Applications of the Shannon-Hartley Theorem to Data Streams and Sparse Recovery

Eric Price
MIT

David P. Woodruff
IBM Almaden

***Abstract*—The Shannon-Hartley theorem bounds the maximum rate at which information can be transmitted over a Gaussian channel in terms of the ratio of the signal to noise power. We show two unexpected applications of this theorem in computer science: (1) we give a much simpler proof of an $\Omega(n^{1-2/p})$ bound on the number of linear measurements required to approximate the p-th frequency moment in a data stream, and show a new distribution which is hard for this problem, (2) we show that the number of measurements needed to solve the k-sparse recovery problem on an n-dimensional vector x with the C-approximate ℓ_2/ℓ_2 guarantee is $\Omega(k \log(n/k)/\log C)$. We complement this result with an almost matching $O(k \log^* k \log(n/k)/\log C)$ upper bound.**

I. Introduction

Let S be a real-valued random variable with $\mathbf{E}[S^2] = \tau^2$. Consider the random variable $S + T$, where $T \sim N(0, \sigma^2)$ is additive white Gaussian noise of variance σ^2. The Shannon-Hartley theorem states that

$$I(S; S+T) \leq \frac{1}{2} \log\left(1 + \frac{\tau^2}{\sigma^2}\right),$$

where $I(X;Y) = h(X) - h(X|Y)$ is the mutual information between X and Y, and $h(X) = -\int_{\mathbb{X}} f(x) \log f(x) dx$ is the differential entropy of a random variable X with probability density function f.

We show two unexpected applications of the Shannon-Hartley theorem in computer science, the first to estimating frequency moments in a data stream, and the second to approximating a vector by a sparse vector.

A. Sketching Frequency Moments

In the data stream literature, a line of work has considered the problem of estimating the frequency moments $F_p(x) = \|x\|_p^p = \sum_{i=1}^n |x_i|^p$, where $x \in \mathbb{R}^n$ and $p \geq 2$. One usually wants a linear sketch, that is, we choose a random matrix $A \in \mathbb{R}^{m \times n}$ from a certain distribution, for $m \ll n$, and compute Ax, from which one can output a constant-factor approximation to $F_p(x)$ with high probability. Linearity is crucial for distributed computation, formalized in the MUD (Massive Unordered Distributed) model [10]. In this model the vector x is split into pieces $x^1, \ldots, x^r$, each of which is handled by a different machine. The machines individually compute $Ax^1, \ldots, Ax^r$, and an aggregation function combines these to compute Ax and estimate $F_p(x)$. Linearity is also needed for network aggregation, which usually follows a bottom-up approach [19]: given a routing tree where the nodes represent sensors, starting from the leaves the aggregation propagates upwards to the root. We refer to the rows of A as *measurements*.

Alon, Matias, and Szegedy [3] initiated the line of work on frequency moments. There is a long line of upper bounds on the number of linear measurements; we refer the reader to the most recent works [4], [12] and the references therein. Similarly, we refer the reader to the most recent lower bounds [17], [23] and the references therein. The best upper and lower bounds for obtaining a $(1+\epsilon)$-approximation with probability at least $1-\delta$ have the form $n^{1-2/p} \cdot \text{poly}(\epsilon^{-1} \log(n\delta^{-1})$.

The existing lower bounds are rather involved, using the direct sum paradigm for information complexity [6]. Moreover, they apply to the number of bits rather than the number of linear measurements, and typically do not provide an explicit distribution which is hard. These issues can be resolved using techniques from [5], [21] and [16]. The resulting hard distribution is: choose $x \in \{-1, 0, 1\}^n$ uniformly at random, and then with probability $1/2$, replace a random coordinate x_i of x with a value in $\Theta(n^{1/p})$. $F_p(x)$ changes by a constant factor in the two cases, and so the approximation algorithm must determine which case we are in.

We instead consider the following continuous distribution: choose x to be a random $N(0, I_n)$ vector, i.e., a vector whose coordinates are independent standard normal random variables. With probability $1/2$, replace a random coordinate x_i of x with a value in $\Theta(n^{1/p})$. The use of Gaussians instead of signs allows us to derive our lower bound almost immediately from the Shannon-Hartley theorem. We obtain an $\Omega(n^{1-2/p})$ bound on the number of linear measurements required for estimating F_p, matching known bounds up to $\text{poly}(\epsilon^{-1} \log(Mn\delta^{-1})$ factors. Our proof is much simpler than previous proofs.

