Sublinear Time and Space Algorithms 2018B – Lecture 5
Heavy Hitters and Compressed Sensing*

Robert Krauthgamer

Today we will see some applications of point queries.

1 Application 1: Heavy Hitters (Frequent Items)

Problem Definition: For parameter $\phi \in (0,1)$ and $p \in [1,\infty)$, define

$$HH^p_\phi(x) = \{i \in [n] : \ |x_i| \geq \phi \|x\|_p\}.$$

Observe that its cardinality is bounded by $\left|HH^p_\phi(x)\right| \leq 1/\phi^p$.

We will focus on $p = 1$ and ϕ is "not too small".

Approximate Heavy Hitters:

Parameters: $\phi, \varepsilon \in (0,1)$.

Goal: return a set $S \subseteq [n]$ such that

$$HH^p_\phi \subseteq S \subseteq HH^p_{\phi(1-\varepsilon)}.$$

Reduction from HH to point query (for $p = 1$):

Assume we have an algorithm for ℓ_1 point queries with parameter $\alpha = \varepsilon\phi/2$, and amplify its success probability to $1 - \frac{1}{3n}$ if needed.

1. compute an estimate $\tilde{x}_i$ for every $i \in [n]$ using this algorithm (this step takes time $O(n \log n)$ or even more)

2. report the set $S = \{i : \ \tilde{x}_i \geq (\phi - \varepsilon\phi/2)\|x\|_1\}$ (it is easy to know $\|x\|_1$ when $x \geq 0$, but more difficult in general)

Storage requirement: We can employ algorithm CountMin+ for ℓ_1 point queries, which requires $O(\alpha^{-1} \log n)$ words, and has error probability $1/n^2$, which is small enough. Then our approximate HH algorithm will take $O(\phi^{-1}\varepsilon^{-1} \log^2 n)$ bits.

*These notes summarize the material covered in class, usually skipping proofs, details, examples and so forth, and possibly adding some remarks, or pointers. The exercises are for self-practice and need not be handed in. In the interest of brevity, most references and credits were omitted.

Correctness: With probability $\geq 2/3$, all the n estimates are correct within additive $\varepsilon/2$. In this case, S contains all the ϕ-HH, and is contained in the $(\phi(1-\varepsilon))$-HH.

Exer: Extend the above approach to $p = 2$ (using CountSketch). How much storage it requires? Use the AMS sketch to estimate the ℓ_2-norm.

2 Application 2: Compressed Sensing (or Sparse Recovery)

Problem Definition: The input is a "signal" $x \in \mathbb{R}^n$, but instead of reading it directly we have only via linear measurements, i.e., we can observe/access $y_i = \langle A_i, x \rangle$ for $A_1, \ldots, A_m \in \mathbb{R}^n$ of our choice. Informally, the goal is to design few A_i's and then to use them recover x. We shall focus on non-adaptive A_i, i.e., the entire sequence has to be determined in advance.

Let $A_{m \times n}$ be a matrix whose rows are the A_i's, then we know that $Ax = y$. A trivial solution is to choose A that is invertible, which requires $m = n$. In general, this is optimal, because for smaller m there might be infinitely many solutions x to $Ax = y$.

Initial goal: Suppose that x is k-sparse (has at most k nonzeros, i.e., $\|x\|_0 = k$). What $m = m(n,k)$ is needed to recover x?

True goal: Suppose x is approximately k-sparse. For what m can we recover an approximation to x?

Remark: In most applications, it's preferable that A has bounded precision (i.e., the entries of A are integers of bounded magnitude), as otherwise y must be "acquired" with very high precision. Sometimes it's even important that A's entries are nonnegative.

CountMin Approach: Recall that CountMin is a (randomized) linear sketch of $x \in \mathbb{R}^n$, hence it can be viewed as multiplying x by some matrix A with $p = O(\alpha^{-1} \log n)$ rows.

Sparse 0-1 vector: Suppose first $x \in \{0,1\}^n$ and is k-sparse. Then $\|x\|_1 = k$, and a CountMin+ sketch of accuracy $\alpha = \frac{1}{3k}$ succeeds with probability at least $1 - 1/n$ in estimating all x_i's within additive $\pm\alpha\|x\|_1 \leq \pm\frac{1}{3}$, which can distinguish whether x_i is 0 or 1.

Sparse vector: If the nonzeros of x have different magnitudes, the above approach might require $\alpha \ll \frac{1}{k}$.

But a deeper inspection of CountMin shows that every coordinate has a good chance to "not collide" with any nonzero coordinate. This behavior is amplified by the repetitions + median trick's, and then WHP the estimator is exact, i.e., $\hat{x}_i = x_i$.

Exer: Show that a sketching matrix A with $m = O(k)$ rows (linear measurements) and whose entries are random Gaussians (or chosen uniformly from $[0,1]$) can recover with high probability every k-sparse input x. Show it also for an ε-coherent matrix for $\varepsilon = \frac{1}{10k}$.

Hint: It suffices that every $2k$ columns are linearly independent.

Approximately sparse vector: We will now prove an even more general result.

For $z \in \mathbb{R}^n$, denote by $z_{top(k)}$ the vector z after zeroing all *but* the k heaviest entries (largest in absolute value), breaking ties arbitrarily. Notice this vector is the "best" k-sparse approximation to z. Similarly, denote by $z_{tail(k)} \in \mathbb{R}^n$ the vector z after zeroing the k heaviest entries. Then $z_{tail(k)} = z - z_{top(k)}$ is the "error" of approximating z by a k-sparse vector.

