L2 Norm Estimation

MIT

Piotr Indyk

Lecture 2
L2 Norm Estimation

Vector $x$:  

- A stream is a sequence of updates $(i,a)$  
  $x_i = x_i + a$
- Want to estimate $\|x\|_2$ up to $1 \pm \epsilon$
- Last week, we have seen how to do that for $\|x\|_0$:
  - Space: $(1/\epsilon + \log m)^{O(1)}$
  - Technique:
    - Linear sketches $\sum_{S} x_i = \sum_{i \in S} x_i$ for "random" sets $S$
    - (Somewhat messy) estimator
- Today: two methods for estimating $\|x\|_2$ + applications
  - Alon-Matias-Szegedy - Really cute and simple
  - Johnson-Lindenstrauss - Need in future lectures
- First: two digressions
Digression 1

• Our algorithm computes a linear sketch of the vector $x$:
  – Linear sketches $\text{Sum}_S(x) = \sum_{i \in S} x_i$ for “random” sets $S$
    • $\log(m)/\varepsilon$ values of $T = 1, 1 + \varepsilon, \ldots, m$
    • $k$ sets $S_j$ such that $\Pr[i \in S_j] = 1/T$
  – Can represent as a product of $Ax$, for a $(\log(m)/\varepsilon \times k) \times m$ 0-1 matrix $A$
• Our setup:
  – World: provides a stream, defining $x$
  – We: choose a random $A$
  – The method works with “high probability”

• Comments:
  – Do not need to assume that a “source” generates $x$
  – Useful for composing algorithms, i.e., when $x$ is itself an output of another algorithm (later in the course)
L2 norm
Why $L_2$ norm?

- Database join (on $A$):
  - All triples $(\text{Rel1.A}, \text{Rel1.B}, \text{Rel2.B})$
    s.t. $\text{Rel1.A} = \text{Rel2.A}$
- Self-join: if $\text{Rel1} = \text{Rel2}$
- Size of self-join:
  $$\sum_{\text{val of } A} \text{Rows(val)}^2$$
- Updates to the relation increment/decrement
  $\text{Rows(val)}$

### Example Table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lec1</td>
<td>distinct</td>
<td></td>
</tr>
<tr>
<td>Lec1</td>
<td>elements</td>
<td></td>
</tr>
<tr>
<td>Lec1</td>
<td>norm</td>
<td></td>
</tr>
<tr>
<td>Lec2</td>
<td>L2</td>
<td></td>
</tr>
<tr>
<td>Lec2</td>
<td>norm</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Example Table

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>Rel1.B</th>
<th>Rel2.B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lec1</td>
<td>distinct</td>
<td>distinct</td>
<td></td>
</tr>
<tr>
<td>Lec1</td>
<td>distinct</td>
<td>elements</td>
<td></td>
</tr>
<tr>
<td>Lec1</td>
<td>distinct</td>
<td>norm</td>
<td></td>
</tr>
<tr>
<td>Lec1</td>
<td>elements</td>
<td>distinct</td>
<td></td>
</tr>
<tr>
<td>Lec1</td>
<td>elements</td>
<td>elements</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Algorithm I: AMS
Choose $r_1 \ldots r_m$ to be i.i.d. r.v., with $\Pr[r_i=1]=\Pr[r_i=-1]=1/2$

Maintain

$$Z = \sum_i r_i x_i$$

under increments/decrements to $x_i$

Algorithm A:

$$Y = Z^2$$

“Claim”: $Y$ “approximates” $||x||_2^2$ with “good” probability
Analysis

- The expectation of $Z^2 = (\sum_i r_i x_i)^2$ is equal to
  $$E[Z^2] = E[\sum_{i,j} r_i x_i r_j x_j] = \sum_{i,j} x_i x_j E[r_i r_j]$$

- We have
  - For $i \neq j$, $E[r_i r_j] = E[r_i] E[r_j] = 0$ – term disappears
  - For $i = j$, $E[r_i r_j] = 1$

- Therefore
  $$E[Z^2] = \sum_i x_i^2 = ||x||_2^2$$
  (unbiased estimator)
Analysis, ctd.

• The second moment of $Z^2 = (\sum_i r_i x_i)^2$ is equal to the expectation of $Z^4 = (\sum_i r_i x_i)(\sum_i r_i x_i)(\sum_i r_i x_i)(\sum_i r_i x_i)$

• This can be decomposed into a sum of
  - $\sum_i (r_i x_i)^4$ → expectation $= \sum_i x_i^4$
  - $6 \sum_{i<j} (r_i r_j x_i x_j)^2$ → expectation $= 6\sum_{i<j} x_i^2 x_j^2$
  - Terms involving single multiplier $r_i x_i$ (e.g., $r_1 x_1 r_2 x_2 r_3 x_3 r_4 x_4$) → expectation $= 0$

  Total: $\sum_i x_i^4 + 6\sum_{i<j} x_i^2 x_j^2$

• The variance of $Z^2$ is equal to

$$E[Z^4] - E^2[Z^2] = \sum_i x_i^4 + 6\sum_{i<j} x_i^2 x_j^2 - (\sum_i x_i^2)^2$$

$$= \sum_i x_i^4 + 6\sum_{i<j} x_i^2 x_j^2 - \sum_i x_i^4 - 2 \sum_{i<j} x_i^2 x_j^2$$

$$= 4\sum_{i<j} x_i^2 x_j^2$$

$$\leq 2(\sum_i x_i^2)^2$$
Analysis, ctd.

