Space-optimal $L_k$ norm estimation

Piotr Indyk

MIT
**Lₖ norm**

- A stream is a sequence of updates \((i, a), i=1...m\)
  \[x_i = x_i + a\]
  (a integer)
- Want to estimate \(||x||_k\)
- Today, we will see an algorithm:
  - Space \(O(m^{1-2/k} \text{ polylog})\), polylog=(log m + log M)^{O(1)}
  - Approximation: \(O(\log M \log m)\)
    (can be improved to \(c=1±\varepsilon\))
  - Two passes
    (can be easily reduced to one)
  - For the purpose of \(O()\) notation we assume \(k\) is constant
Tool: heavy hitters

• Define

\[ HH_p^\varphi (x) = \{ i: |x_i|^p \geq \varphi \|x\|_p^p \} \]

• \( L_p \) Heavy Hitter Problem:
  – Parameters: \( \varphi \) and \( \varphi' \) (today: \( \varphi' = \varphi/2 \))
  – Goal: return a set \( S \) of coordinates s.t.
    • \( S \) contains \( HH_p^\varphi (x) \)
    • \( S \) is included in \( HH_p^{\varphi'} (x) \)

• If we have two passes, then
  – \( S=HH_p^\varphi (x) \)
  – We know the value of \( x_i \) for each \( i \in S \)

• Algorithms:
  – \( L1 \) heavy hitters using space \( O(1/\varphi \text{ polylog}) \)
  – \( L2 \) heavy hitters using same space [Charikar-Chen-Farach-Colton’02]
Point Queries/Heavy Hitters

• L1 point queries:
  – \( w = O(1/\varphi) \)
  – Prepare a random hash function \( h: \{1..m\} \rightarrow \{1..w\} \)
  – Maintain an array \( Z = [Z_1, \ldots, Z_w] \) such that \( Z_j = \sum_{i: h(i)=j} x_i \)
  – To estimate \( x_i \) return \( x^*_i = Z_{h(i)} \)
  – Amplify probability of success using median of several estimators

• L2 point queries:
  – As before, but \( Z_j = (1 \pm \alpha) \sum_{i: h(i)=j} x_i^2 \)
    and use \( x^*_i^2 = Z_{h(i)} \) to estimate \( x_i^2 \)
  – Each \( Z_j \) can be maintained using L2 sketches
Main ideas (*)

• Bucketing:
  - Define $B_j=\{i: |x_i| \in [2^j, 2^{j+1})\}$, $b_j=|B_i|$, $j=0…\log M -1$
  - We have $\sum_j 2^k b_j \leq ||x||_k \leq \sum_j 2^{k(j+1)} b_j$
  - Unfortunately, cannot estimate $b_j$’s in general
    • Take the largest $j$ such that $b_j>0$
    • Then $2^j$ is a 2-approximation of $||x||_\infty$

• “Important” buckets:
  - $J=\{j: 2^k b_j \geq 1/(2^{k+1} \log M) ||x||_k \}$
  - Any term $2^k b_j$ provides an $O(\log M)$-approximation to $||x||_k$
  - There exists at least one important bucket

• Turns out we can estimate $b_j$ for $j \in J$
(although our algorithm does not do that explicitly)
Algorithm overview

• “Algorithm”:
  – For $t=1\ldots\log m$
    • Choose $S(t) \subseteq \{1\ldots m\}$ be a random set such that $\Pr[i \in S(t)] = 2^{-t}$ (independently for each $i$)
    • Find $\text{val}_t$ whose value is “likely” to be “close to” $\max_{i \in S(t)} |x_i|
    • Define $Z_t = 2^t \text{val}_t^k$
  – Estimator: $Z = \sum_t Z_t$

• Intuition:
  – If $x_{S(t)}$ has an entry of magnitude $\text{val}_t$, then it is likely that the whole vector $x$ contains at least $2^t$ entries of magnitude $\text{val}_t$ or higher
  – This contributes $Z_t = 2^t \text{val}_t^k$ to $\|x\|_k^k$
  – We take care of all ranges of bucket sizes by trying out $t=1\ldots\log m$
Dealing with important buckets

- Assume
  - \(2^{k_j} b_j \geq 1/(2^{k+1} \log M) \|x\|_k^k \rightarrow 2^{k_j} \geq 1/(2^{k+1} \log M) \|x\|_k^k / b_j\)
  - \(2^t \leq b_j \leq 2^{t+1}\)

- Projection: let \(S(t) \subseteq \{1\ldots m\}\) be a random set such that \(\Pr[i \in S(t)] = 2^{-t}\) (independently for each \(i\))