Our new hard distribution may also more accurately model those signals x arising in practice, since it corresponds to a signal with support 1 which is corrupted by independent Gaussian noise in each coordinate. Identifying natural hard distributions has been studied for other data stream problems, see, e.g., [20] and [18].

B. Sparse Recovery

In the field of compressed sensing, a standard problem is that of stable sparse recovery: we want a distribution $\mathcal{A}$ of matrices $A \in \mathbb{R}^{m\times n}$ such that, for any $x \in \mathbb{R}^n$ and with probability $1-\delta > 2/3$ over $A \in \mathcal{A}$, there is an algorithm to recover $\hat{x}$ from Ax with

$$\|\hat{x}-x\|_p \le (1+\epsilon) \min_{k-\text{sparse } x'} \|x-x'\|_p$$

for some $\epsilon > 0$ and norm p. We call this a $(1+\epsilon)$-approximate ℓ_p/ℓ_p recovery scheme with failure probability δ. We will focus on the popular case of $p = 2$.

For any constant $\delta > 0$ and any ϵ satisfying $\epsilon = O(1)$ and $\epsilon = \Omega(n^{-1/2})$, the optimal number of measurements is $\Theta(k \log(n/k)/\epsilon)$. The upper bound is in [13], and the lower bound is given by [1], [7], [15], [21]; see [21] for a comparison of these works.

One question is if the number of measurements can be improved when the approximation factor $C = 1+\epsilon$ is very large (i.e. $\omega(1)$). In the limiting case of $C = \infty$, corresponding to sparse recovery in the absence of noise, it is known that $O(k)$ measurements are sufficient [8]. However, the intermediate regime has not been well studied.

Using the Shannon-Hartley theorem, we prove an $\Omega(k\log(n/k)/\log C)$ lower bound on the number of measurements. We complement this with a novel sparse recovery algorithm, which builds upon [13] and [14], but is the first to obtain an improved bound for $C > 1$. Our bound is $O(k + k\log(n/k)\log^* k/\log C)$, which matches our lower bound up to a $\log^* k$ factor. Because $\log(1+\epsilon) \approx \epsilon$, these results match the $\Theta(\frac{1}{\epsilon}k\log(n/k))$ results for $\epsilon \ll 1$.

Related work. Related lower bounds have appeared in a number of recent works, including [7], [15], [1], [22], and [11]. See [21] for a comparison.

II. Lower Bound for Frequency Moments

This section is devoted to proving the following theorem:

Theorem 1: Any sketching algorithm for F_p up to a factor of $(1 \pm \epsilon)$ for $\epsilon < 1/2$, which succeeds with probability $1-\delta$ for a sufficiently small constant $\delta > 0$, requires $m = \Omega(n^{1-2/p})$.

Let $G_p = \mathbf{E}[|X|^p]$ where $X \sim N(0,1)$. For constant p, G_p is $\Theta(1)$.

Consider the following communication game between two players, Alice and Bob. Alice chooses a random $\ell \in [n]$ and associated standard unit vector $e_\ell = (0,\ldots,0,1,0,\ldots,0) \in \mathbb{R}^n$. She also chooses $w \sim N(0, I_n)$. Then she chooses $Z \in \{0,1\}$ uniformly at random. If $Z = 0$, then Alice sets $x = w$. If $Z = 1$, then Alice sets $x = (4G_p)^{1/p} \cdot n^{1/p} e_\ell + w$. She sets $y = Ax$, where A is the random matrix used for estimating F_p. She sends y to Bob, who runs the estimation procedure associated with A to recover an estimate r to $F_p(x)$. If $r \ge 2G_p n$, then Bob sets $Z' = 1$, else Bob sets $Z' = 0$. We thus have a Markov chain $\ell, Z \to x \to y \to Z'$.

