Lower Bounds for Adaptive Sparse Recovery

Eric Price
MIT
ecprice@mit.edu

David P. Woodruff
IBM Almaden
dpwoodru@us.ibm.com

Abstract

We give lower bounds for the problem of stable sparse recovery from *adaptive* linear measurements. In this problem, one would like to estimate a vector $x \in \mathbb{R}^n$ from m linear measurements $A_1 x, \ldots, A_m x$. One may choose each vector A_i based on $A_1 x, \ldots, A_{i-1} x$, and must output $\hat{x}$ satisfying

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

with probability at least $1 - \delta > 2/3$, for some $p \in \{1, 2\}$. For $p = 2$, it was recently shown that this is possible with $m = O(\frac{1}{\epsilon} k \log \log(n/k))$, while nonadaptively it requires $\Theta(\frac{1}{\epsilon} k \log(n/k))$. It is also known that even adaptively, it takes $m = \Omega(k/\epsilon)$ for $p = 2$. For $p = 1$, there is a non-adaptive upper bound of $\widetilde{O}(\frac{1}{\sqrt{\epsilon}} k \log n)$. We show:

- For $p = 2$, $m = \Omega(\log \log n)$. This is tight for $k = O(1)$ and constant ϵ, and shows that the $\log \log n$ dependence is correct.

- If the measurement vectors are chosen in R "rounds", then $m = \Omega(R \log^{1/R} n)$. For constant ϵ, this matches the previously known upper bound up to an $O(1)$ factor in R.

- For $p = 1$, $m = \Omega(k/(\sqrt{\epsilon} \cdot \log k/\epsilon))$. This shows that adaptivity cannot improve more than logarithmic factors, providing the analogue of the $m = \Omega(k/\epsilon)$ bound for $p = 2$.

1 Introduction

Compressed sensing or *sparse recovery* studies the problem of solving underdetermined linear systems subject to a sparsity constraint. It has applications to a wide variety of fields, including data stream algorithms [Mut05], medical or geological imaging [CRT06, Don06], and genetics testing [SAZ10]. The approach uses the power of a *sparsity* constraint: a vector x' is *k-sparse* if at most k coefficients are non-zero. A standard formulation for the 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

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

for some parameter $\epsilon > 0$ and norm p. We refer to the elements of Ax as *measurements*. We say Equation (1) denotes ℓ_p/ℓ_p *recovery*.

The goal is to minimize the number of measurements while still allowing efficient recovery of x. This problem has recently been largely closed: for $p = 2$, it is known that $m = \Theta(\frac{1}{\epsilon} k \log(n/k))$ is tight (upper bounds in [CRT06, GLPS10], lower bounds in [PW11, CD11]), and for $p = 1$ it is known that $m = \widetilde{O}(\frac{1}{\sqrt{\epsilon}} k \log n)$ and $m = \widetilde{\Omega}(\frac{k}{\sqrt{\epsilon}})$ [PW11] (recall that $\widetilde{O}(f)$ means $O(f \log^c f)$ for some constant c, and similarly $\widetilde{\Omega}(f)$ means $\Omega(f/\log^c f)$).

In order to further reduce the number of measurements, a number of recent works have considered making the measurements *adaptive* [JXC08, CHNR08, HCN09, HBCN09, MSW08, AWZ08, IPW11]. In this setting, one may choose each row of the matrix after seeing the results of previous measurements. More generally, one may split the adaptivity into R "rounds", where in each round r one chooses $A^r \in \mathbb{R}^{m_r \times n}$ based on $A^1 x, \ldots, A^{r-1} x$. At the end, one must use $A^1 x, \ldots, A^R x$ to output $\hat{x}$ satisfying Equation (1). We would still like to minimize the total number of measurements $m = \sum m_i$. In the $p = 2$ setting, it is known that for arbitrarily many rounds $O(\frac{1}{\epsilon} k \log \log(n/k))$ measurements suffice, and for $O(r \log^* k)$ rounds $O(\frac{1}{\epsilon} kr \log^{1/r}(n/k))$ measurements suffice [IPW11].