- Claim: with constant probability \(p > 1 - 1/e - 1/2\), both of the following events hold:
  - There exists \(i \in B_j \cap S(t)\)
  - We have \(\|x_{S(t)}\|_k^k \leq 2 \|x\|_k^k / 2^t\)

- In that case
  \[|x_i|_k^k \geq 2^{k_j} \geq 1/(2^{k+1} \log M) \|x\|_k^k / 2^{t+1} \geq 1/(2^{k+3} \log M) \|x_{S(t)}\|_k^k = d \|x_{S(t)}\|_k^k\] (\(d = 1/(2^{k+3} \log M)\))

- \(x_i\) is an \(L_k\) heavy hitter in \(x_{S(t)}\)
Claim proof

• There exists \( i \in B_j \cap S(t) \) with prob. \( >1-1/e \)
  – The probability that \( B_j \cap S(t) \) is empty is at most \( (1-2^{-t})^{bj} < 1/e \)

• We have \( ||x_{S(t)}||_k \leq 2 ||x||_k / 2^t \), with prob. \( \geq 1/2 \)
  – We have \( E[||x_{S(t)}||_k] = ||x||_k / 2^t \)
  – Markov inequality
Finding $L_k$ HH via L2 HH

- We have $|x_i|^k \geq d \|x_{S(t)}\|_k^k$, $i \in S(t)$
- Then
  
  $|x_i|^2 \geq d^{2/k} \|x_{S(t)}\|_k^2$
  
  $\geq d \left[\|x_{S(t)}\|_2 / m^{1/2 - 1/k}\right]^2$
  
  $= d / m^{1 - 2/k} \|x_{S(t)}\|_2^2$

- Therefore, we can find $i$ and $x_i$ using $m^{1 - 2/k}$ polylog space L2 heavy hitter algorithm
Algorithm

• Sketching:
  – For $t=1\ldots \log m$, in parallel
    • Choose the random set $S(t)$
    • Prepare the L2 heavy hitter algorithm that finds the set
      $$I(t) = \{i: |x_i| \geq d/m^{1-2/k} \|x_{S(t)}\|_2^2\}$$
      as well as the values $x_i, i \in I(t)$
  
• Estimation:
  – For $t=1\ldots \log m$
    • Set $val_t = \max_{i \in I(t)} |x_i|$ (or 0 if $I(t)$ empty)
    • $Z_t = 2^t val_t^k$
    – Estimator: $Z = \sum_t Z_t$
Analysis: Lower Bound

• Lemma 1: For any important $B_j$, $2^{t} \leq b_j \leq 2^{t+1}$, we have $Z_t \geq \|x\|_k^k / O(\log M)$ with constant probability

• Proof:
  – With probability $p$, we find some $i \in B_j$
  – Then $\text{val}_t \geq |x_i| \geq 1/(2^{k+1}\log M) \|x\|_k^k / b_j$ and $Z_t = 2^t \text{val}_t = \|x\|_k^k / O(\log M)$

• Therefore, we have $Z \geq \|x\|_k^k / O(\log M)$ with constant probability
Analysis: Upper Bound

• Lemma 2: For each $t$, we have $E[Z_t] \leq \|x\|_k^k$

• Proof:
  
  $- Z_t \leq 2^t \|x_{S(t)}\|_\infty^k$
  
  $- E[2^t \|x_{S(t)}\|_\infty^k] \leq 2^t \sum_i |x|_i^k (1-1/2^t)^i^{-1}1/2^t \leq \|x\|_k^k$

• Therefore, we have $E[Z] \leq \log m \|x\|_k^k$
Discussion

• How can this be improved?
• Lower bound: can shave off the log M factor by analyzing all important buckets, not just one
• Upper bound: more tricky. Consider $x=(1,0,\ldots,0)$

For each $Z_t$ we have:
– $E[Z_t]=1$ (with probability $1/2^t$ we have $Z_i=2^t$), which means $E[Z]=\log m$
– $Pr[Z_t>0] = 1/2^t$, which means $Z$ is constant with arbitrarily high (constant) probability
– A logarithmic gap

• The gap can be avoided by using a more careful (but also more messy) estimator
State of the art

- \((1/\varepsilon + \log (nm))^{O(1)} n^{1-2/k}\) [Indyk-Woodruff’05]
- \(1/\varepsilon^{2+4/k} \log^2 n \log(nm) n^{1−2/k}\) [Bhuvanagiri, Ganguly, Kesh and Saha’06]
- \(1/\varepsilon^{2+6/k} \log n \log(nm) n^{1−2/k}\) [Andoni-Krauthgamer-Onak, July’10]
- \(1/\varepsilon^{2+4/k} g(n) \log (m) \log(nm) n^{1-2/k}\) [Braverman-Ostrovsky, November’10]