If A works for any x with probability $1-\delta$, as a distribution over A, then there is a specific A and random seed such that A, together with the associated estimation procedure, succeeds with probability $1-\delta$ over x drawn from the distribution described above. Let us fix this choice of A and associated random seed, so that Alice and Bob run deterministic algorithms. Let m be the number of rows of A. We can assume the rows of A are orthonormal since this can be done in post-processing.

Lemma 2: $I(Z; Z') = O(m/n^{1-2/p})$.

Proof: Let the rows of A be denoted $v^1, \ldots, v^m$. Then we have that

$$y_i = \langle v^i, x\rangle = (4G_p)^{1/p} n^{1/p} \cdot \langle v^i, e_\ell\rangle Z + w'_i,$$

where $w'_i \sim N(0,1)$. Define $z_i = (4G_p)^{1/p} n^{1/p} \cdot \langle v^i, e_\ell\rangle Z$ so $y_i = z_i + w'_i$. Then

$$\begin{aligned}\mathbf{E}_{Z,\ell}[z_i^2] &= \frac{1}{2}\cdot(4G_p n)^{2/p}\mathbf{E}_\ell[(v^i_\ell)^2] \\ &= \frac{1}{2}\frac{(4G_p n)^{2/p}}{n} = \frac{1}{2}\frac{(4G_p)^{2/p}}{n^{1-2/p}} = \Theta(1/n^{1-2/p}).\end{aligned}$$

Hence, $y_i = z_i + w'_i$ is a Gaussian channel with power constraint $\mathbf{E}[z_i^2] = \Theta(1/n^{1-2/p})$ and noise variance $\mathbf{E}[(w'_i)^2] = 1$. By the Shannon-Hartley theorem,

$$\begin{aligned}\max_{v^i} I(z_i; y_i) &\le \frac{1}{2}\log\left(1 + \frac{\mathbf{E}[z_i^2]}{\mathbf{E}[(w'_i)^2]}\right) \\ &= \frac{1}{2}\log\left(1+\Theta(1/n^{1-2/p})\right) = \Theta(1/n^{1-2/p}).\end{aligned}$$

By the data processing inequality for Markov chains and the chain rule for entropy,

$$\begin{aligned} I(Z;Z') &\le I(z;y) = h(y) - h(y|z) \\ &= h(y) - h(y-z|z) \\ &= h(y) - \sum_i h(w'_i|z, w'_1, \ldots, w'_{i-1}) \\ &= h(y) - \sum_i h(w'_i) \le \sum_i h(y_i) - h(w'_i) \\ &= \sum_i h(y_i) - h(y_i|z_i) = \sum_i I(y_i; z_i) \\ &\le O(m/n^{1-2/p}). \end{aligned}$$

■

Proof of Theorem 1: If $Z = 1$, then $\|x\|_p^p \ge 4G_p \cdot n$, and so any $(1\pm\epsilon)$-approximation is at least $2G_p n$ for $\epsilon < 1/2$. On the other hand, if $Z = 0$, then $\mathbf{E}[\|x\|_p^p] = G_p \cdot n$, and since the $|x_i|^p$ are i.i.d. (as we range over i) with bounded variance, by Bernstein's inequality, with probability at least $1-1/n$, $\|x\|_p^p \le \frac{4}{3}\cdot G_p \cdot n$. Hence, any $(1\pm\epsilon)$-approximation is less than $2G_p n$ for $\epsilon < 1/2$. So if the algorithm succeeds with probability at least $1-\delta$, then $Z = Z'$ with probability at least $1-\delta-1/n$.

By Fano's inequality and the fact that $Z, Z' \in \{0,1\}$, if $q = \Pr[Z' \neq Z]$ then we have $H(Z|Z') \leq H(q) + q$. Hence,

$$I(Z;Z') = H(Z) - H(Z \mid Z') = 1 - (H(q) + q) \geq 1/2$$

if q is less than a sufficiently small constant, which follows from δ being a sufficiently small constant. But by Lemma 2, $I(Z;Z') = O(m/n^{1-2/p})$. Hence $m = \Omega(n^{1-2/p})$. ■

III. Bounds for Sparse Recovery

A. Lower bound for $C \gg 1$

For $C = 1+\epsilon$ a lower bound of $\Omega(k\log(n/k)/\epsilon)$ was shown in [1], [7], [15], [21] for any constant $\delta > 0$ and ϵ satisfying $\epsilon = O(1)$ and $\epsilon = \Omega(n^{-1/2})$. As in the lower bound of [21], ours uses the Shannon-Hartley theorem, but this proof is simpler because it can use that C is large. We explain the approach and our modification, and refer the reader to [21] for more details.

