Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2008-024 May 2, 2008

Block Heavy Hitters

Alexandr Andoni, Khanh Do Ba, and Piotr Indyk

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Block Heavy Hitters

Alexandr Andoni MIT
Khanh Do Ba MIT
Piotr Indyk MIT

April 11, 2008

Abstract

We study a natural generalization of the heavy hitters problem in the streaming context. We term this generalization *block heavy hitters* and define it as follows. We are to stream over a matrix A, and report all *rows* that are heavy, where a row is heavy if its ℓ_1-norm is at least ϕ fraction of the ℓ_1 norm of the entire matrix A. In comparison, in the standard heavy hitters problem, we are required to report the matrix *entries* that are heavy. As is common in streaming, we solve the problem approximately: we return all rows with weight at least ϕ, but also possibly some other rows that have weight no less than $(1-\epsilon)\phi$. To solve the block heavy hitters problem, we show how to construct a linear sketch of A from which we can recover the heavy rows of A.

The block heavy hitters problem has already found applications for other streaming problems. In particular, it is a crucial building block in a streaming algorithm of [AIK08] that constructs a small-size sketch for the Ulam metric, a metric on non-repetitive strings under the edit (Levenshtein) distance.

We prove the following theorem. Let $M_{n,m}$ be the set of real matrices A of size n by m, with entries from $E = \frac{1}{nm} \cdot \{0, 1, \ldots nm\}$. For a matrix A, let A_i denote its i^{th} row.

Theorem 0.1. *Fix some $\epsilon > 0$, and $n, m \geq 1$, and $\phi \in [0, 1]$. There exists a randomized linear map (sketch) $\mu : M_{n,m} \to \{0,1\}^s$, where $s = O(\frac{1}{\epsilon^5 \phi^2} \log n)$, such that the following holds. For a matrix $A \in M_{n,m}$, it is possible, given $\mu(A)$, to find a set $W \subset [n]$ of rows such that, with probability at least $1 - 1/n$, we have:*

- *for any $i \in W$, $\frac{\|A_i\|_1}{\|A\|_1} \geq (1-\epsilon)\phi$ and*
- *if $\frac{\|A_i\|_1}{\|A\|_1} \geq \phi$, then $i \in W$.*

Moreover, μ can be of the form $\mu(A) = \mu'(\rho(A_1), \rho(A_2), \ldots \rho(A_n))$, where $\rho : E^m \to \mathbb{R}^k$ and $\mu' : \mathbb{R}^{kn} \to \{0,1\}^s$ are randomized linear mappings. That is, the sketch μ is obtained by first sketching the rows of A (using the same function ρ) and then sketching those sketches.

Our construction is inspired by the CountMin sketch of [CM05], and may be seen as a CountMin sketch on the projections of the rows of A.

Proof. **Construction of the sketch.** We define the function ρ as an ℓ_1 projection into a space with $k = O(\frac{1}{\epsilon^2} \log n)$ dimensions, achieved through a standard Cauchy distribution projection.

Namely, the function ρ is determined by k vectors $\vec{c}_1, \ldots \vec{c}_k \in \mathbb{R}^m$, with coordinates chosen iid from the Cauchy distribution with pdf $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. Then $\rho(\vec{x})$, for some $\vec{x} \in E^m$, is given by

$$\rho(\vec{x}) = (\vec{c}_1\vec{x}, \vec{c}_2\vec{x}, \ldots \vec{c}_k\vec{x}).$$

The function μ' takes as input $\rho(A_1), \ldots \rho(A_n)$, and produces k hash tables, each having $l = O(\frac{1}{\epsilon^2\phi})$ cells. The j^{th} cell of the ith hash table $H^{(i)}$, for $j \in [l]$, is given by

$$H^{(i)}_j = \sum_{q:h_i(q)=j} [\rho(A_q)]_i.$$

See Figure 1 for an illustration.

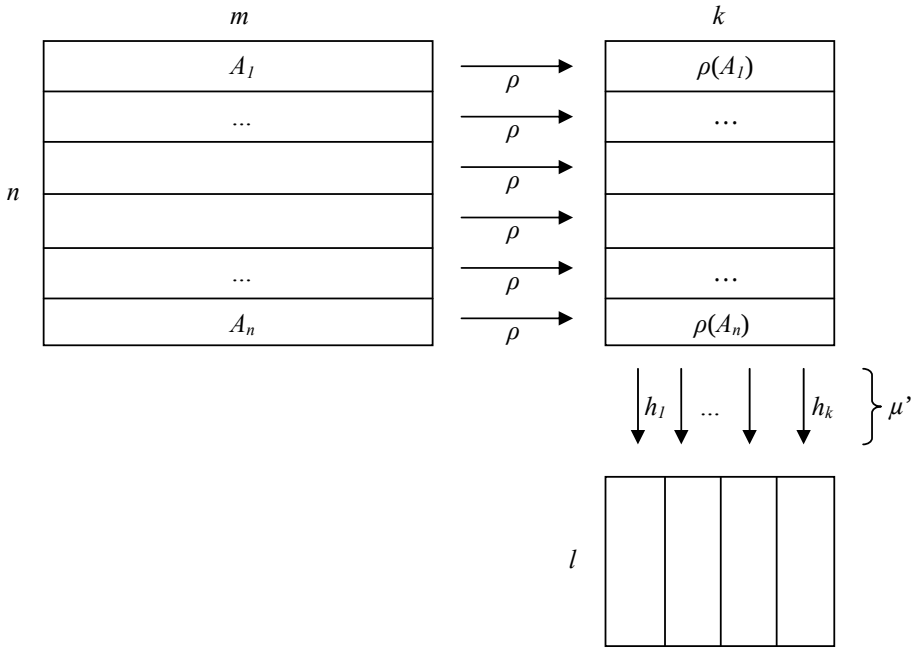

Figure 1: Illustration of μ as a double sketch.

