
- $E_D[|\text{Dy}|_2^2] = \sum_i y_i^2 \cdot E[D_{i,i}^2]$
- $$\begin{aligned} E[D_{i,i}^2] &= \int_{t \geq 0} t^{-2/p} e^{-t} dt \\ &= \int_{t \in [0,1]} t^{-2/p} e^{-t} dt + \int_{t > 1} t^{-2/p} e^{-t} dt \\ &\leq \int_{t \in [0,1]} t^{-2/p} dt + \int_{t > 1} e^{-t} dt \\ &= \left(\frac{1}{1-\frac{2}{p}}\right) \cdot t^{1-2/p} \Big|_0^1 - e^{-t} \Big|_1^\infty \\ &= O(1) \end{aligned}$$
- So, $E[(\text{PDy})_i^2] = O\left(\frac{1}{s}\right) |y|_2^2 = O\left(\frac{1}{s}\right) (n^{1-\frac{2}{p}} |y|_p^2)$. **Why?**

Understanding $|\text{PDy}|_\infty$

- $E[(\text{PDy})_i] = 0$ for each hash bucket i , and $E[(\text{PDy})_i^2] = O\left(\frac{1}{s}\right) (n^{1-\frac{2}{p}} |y|_p^2)$
- Bernstein's bound: Suppose R_1, \dots, R_n are independent, and for all j , $|R_j| \leq K$, and $\text{Var}[\sum_j R_j] = \sigma^2$. There are constants C, c , so that for all $t > 0$,

$$\Pr\left[\left|\sum_j R_j - E\left[\sum_j R_j\right]\right| > t\right] \leq C \left(e^{-\frac{ct^2}{\sigma^2}} + e^{-\frac{ct}{K}}\right)$$
- Recall $(\text{PDy})_i = \sum_j \delta(h(j) = i) \cdot \sigma_j \cdot (\text{Dy})_j$, and set $R_j = \delta(h(j) = i) \cdot \sigma_j \cdot (\text{Dy})_j$
- Want $|\text{PDy}|_\infty \approx |\text{Dy}|_\infty$, where $|\text{Dy}|_\infty \in \left[\frac{|y|_p}{10^{1/p}}, 10^{1/p} |y|_p\right]$ with probability $> 4/5$
- Set $t = \frac{|y|_p}{100}$ and $s = \Theta\left(n^{1-\frac{2}{p}} \log n\right)$, to get $\frac{1}{n^2}$ error probability in Bernstein's bound
- But what is $K = \max_j |R_j|$?

Understanding the Large Elements

- Recall $(PDy)_i = \sum_j \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$, and set $R_j = \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$
- We will separately handle those R_j for which $|R_j| > \frac{\alpha |y|_p}{\log n}$, for a sufficiently small constant $\alpha > 0$. If $|R_j| > \frac{\alpha |y|_p}{\log n}$, then necessarily $|(Dy)_j| \geq \frac{\alpha |y|_p}{\log n}$
 - We call such a j **large** if $|(Dy)_j| \geq \frac{\alpha |y|_p}{\log n}$, otherwise j is **small**. How many indices j are large?
- Recall: $|(Dy)_j| = |y_j|/E_j^{1/p}$
- $\Pr_D[|(Dy)_j| \text{ is large}] = \Pr \left[\frac{|y_j|}{E_j^{1/p}} \geq \frac{\alpha |y|_p}{\log n} \right] = \Pr \left[\frac{|y_j|^p}{\alpha^p |y|_p^p} (\log^p n) \geq E_j \right]$
 $= 1 - e^{-\frac{|y_j|^p (\log^p n)}{\alpha^p |y|_p^p}} \leq \frac{|y_j|^p (\log^p n)}{\alpha^p |y|_p^p}$, so the expected number of large j is $O(\log^p n)$

Understanding the Large Elements

- Recall $(PDy)_i = \sum_j \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$, and set $R_j = \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$
- We have shown the expected number of large j is $O(\log^p n)$, so by a Markov bound we have $O(\log^p n)$ large j with constant probability and we condition on D satisfying this
- We also condition on $|Dy|_\infty \in \left[\frac{|y|_p}{10^{\frac{1}{p}}}, 10^{\frac{1}{p}} |y|_p \right]$, which held with probability $> 4/5$
- All the large j get perfectly hashed into separate hash buckets by P
 - We are throwing $O(\log^p n)$ balls into $s \geq n^{1-2/p}$ bins
- We apply Bernstein for each hash bucket separately
 - We apply Bernstein on the small indices j inside a hash bucket!