This section will prove the following theorem:

Theorem 3: Any C-approximate ℓ_2/ℓ_2 recovery scheme with failure probability $\delta < 1/2$ requires $m = \Omega(k\log(n/k)/\log C)$.

As in [21], let $\mathcal{F} \subset \{S \subset [n] \mid |S| = k\}$ be a family of k-sparse supports such that:

- $|S \Delta S'| \geq k$ for $S \neq S' \in \mathcal{F}$,
- $\Pr_{S\in\mathcal{F}}[i \in S] = k/n$ for all $i \in [n]$, and
- $\log|\mathcal{F}| = \Omega(k\log(n/k))$.

A random linear code on $[n/k]^k$ with relative distance $1/2$ has these properties (see discussion in [21]).

Let $X = \{x \in \{0, \pm1\}^n \mid \mathrm{supp}(x) \in \mathcal{F}\}$. Let $w \sim N(0, \alpha\frac{k}{n}I_n)$ be i.i.d. normal with variance $\alpha k/n$ in each coordinate. Consider the following process.

Alice chooses $S \in \mathcal{F}$ uniformly at random, then $x \in X$ uniformly at random subject to $\mathrm{supp}(x) = S$, then $w \sim N(0, \alpha\frac{k}{n}I_n)$. She sets $y = A(x+w)$ and sends y to Bob. Bob performs sparse recovery on y to recover $x' \approx x$, rounds to X by $\hat{x} = \arg\min_{\hat{x}\in X}\|\hat{x} - x'\|_2$, and sets $S' = \mathrm{supp}(\hat{x})$. This gives a Markov chain $S \to x \to y \to x' \to S'$.

If sparse recovery works for $x + w$ with probability $1-\delta$ over A, then there is a fixed A and random seed such that sparse recovery works with probability $1-\delta$ over $x+w$; choose this A and random seed, so that Alice and Bob run deterministic algorithms on their inputs.

The next lemma uses the Shannon-Hartley theorem.

Lemma 4: (4.1 of [21]) $I(S,S') = O(m\log(1+\frac{1}{\alpha}))$.

We modify Lemma 4.3 of [21] to obtain our main lemma and theorem. It is simpler than [21] since when C is large, the recovery algorithm cannot try to output many of the Gaussian coordinates in lieu of finding x.

Lemma 5: $I(S,S') = \Omega(k\log(n/k))$ if $\alpha = \Omega(1/C)$.

Proof: The claim is that with probability at least $1/2$, $\hat{x} = x$, and so $S = S'$. By Fano's inequality we will then have $H(S|S') \leq 1 + \Pr[S' \neq S]\log|\mathcal{F}|$, and so $I(S;S') = H(S) - H(S|S') \geq -1 + \frac{1}{2}\log|\mathcal{F}| = \Omega(k\log n/k)$.

To show the claim, we condition on successful sparse recovery, which happens with probability $1-\delta \geq 2/3$. Let $z = x + w$ be the transmitted signal. We also condition on $\|w\|_\infty^2 \leq O(\frac{\alpha k}{n}\log n)$ and $\|w\|_2^2/(\alpha k) \leq 2$, which happen with probability at least $1 - o(1)$. So both events occur with probability at least $2/3 - o(1) > 1/2$. Given this conditioning and that $\alpha = \Omega(1/C)$, the best k-sparse approximation to z is $x + w_S$, where w_S is the restriction of w to coordinates in S.