Given these upper bounds, two natural questions arise: first, is the improvement in the dependence on n from $\log(n/k)$ to $\log \log(n/k)$ tight, or can the improvement be strengthened? Second, can adaptivity help by more than a logarithmic factor, by improving the dependence on k or ϵ?

A recent lower bound showed that $\Omega(k/\epsilon)$ measurements are necessary in a setting essentially equivalent to the $p = 2$ case [ACD11][1]. Thus, they answer the second question in the negative for $p = 2$. Their techniques rely on special properties of the 2-norm; namely, that it is a rotationally invariant inner product space and that the Gaussian is both 2-stable and a maximum entropy distribution. Such techniques do not seem useful for proving lower bounds for $p = 1$.

Our results. For $p = 2$, we show that any adaptive sparse recovery scheme requires $\Omega(\log \log n)$ measurements, or $\Omega(R \log^{1/R} n)$ measurements given only R rounds. For $k = O(1)$, this matches the upper bound of [IPW11] up to an $O(1)$ factor in R. It thus shows that the $\log \log n$ term in the adaptive bound is necessary.

For $p = 1$, we show that any adaptive sparse recovery scheme requires $\widetilde{\Omega}(k/\sqrt{\epsilon})$ measurements. This shows that adaptivity can only give polylog(n) improvements, even for $p = 1$. Additionally, our bound of $\Omega(k/(\sqrt{\epsilon} \cdot \log(k/\sqrt{\epsilon})))$ improves the previous *non-adaptive* lower bound for $p = 1$ and small ϵ, which lost an additional $\log k$ factor [PW11].

---

[1] Both our result and their result apply in both settings. See Appendix A for a more detailed discussion of the relationship between the two settings.

Related work. Our work draws on the lower bounds for non-adaptive sparse recovery, most directly [PW11].

The main previous lower bound for adaptive sparse recovery gets $m = \Omega(k/\epsilon)$ for $p = 2$ [ACD11]. They consider going down a similar path to our $\Omega(\log \log n)$ lower bound, but ultimately reject it as difficult to bound in the adaptive setting. Combining their result with ours gives a $\Omega(\frac{1}{\epsilon}k + \log \log n)$ lower bound, compared with the $O(\frac{1}{\epsilon}k \cdot \log \log n)$ upper bound. The techniques in their paper do not imply any bounds for the $p = 1$ setting.

For $p = 2$ in the special case of adaptive Fourier measurements (where measurement vectors are adaptively chosen from among n rows of the Fourier matrix), [HIKP12] shows $\Omega(k \log(n/k)/\log \log n)$ measurements are necessary. In this case the main difficulty with lower bounding adaptivity is avoided, because all measurement rows are chosen from a small set of vectors with bounded ℓ_∞ norm; however, some of the minor issues in using [PW11] for an adaptive bound were dealt with there.

Our techniques. We use very different techniques for our two bounds.

To show $\Omega(\log \log n)$ for $p = 2$, we reduce to the information capacity of a Gaussian channel. We consider recovery of the vector $x = e_{i^*} + w$, for $i^* \in [n]$ uniformly and $w \sim N(0, I_n/\Theta(n))$. Correct recovery must find i^*, so the mutual information $I(i^*; Ax)$ is $\Omega(\log n)$. On the other hand, in the nonadaptive case [PW11] showed that each measurement $A_j x$ is a power-limited Gaussian channel with constant signal-to-noise ratio, and therefore has $I(i^*; A_j x) = O(1)$. Linearity gives that $I(i^*; Ax) = O(m)$, so $m = \Omega(\log n)$ in the nonadaptive case. In the adaptive case, later measurements may "align" the row A_j with i^*, to increase the signal-to-noise ratio and extract more information—this is exactly how the upper bounds work. To deal with this, we bound how much information we can extract as a function of how much we know about i^*. In particular, we show that given a small number b bits of information about i^*, the posterior distribution of i^* remains fairly well "spread out". We then show that any measurement row A_j can only extract $O(b + 1)$ bits from such a spread out distribution on i^*. This shows that the information about i^* increases at most exponentially, so $\Omega(\log \log n)$ measurements are necessary.