Understanding the Large Elements

- $E[(PDy)_i] = 0$ for each hash bucket i , and $E[(PDy)_i^2] = O\left(\frac{1}{s}\right) (n^{1-\frac{2}{p}} |y|_p^2)$
- Bernstein's bound: Suppose R_1, \dots, R_n are independent, and for all j , $|R_j| \leq K$, and $\text{Var}[\sum_j R_j] = \sigma^2$. There are constants C, c , so that for all $t > 0$,
 - $\Pr[|\sum_j R_j - E[\sum_j R_j]| > t] \leq C (e^{-\frac{ct^2}{\sigma^2}} + e^{-\frac{ct}{K}})$
- $(PDy)_i = \sum_j \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$, and $R_j = \delta(h(j) = i) \cdot \sigma_j \cdot (Dy)_j$
- Can assume $K = \max_j |R_j| \leq \frac{\alpha |y|_p}{\log n}$, since there is at most one large j in any hash bucket $(PDy)_i$
- Set $t = \frac{|y|_p}{100}$, and $s = \Theta(n^{1-\frac{2}{p}} \log n)$ in Bernstein's bound, to get for a bucket $(PDy)_i$:

$$\Pr \left[\left| \sum_{\text{small } j} \delta(h(j) = i) \sigma_j (Dy)_j \right| > \frac{|y|_p}{100} \right] \leq C \left(e^{-\Theta(\log n)} + e^{-c \frac{(\log n)}{100\alpha}} \right) \leq \frac{1}{n^2}$$
- By a union bound over all the s buckets, the "signed sum" of small j in every bucket will be at most $\frac{|y|_p}{100}$ 24

Wrapping Up

- For all i ,
 - $|(PDy)_i| \leq \frac{|y|_p}{100}$ if no large indices in i -th bucket
 - $|(PDy)_i| = |\sigma_j(Dy)_j| \pm \frac{|y|_p}{100}$ if exactly one large index j in i -th bucket
 - No bucket contains more than 1 large index j
- We conditioned on $|Dy|_\infty \in \left[\frac{|y|_p}{10^{\frac{1}{p}}}, 10^{\frac{1}{p}}|y|_p \right]$
- What is $|PDy|_\infty$?
 - $|PDy|_\infty \leq 10^{\frac{1}{p}}|y|_p + \frac{|y|_p}{100}$ and $|PDy|_\infty \geq \frac{|y|_p}{10^{\frac{1}{p}}} - \frac{|y|_p}{100}$
- So just output $|PDy|_\infty$ as your estimate to $|y|_p$
- Total space is $s = O(n^{1-\frac{2}{p}} \log n)$ words, which is $O(n^{1-\frac{2}{p}} \log^2 n)$ bits

Outline

- Quick recap of ℓ_1 -regression, and how to speed it up
- Introduction to the Streaming Model
- Estimating Norms in the Streaming Model
- **Heavy Hitters in a Stream**
- Estimating Number of Non-Zero Entries (ℓ_0)

Heavy Hitter Guarantees

- l_1 – guarantee
 - output a set containing all items j for which $|x_j| \geq \phi |x|_1$
 - the set should not contain any j with $|x_j| \leq (\phi - \epsilon) |x|_1$
- l_2 – guarantee
 - output a set containing all items j for which $x_j^2 \geq \phi |x|_2^2$
 - the set should not contain any j with $x_j^2 \leq (\phi - \epsilon) |x|_2^2$
- l_2 – guarantee can be much stronger than the l_1 – guarantee
 - Suppose $x = (\sqrt{n}, 1, 1, 1, \dots, 1)$
 - Item 1 is an l_2 -heavy hitter for constant ϕ, ϵ , but not an l_1 -heavy hitter
 - If $|x_j| \geq \phi |x|_1$, then $x_j^2 \geq \phi^2 |x|_1^2 \geq \phi^2 |x|_2^2$

Heavy Hitter Intuition

- Suppose you are promised at the end of the stream, $x_i = n$, and $x_j \in \{0,1\}$ for $j \in \{1, 2, \dots, n\}$ with $j \neq i$
- How would you find the identity i ?
- For each j in $\{1, 2, 3, \dots, \log n\}$, let $A_j \subset [n]$ be the set of indices with j -th bit in their binary representation equal to 0, and B_j be the set with j -th bit equal to 1
- Compute $a_j = \sum_{i \in A_j} x_i$ and $b_j = \sum_{i \in B_j} x_i$ for each j in $\{1, 2, \dots, \log n\}$
- Read off the identity of item i

Heavy Hitter Intuition Continued

- Suppose you are promised at the end of the stream, $x_i = 100\sqrt{n \log(n)}$, and $x_j \in \{0,1\}$ for $j \in \{1, 2, \dots, n\}$ with $j \neq i$
- How would you find the identity i ?
- For each j in $\{1, 2, 3, \dots, \log n\}$, let $A_j \subset [n]$ be the set of indices with j -th bit in their binary representation equal to 0, and B_j be the set with j -th bit equal to 1
- Compute $a_j = \sum_{i \in A_j} \sigma_i \cdot x_i$ and $b_j = \sum_{i \in B_j} \sigma_i \cdot x_i$ for each j in $\{1, 2, \dots, \log n\}$
- Read off the identity of item i ?
- Additive Chernoff bound implies magnitude of “noise” in a count is at most $\sqrt{n \log(n)}$ w.h.p.
- Remove assumptions: (1) $x_i = 100\sqrt{n \log(n)}$ and (2) and $x_j \in \{0,1\}$ for $j \in \{1, 2, \dots, n\}$ with $j \neq i$ ⁹