Suppose $\hat{x} \neq x$, so $\|\hat{x} - x'\|_2 \leq \|x - x'\|_2$. Then because sparse recovery was successful, $\|z - x'\|_2 \leq C\|w - w_S\|_2 \leq C\|w\|_2$. Hence

$$\begin{aligned}
\|\hat{x} - x\|_2 &\leq \|\hat{x} - x'\|_2 + \|x' - x\|_2 \\
&\leq 2\|x' - x\|_2 \\
&\leq 2(\|x' - z\|_2 + \|z - x\|_2) \\
&\leq 2(C+1)\|w\|_2 \\
&\leq 2(C+1)\sqrt{2\alpha k},
\end{aligned}$$

which is less than $\sqrt{k}$ for appropriate $\alpha = \Omega(1/C)$. This is a contradiction, and so $\hat{x} = x$, as desired. ■

Proof of Theorem 3: Combining Lemma 4 with Lemma 5, $\Omega(k\log(n/k)) = I(S,S') = O(m\log C)$, from which $m = \Omega(k\log(n/k)/\log C)$. ■

B. Upper bound for $C \gg 1$

We first focus on recovery of a single heavy coordinate. We then study recovery of 90% of the heavy hitters for general k. We conclude with recovery of all the heavy hitters.

1) k=1: We observe $2r$ measurements, for some $r = O(\log_C n)$. Let $D = C/16$. For $i \in [r]$, we choose pairwise independent hash functions $h_i\colon [n] \to [D]$ and $s_i\colon [n] \to \{\pm1\}$. We then observe

$$y_{2i} = \sum_j h_i(j)s_i(j)x_j \qquad y_{2i+1} = \sum_j s_i(j)x_j$$

```
procedure IdentifySingle($y$, $h$)
    $\alpha_i \leftarrow \mathrm{ROUND}(y_{2i}/y_{2i+1})$ for $i \in [r]$.
    $c_j \leftarrow |\{i \in [r] \mid h_i(j) = \alpha_i\}|$ for $j \in [n]$.
    $S \leftarrow \{j \in [n] \mid c_j > 5r/8\}$.
    if $|S| = 1$ then
        return $j \in S$
    else
        return $\perp$
    end if
end procedure
```

Algorithm III.1: 1-sparse identification

Define x_{-j} to equal x over $[n] \setminus \{j\}$ and 0 at j.

Lemma 6: Suppose there exists a $j^* \in [n]$ such that $|x_{j^*}| \geq C\left\|x_{-j^*}\right\|_2$. Then if C is a sufficiently large constant, we can choose $r = O(\log_C n + \log 1/\delta)$ and

$D = C/16$ so that IdentifySingle returns j^* with probability $1 - \delta$.

Proof: The key claim is that, for $\alpha_i = \text{round}(\frac{y_{2i}}{y_{2i+1}})$, we have

$$\Pr[\alpha_i \neq h_i(j^*)] \leq 1/4. \qquad (1)$$

Straightforward concentration inequalities then give the result. To get (1), define the "noise" $\beta_i = \sum_{j \neq j^*} h_i(j) s_i(j) x_j$ and $\gamma_i = \sum_{j \neq j^*} s_i(j) x_j$. Then

$$\mathbf{E}[\gamma_i^2] = \|x_{-j^*}\|_2^2 \qquad \mathbf{E}[\beta_i^2] \leq D^2 \|x_{-j^*}\|_2^2 .$$

Thus with probability at least $1 - 2/9 > 3/4$, $\gamma_i \leq 3\|x_{-j^*}\|_2$ and $\beta_i \leq 3D\|x_{-j^*}\|_2$. But then

$$\frac{y_{2i}}{y_{2i+1}} = \frac{h_i(j^*) + s_i(j^*)\beta_i/x_j}{1 + s_i(j^*)\gamma_i/x_j} = \frac{h_i(j^*) \pm 3D/C}{1 \pm 3/C}$$

$$\left|\frac{y_{2i}}{y_{2i+1}} - h_i(j^*)\right| \leq \frac{3D/C + 3h_i(j^*)/C}{1 - 3/C} \leq \frac{6D}{C-2}$$

so if $D = C/16 < (C-2)/12$, as happens for sufficiently large C, this is less than $1/2$ so $\alpha_i = \text{round}(\frac{y_{2i}}{y_{2i+1}}) = h_i(j^*)$, giving (1).