Reconstruction. Given a sketch $\mu(A) = \mu'(\rho(A_1), \ldots \rho(A_n))$, we construct the desired set W as follows. For each $w \in [n]$, consider the vector $\vec{r}_w = \left(|H^{(i)}_{h_i(w)}|\right)_{i\in[k]}$. Then w is included in W iff median$(\vec{r}_w) > (1 - \epsilon/2)\phi$. In words, for any block w we consider the cell of a hash table $H^{(i)}$ into which w falls (one for each i). If the majority of these cells contain a value greater or equal to $(1 - \epsilon/2)\phi$ (in magnitude), then w is included in W.

Sketch size. As described, the sketch $\mu(A) = \mu'(\rho(A_1), \ldots \rho(A_n))$ consists of $k \cdot l = O(\frac{1}{\epsilon^4\phi}\log n)$ real numbers. We note that, by usual arguments, it is enough to store all the real numbers up to precision $O(\epsilon\phi)$ and cut off when the absolute value is beyond a constant such as 2. The resulting size of the sketch (in bits) is $s = O(\frac{1}{\epsilon^5\phi^2}\log n)$.

Analysis of correctness. We proceed to proving that the set W satisfies the desired properties. Since our sketches are linear, we assume without loss of generality that $\|A\|_1 = 1$.

First, consider any w such that $\|A_w\|_1 \geq \phi$. We would like to prove that $w \in W$ w.h.p. For this purpose, it is sufficient to prove that, for fixed $i \in [k]$, we have that $|H^{(i)}_{h_i(w)}| > (1-\epsilon/2)\phi$ with probability $\geq 1/2 + \Omega(\epsilon)$. Then, a standard application of the Chernoff bound will imply that $\text{median}(\vec{r}_w) > (1-\epsilon/2)\phi$ w.h.p.

So fix some $i \in [k]$, and consider the cell $h_i(w)$ of the hash table $H^{(i)}$. Let $\chi[E]$ denote the indicator variable of an event E. The mass that falls into the cell $h_i(w)$ is equal to the following quantity:

$$\begin{aligned} H^{(i)}_{h_i(w)} &= [\rho(A_w)]_i + \sum_{j \in [n], j \neq w} [\rho(A_j)]_i \cdot \chi[h_i(j) = h_i(w)] \\ &= \vec{c}_i \cdot A_w + \vec{c}_i \cdot \left(\sum_{j \in [n], j \neq w} A_j \cdot \chi[h_i(j) = h_w(j)] \right) \\ &= \vec{c}_i \cdot \left(A_w + \left(\sum_{j \in [n], j \neq w} A_j \cdot \chi[h_i(j) = h_w(j)] \right) \right). \end{aligned}$$

Now, consider the vector $\vec{z} = \left(\sum_{j \in [n], j \neq w} A_j \cdot \chi[h_i(j) = h_w(j)] \right)$. The expected norm of $\vec{z}$ is at most

$$\mathbb{E}_{h_i}\left[\|\vec{z}\|_1\right] \leq \frac{1}{l} \sum_{j \in [n], j \neq w} \|A_j\|_1 \leq 1/l = O(\epsilon^2 \phi).$$

By Markov's inequality, with probability at least $1 - O(\epsilon)$, we have $\|\vec{z}\|_1 \leq \epsilon\phi/4$ and thus $\|A_w + \vec{z}\|_1 \geq (1-\epsilon/4)\phi$. It follows that the random variable $|(A_w + \vec{z}) \cdot \vec{c}_i|$ has a Cauchy distribution with median $\|A_w + \vec{z}\|_1 \geq (1-\epsilon/4)\phi$. By standard properties of Cauchy distributions we have

$$\left| H^{(i)}_{h_i(w)} \right| \geq (1-\epsilon/4) \cdot (1-\epsilon/4)\phi > (1-\epsilon/2)\phi$$

with probability at least $(1/2 + \Omega(\epsilon))(1 - O(\epsilon)) = 1/2 + \Omega(\epsilon)$.

Next we prove that if $\|A_w\|_1 \leq (1-\epsilon)\phi$, then $w \notin W$ w.h.p. As above, we just need to prove that $|H^{(i)}_{h_i(w)}| < (1-\epsilon/2)\phi$ with probability $\geq 1/2 + \Omega(\epsilon)$. We again consider the vector $\vec{z} = \left(\sum_{j \in [n], j \neq w} A_j \cdot \chi[h_i(j) = h_j(w)] \right)$, and similarly deduce that, with probability at least $1 - O(\epsilon)$, we have $\|\vec{z}\|_1 \leq \epsilon\phi/4$ and thus $\|A_w + z\|_1 \leq (1 - \frac{3}{4}\epsilon)\phi$. Again by standard properties of Cauchy distributions, we conclude that

$$\left| H^{(i)}_{h_i(w)} \right| \leq (1+\epsilon/4) \cdot (1 - \tfrac{3}{4}\epsilon)\phi < (1-\epsilon/2)\phi$$

with probability at least $(1/2 + \Omega(\epsilon))(1 - O(\epsilon)) = 1/2 + \Omega(\epsilon)$. □

References

[AIK08] Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer. Overcoming the ℓ_1 non-embeddability barrier: Algorithms for product metrics. *Manuscript*, 2008.

[CM05] G. Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms*, 55(1):58–75, 2005.