• We have an estimator $Y = Z^2$
  – $E[Y] = \sum_i x_i^2$
  – $\sigma^2 = \text{Var}[Y] \leq 2 (\sum_i x_i^2)^2$

• Chebyshev inequality:
  $\Pr[|E[Y] - Y| \geq c \sigma] \leq 1/c^2$

• Algorithm B:
  – Maintain $Z_1 \ldots Z_k$ (and thus $Y_1 \ldots Y_k$), define $Y' = \sum_i Y_i / k$
  – $E[Y'] = k \sum_i x_i^2 / k = \sum_i x_i^2$
  – $\sigma'^2 = \text{Var}[Y'] \leq 2k(\sum_i x_i^2)^2 / k^2 = 2 (\sum_i x_i^2)^2 / k$

• Guarantee:
  $\Pr[|Y' - \sum_i x_i^2| \geq c (2/k)^{1/2} \sum_i x_i^2] \leq 1/c^2$

• Setting $c$ to a constant and $k = O(1/\epsilon^2)$ gives $(1 \pm \epsilon)$-approximation with const. probability
Digression 3

• Only needed that \( r_1 \ldots r_m \) are 4-wise independent

• **Definition:** identically distributed random variables \( r_1 \ldots r_m \), with each \( r_i \) chosen uniformly at random from \( \{0 \ldots P-1\} \), are \( t \)-wise independent if for any \( S \subseteq \{1 \ldots m\} \), \( |S|=t \), and \( u \in \{0 \ldots P-1\}^t \), we have

\[
\Pr[r_S = u] = \frac{1}{P^t}
\]

• Can generate such random variables using only \( O(t \log(Pm)) \) truly random bits
Digression 3 ctd

• Example I: \( k=2 \), for \( m=P \), \( P \) prime
  – Choose \( a,b \) independently uniformly at random from \( \{0\ldots P-1\} \)
  – Define \( r_i = a_i + b \mod P \)
  – For \( S = \{i,j\}, i \neq j \) and \( u = (u_1, u_2) \in \{0\ldots P-1\}^2 \), there exists exactly one pair \( (a,b) \) such that
    \[
    a_i + b \mod P = u_1 \\
    a_j + b \mod P = u_2
    \]
  – Therefore, \( \Pr[r_{\{i,j\}} = (u_1, u_2)] = 1/P^2 \)

• Example II: any \( k \), for \( m=P \), \( P \) prime
  – Use polynomials of degree \( k-1 \)
Recap

• What we did:
  – Maintain a “linear sketch” vector $Z = [Z_1 ... Z_k] = R\ x$
  – Estimator for $\|x\|_2^2$ : $\frac{(Z_1^2 + ... + Z_k^2)}{k} = \frac{\|Rx\|_2^2}{k}$
  – “Dimensionality reduction”: $x \rightarrow Rx$
    … but the tail somewhat “heavy”
  – Reason: only used second moment of the estimator
Algorithm II: Dim. Reduction (JL)
Interlude: Normal Distribution

• Normal distribution $N(0,1)$:
  – Range: $(-\infty, \infty)$
  – Density: $f(x)=e^{-x^2/2} / (2\pi)^{1/2}$
  – Mean=0, Variance=1

• Basic facts:
  – If $X$ and $Y$ independent r.v. with normal distribution, then $X+Y$ has normal distribution
  – $\text{Var}(cX)=c^2 \text{Var}(X)$
  – If $X,Y$ independent, then $\text{Var}(X+Y)=\text{Var}(X)+\text{Var}(Y)$
A different linear sketch

• Instead of ±1, let $r_i$ be i.i.d. random variables from $\text{N}(0,1)$
• Consider $Z = \sum_i r_i x_i$

We still have that $E[Z^2] = \sum_i x_i^2 = \|x\|_2^2$, since:
- $E[r_i] E[r_j] = 0$
- $E[r_i^2]$ = variance of $r_i$, i.e., 1

• As before we maintain $Z=[Z_1 \ldots Z_k]$ and define $Y = \|Z\|_2^2 = \sum_j Z_j^2$ (so that $E[Y]=k\|x\|_2^2$ )
• We show that there exists $C>0$ s.t. for small enough $\varepsilon>0$

$$\Pr[ |Y - k\|x\|_2^2| > \varepsilon k\|x\|_2^2 ] \leq \exp(-C \varepsilon^2 k)$$
Proof

• See the attached notes, by Ben Rossman and Michel Goemans