Then by a Chernoff bound, $\Pr[j^* \notin S] = \Pr[c_{j^*} < 5r/8] = e^{-\Omega(r)} < \delta/2$ for $r = \Omega(\log(1/\delta))$. Suppose that $j^* \in S$. In order for any $j \neq j^*$ to lie in S, it must have $h_i(j) = h_i(j^*)$ for at least $r/4$ different i (because both match α for $5r/8$ coordinates). But $\Pr[h_i(j) = h_i(j^*)] = 1/D$ independently over i, so

$$\Pr[j \in S] \leq \binom{r}{r/4}(1/D)^{r/4} \leq (4e/D)^{r/4} = C^{-\Omega(r)}$$

as long as C is larger than a fixed constant. But for $r = O(\log_C(n/\delta))$ this gives $\Pr[j \in S] < \delta/(2n)$, so a union bound gives that $S = \{j\}$ with probability $1 - \delta$. ■

2) General k, finding most coordinates: For general k, we identify a set L of $O(k)$ coordinates by partitioning the coordinates into $O(k)$ sets of size $\Theta(n/k)$ and applying IdentifySingle. To be specific, we use a pairwise independent hash function $h\colon [n] \to [l]$ to partition into l sets.

To analyze how this performs, define the "error"

$$\text{Err}^2(x, k) = \min_{k\text{-sparse } x'} \|x - x'\|_2^2$$

and the "heavy hitters"

$$S = \{i \in [n] \mid |x_i|^2 > \frac{C^2}{k}\text{Err}^2(x, k)\}.$$

Lemma 7: With $O(k \log_C(n/k))$ measurements, this algorithm returns a set L of size $O(k)$ such that each $j \in S$ has $j \in L$ with probability at least $3/4$.

Proof: For each coordinate $j \in S$, Lemma 6 shows it will be recovered as long as three events hold: none of the other elements of the top k coordinates hash to the same value as j, the ℓ_2^2 norm of the mass that hashes to the same value as j is no more than a constant factor times its expectation $\text{Err}^2(x, k)/k$, and the algorithm in Lemma 6 does not fail. All these occur with constant probability if $l = O(k)$, giving the result. ■

Corollary 8: With $O(k \log_C(n/k) \log(1/\delta))$ measurements, IdentifyMost returns a set L of size $O(k)$ such that each $j \in S$ has $j \in L$ with probability at least $1 - \delta$.

Proof: We repeat the method of Lemma 7 $O(\log(1/\delta))$ times, and take all coordinates that are listed in more than half the sets L_i. This at most doubles the output size, and by a Chernoff bound each $j \in S$ lies in the output with probability at least $1 - \delta$. ■

Corollary 8 gives a good method for finding the heavy hitters, but we also need to estimate them.

3) Estimating coordinates: We estimate using Count-Sketch [9], with $R = O(\log(1/\delta))$ hash tables of size $O(k/\epsilon)$.

```
procedure IdentifyMost(y)
    for $r \leftarrow [R]$ do                ▷ $R = O(\log(1/\delta))$
        $L_r \leftarrow \{\text{IdentifySingle}(y^{(i)}) \mid i \in [k]\}$
    end for
    $c_j \leftarrow |\{r \mid j \in L_r\}|$ for $j \in [n]$.
    $L \leftarrow \{j \mid c_j > R/2\}$
    return $\widehat{x}_L$
end procedure
procedure EstimateMost(y, L)
    for $r \leftarrow [R]$ do                ▷ $R = O(\log(1/\delta))$
        $\widehat{x}_j^{(r)} \leftarrow s(j) y_{h(j)}$.
    end for
    $\widehat{x}_j \leftarrow \text{median}_r\, \widehat{x}_j^{(r)}$
    return $\widehat{x}_L$
end procedure
```

Algorithm III.2: Estimating most coordinates well

Lemma 9: Suppose $|L| \leq O(k)$. With $O(\frac{1}{\epsilon} k \log(\frac{1}{f\delta}))$ measurements, EstimateMost returns $\widehat{x}_L$ so that for any j, with probability $1 - \delta$ we have

$$\text{Err}^2(x_L - \widehat{x}_L, fk) \leq \epsilon \text{Err}^2(x, k)$$

Proof: The analysis of Count-Sketch [9] gives

$$\Pr[|\widehat{x}_j - x_j|^2 > \frac{\epsilon}{k}\text{Err}^2(x, k)] < f\delta.$$