Theorem 1 [Cormode and MuthuKrishnan, 2006]: CountMin+ with parameter $\alpha = \varepsilon/k$ can recover, with high probability, a vector $x' \in \mathbb{R}^n$ that satisfies

$$\|x - x'\|_1 \le (1 + 3\varepsilon)\|x_{tail(k)}\|_1.$$

In fact, $x' = \hat{x}_{top(k)}$ and is thus k-sparse. (Recall $\hat{x} \in \mathbb{R}^n$ is the estimate of algorithm CountMin.)

The above condition is usually called an ℓ_1/ℓ_1 guarantee.

Remark 1: Observe that if x is k-sparse, then this guarantees exact recovery. In general, it guarantees the output's "quality" (distance from true x) is comparable to the best k-sparse vector.

Remark 2: While in point queries we bounded the error in each coordinate separately, the above guarantee bounds the total error (over all coordinates).

Remark 3: Different constructions achieve/optimize for other guarantees like different norms, deterministic recovery, small explicit description of A, or fast recovery time. Often, the optimal number of measurements is $O(k \log(n/k))$ (ignoring dependence on ε).

Lemma 1a: CountMin+ with parameter $\alpha = \varepsilon/k$ computes, with high probability, an estimate $\hat{x}_i \in x_i \pm \alpha\|x_{tail(k)}\|_1$, i.e.,

$$\|x - \hat{x}\|_\infty \le \alpha\|x_{tail(k)}\|_1.$$

Exer: Prove this lemma.

Hint: Show that with high probability, both (a) coordinate i will not collide with the k (other) heaviest coordinates and (b) the contribution from the rest (tail) is comparable to the expectation.

Lemma 1b: If $\|x - \hat{x}\|_\infty \le \alpha\|x_{tail(k)}\|_1$ then $\|x - \hat{x}_{top(k)}\|_1 \le (1 + 3k\alpha)\|x_{tail(k)}\|_1$.

Proof of lemma: Let z_S denote the vector z after zeroing all coordinates outside $S \subset [n]$.

Let $\hat{T} \subset [n]$ be the indices of the k heaviest coordinates in $\hat{x}$, then by definition $x' = \hat{x}_{top(k)} = \hat{x}_{\hat{T}}$.

Let $T \subset [n]$ be the indices of the k heaviest coordinates in x, hence $x_T = x_{top(k)}$.

We can now bound (all norms are ℓ_1-norms) using the triangle inequality $\|a\| \in \|b\| \pm \|a - b\|$ (think of it as saying $\|b\| \approx \|a\|$)

$$\begin{aligned}
\|x - x'\| &= \|x_{\hat{T}} - \hat{x}_{\hat{T}}\| + \|x_{\neg\hat{T}} - 0\| && \text{separate coordinates of } \hat{T} \\
&= \|x_{\hat{T}} - \hat{x}_{\hat{T}}\| + \|x\| - \|x_{\hat{T}}\| \\
&\le \|x_{\hat{T}} - \hat{x}_{\hat{T}}\| + \|x\| - \|\hat{x}_{\hat{T}}\| + \|x_{\hat{T}} - \hat{x}_{\hat{T}}\| && \text{by } x \approx \hat{x} \text{ on } \hat{T} \\
&= 2\|x_{\hat{T}} - \hat{x}_{\hat{T}}\| + \|x\| - \|\hat{x}_{\hat{T}}\| \\
&\le 2\|x_{\hat{T}} - \hat{x}_{\hat{T}}\| + \|x\| - \|\hat{x}_T\| && \hat{T} \text{ is heaviest in } \hat{x} \\
&\le 2\|x_{\hat{T}} - \hat{x}_{\hat{T}}\| + \|x\| - \|x_T\| + \|\hat{x}_T - x_T\| && \text{by } \hat{x} \approx x \text{ on } T \\
&\le (2|\hat{T}|\alpha + 1 + |\hat{T}|\alpha)\|x_{tail(k)}\|.
\end{aligned}$$

QED.

Exer: Can you extend the above sparse recovery to ℓ_2/ℓ_2 guarantee by using CountSketch (instead of CountMin)? How many measurements would it require?

3 Application 3: Range Queries

Problem Definition: Let $x \in \mathbb{R}^n$ be the frequency vector of an input stream, and let $\varepsilon \in (0,1)$ be a parameter known in advance.

Given a range query $[i,j]$ (where $i,j \in [n]$), report an estimate for $\sum_{l=i}^{j} x_l$ that with high (constant) probability is within additive error $\varepsilon\|x\|_1$.

Observe there are $O(n^2)$ possible queries (compared with n point queries). We thus need to avoid accumulation of errors from the different coordinates.

Exer: Design a streaming algorithm for range queries with storage requirement of $O(\varepsilon^{-1} \operatorname{polylog} n)$ words.

Hint: Consider first a special case where, the range queries are restricted to the natural partition of $[1,n]$ into 2^k intervals of size $n/2^k$ each, for some $k \in \{0,\dots,\log n\}$ known in advance. For the general case, observe that every range $[i,j]$ can be partition it into $O(\log n)$ intervals as above (called dyadic intervals).

Exer: Design a heavy-hitters algorithm for *insertion-only* streams, that reports the heavy hitters faster, in time that is logarithmic (instead of linear) in n.

Hint: assume first there is only one heavy hitter, and do something like "binary search" using the dyadic intervals.