Thus with probability $1 - \delta$, at most $f|L| = O(fk)$ of the j have $|\widehat{x}_j - x_j|^2 > \frac{\epsilon}{k}\text{Err}^2(x, k)$. Rescaling f and ϵ gives the result. ■

4) Recovering all the heavy hitters:

Lemma 10: The result $\widehat{x}_L$ of IdentifyMost followed by EstimateMost satisfies

$$\text{Err}^2(x - \widehat{x}_L, fk) \leq C^2 \text{Err}^2(x, k)$$

with probability $1 - \delta$, and uses $O(k \log_C(n/k) \log(\frac{1}{f\delta}))$ measurements.

Proof: Let T contain the largest k coordinates of x. By Corollary 8, each $j \in S$ has $j \in L$ with probability

```
procedure RecoverAll(y)
    $k' \leftarrow k$, $\delta \leftarrow 1/16$, $\widehat{x}^{(0)} \leftarrow 0$
    for $r \leftarrow [R]$ do
        $y' \leftarrow y^{(r)} - A^{(r)}\widehat{x}^{(r)}$
        $L^{(r)} \leftarrow$ IdentifyMost$(y', k', \delta)$
        $\widehat{v}^{(r)} \leftarrow$ EstimateMost$(y', k', \delta, L)$
        $\widehat{x}^{(r+1)} \leftarrow \widehat{x}^{(r)} + \widehat{v}^{(r)}$
        Decrease $k', \epsilon, \delta$ per Theorem 11
    end for
end procedure
```

Algorithm III.3: Recovering all coordinates

$1 - \delta f$, so with probability $1 - \delta$ we have $|S \setminus L| \le fk$. Then

$$\begin{aligned}\mathrm{Err}^2(x - \widehat{x}_L, 2fk) &\le \mathrm{Err}^2(x_L - \widehat{x}_L, fk) + \left\|x_{[n]\setminus(S\cup L)}\right\|_2^2 \\ &\le \epsilon \mathrm{Err}^2(x, k) + \left\|x_{[n]\setminus T}\right\|_2^2 + \left\|x_{T\setminus S}\right\|_2^2 \\ &\le (\epsilon + 1 + C^2)\mathrm{Err}^2(x, k) \\ &\le 2C^2 \mathrm{Err}^2(x, k)\end{aligned}$$

with probability $1 - \delta$ by Lemma 9. Rescale f, δ, and C to get the result. ■

Theorem 11: RecoverAll achieves C-approximate ℓ_2/ℓ_2 sparse recovery with $O(k + (\log^* k)k \log_C(n/k))$ measurements and $3/4$ success probability.

Proof: We will achieve $D^{O(\log^* k)}$-approximate recovery using $O(k \log_D(n/k))$ measurements. Substituting $\log C = \log D \log^* k$ gives the result.

Define $\delta_i = \frac{1}{8\cdot 2^i}$. Let $f_0 = 1/16$ and $f_{i+1} = 2^{-1/(4^i f_i)}$. Let $k_i = k\prod_{j<i} f_j$. Then for $R = O(\log^* k)$, $k_R < 1$.

We set $\widehat{x}^{(0)} = 0$, and iterate IdentifyMost and EstimateMost on $x - \widehat{x}^{(r)}$ in each round r with δ_r, f_r, k_r, D as parameters, getting update $\widehat{v}^{(r)}$ and setting $\widehat{x}^{(r+1)} = \widehat{x}^{(r)} + \widehat{v}^{(r)}$.

Then Lemma 10 telescopes, giving

$$\mathrm{Err}^2(x - \widehat{x}^{(r)}, k_r) \le D^{2r}\mathrm{Err}^2(x, k)$$

so $\left\|x - \widehat{x}^{(R)}\right\|_2^2 \le D^{2R}\mathrm{Err}^2(x, k)$, which is $D^{O(\log^* k)}$-approximate recovery.

The total number of measurements is

$$\begin{aligned}&\sum_{i=0}^{R} k_i \log_D(n/k_i)\log(\frac{1}{\delta_i f_i}) \\ =&\sum_{i=0}^{R} k(\prod_{j<i} f_j)\log_D(\frac{n}{k}\prod_{j<i}(1/f_j))\frac{3+i}{4^i f_{i-1}} \\ =&\sum_{i=0}^{R} k\frac{3+i}{4^i}(\prod_{j<i-1} f_j)\log_D(\frac{n}{k}\prod_{j<i}(1/f_j)) \\ =&O(k\log_D\frac{n}{k}) + \frac{k}{\log D}\sum_{i=0}^{R}\frac{3+i}{4^i}(\prod_{j<i-1} f_j)\sum_{j<i}\frac{1}{4^j f_{j-1}} \\ =&O(k\log_D\frac{n}{k}).\end{aligned}$$

■

Acknowledgements

E. Price is supported in part by an NSF Graduate Research Fellowship.

References

[1] S. Aeron, V. Saligrama, and M. Zhao. Information theoretic bounds for compressed sensing. *Information Theory, IEEE Transactions on*, 56(10):5111–5130, 2010.

[2] Mehmet Akçakaya and Vahid Tarokh. A frame construction and a universal distortion bound for sparse representations. *IEEE Transactions on Signal Processing*, 56(6):2443–2450, 2008.

[3] Noga Alon, Yossi Matias, and Mario Szegedy. The Space Complexity of Approximating the Frequency Moments. *JCSS*, 58(1):137–147, 1999.

[4] Alexandr Andoni, Robert Krauthgamer, and Krzysztof Onak. Streaming algorithms via precision sampling. In *FOCS*, pages 363–372, 2011.

[5] Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In *SODA*, pages 1190–1197, 2010.

[6] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *JCSS*, 68(4):702–732, 2004.

[7] E.J. Candès and M.A. Davenport. How well can we estimate a sparse vector? *Arxiv preprint arXiv:1104.5246*, 2011.

[8] E.J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

[9] Moses Charikar, Kevin Chen, and Martin Farach-Colton. Finding frequent items in data streams. In *ICALP*, pages 693–703, 2002.

[10] Jon Feldman, S. Muthukrishnan, Anastasios Sidiropoulos, Clifford Stein, and Zoya Svitkina. On distributing symmetric streaming computations. *ACM Transactions on Algorithms*, 6(4), 2010.

[11] Alyson K. Fletcher, Sundeep Rangan, and Vivek K. Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Transactions on Information Theory*, 55(12):5758–5772, 2009.

[12] Sumit Ganguly. Polynomial estimators for high frequency moments. *CoRR*, abs/1104.4552, 2011.

[13] Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. Approximate sparse recovery: optimizing time and measurements. In *STOC*, pages 475–484, 2010.

[14] Piotr Indyk, Eric Price, and David P. Woodruff. On the power of adaptivity in sparse recovery. In *FOCS*, pages 285–294, 2011.

[15] MA Iwen and AH Tewfik. Adaptive group testing strategies for target detection and localization in noisy environments. *IMA Preprint Series*, (2311), 2010.

[16] T. S. Jayram. Unpublished manuscript. 2002.

[17] T. S. Jayram. Hellinger strikes back: A note on the multi-party information complexity of and. In *APPROX-RANDOM*, pages 562–573, 2009.

[18] Ravi Kumar and Rina Panigrahy. On finding frequent elements in a data stream. In *APPROX-RANDOM*, pages 584–595, 2007.

[19] Samuel Madden, Michael J. Franklin, Joseph M. Hellerstein, and Wei Hong. Tag: A tiny aggregation service for ad-hoc sensor networks. In *OSDI*, 2002.

[20] Rajeev Motwani and Sergei Vassilvitskii. Distinct value estimators in power law distributions. ANALCO, 2006.

[21] Eric Price and David P. Woodruff. (1 + eps)-approximate sparse recovery. In *FOCS*, pages 295–304, 2011.

[22] Martin J. Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE Transactions on Information Theory*, 55(12):5728–5741, 2009.

[23] David P. Woodruff and Qin Zhang. Tight bounds for distributed functional monitoring. *CoRR*, abs/1112.5153, 2011.