To show an $\widetilde{\Omega}(k/\sqrt{\epsilon})$ bound for $p = 1$, we first establish a lower bound on the multiround distributional communication complexity of a two-party communication problem that we call $\mathsf{Multi}\ell_\infty$, for a distribution tailored to our application. We then show how to use an adaptive $(1+\epsilon)$-approximate ℓ_1/ℓ_1 sparse recovery scheme $\mathcal{A}$ to solve the communication problem $\mathsf{Multi}\ell_\infty$, modifying the general framework of [PW11] for connecting non-adaptive schemes to communication complexity in order to now support adaptive schemes. By the communication lower bound for $\mathsf{Multi}\ell_\infty$, we obtain a lower bound on the number of measurements required of $\mathcal{A}$.

In the $\mathsf{Gap}\ell_\infty$ problem, the two players are given x and y respectively, and they want to approximate $\|x - y\|_\infty$ given the promise that all entries of $x - y$ are small in magnitude or there is a single large entry. The $\mathsf{Multi}\ell_\infty$ problem consists of solving multiple independent instances of $\mathsf{Gap}\ell_\infty$ in parallel. Intuitively, the sparse recovery algorithm needs to determine if there are entries of $x - y$ that are large, which corresponds to solving multiple instances of $\mathsf{Gap}\ell_\infty$. We prove a multiround direct sum theorem for a distributional version of $\mathsf{Gap}\ell_\infty$, thereby giving a distributional lower bound for $\mathsf{Multi}\ell_\infty$. A direct sum theorem for $\mathsf{Gap}\ell_\infty$ has been used before for proving lower bounds for non-adaptive schemes [PW11], but was limited to a bounded number of rounds due to the use of a bounded round theorem in communication complexity [BR11]. We instead use the information complexity framework [BJKS04] to lower bound the conditional mutual information between the inputs to $\mathsf{Gap}\ell_\infty$ and the transcript of any correct protocol for $\mathsf{Gap}\ell_\infty$ under a certain input distribution, and prove a direct sum theorem for solving k instances of this problem. We

need to condition on "help variables" in the mutual information which enable the players to embed single instances of $\mathsf{Gap}\ell_\infty$ into $\mathsf{Multi}\ell_\infty$ in a way in which the players can use a correct protocol on our input distribution for $\mathsf{Multi}\ell_\infty$ as a correct protocol on our input distribution for $\mathsf{Gap}\ell_\infty$; these help variables are in addition to help variables used for proving lower bounds for $\mathsf{Gap}\ell_\infty$, which is itself proved using information complexity. We also look at the conditional mutual information with respect to an input distribution which doesn't immediately fit into the information complexity framework. We relate the conditional information of the transcript with respect to this distribution to that with respect to a more standard distribution.

2   Notation

We use lower-case letters for fixed values and upper-case letters for random variables. We use $\log x$ to denote $\log_2 x$, and $\ln x$ to denote $\log_e x$. For a discrete random variable X with probability p, we use $H(X)$ or $H(p)$ to denote its entropy

$$H(X) = H(p) = \sum -p(x) \log p(x).$$

For a continuous random variable X with pdf p, we use $h(X)$ to denote its differential entropy

$$h(X) = \int_{x \in X} -p(x) \log p(x) dx.$$

Let y be drawn from a random variable Y. Then $(X \mid y) = (X \mid Y = y)$ denotes the random variable X conditioned on $Y = y$. We define $h(X \mid Y) = \mathbb{E}_{y \sim Y} h(X \mid y)$. The mutual information between X and Y is denoted $I(X;Y) = h(X) - h(X \mid Y)$.

For $p \in \mathbb{R}^n$ and $S \subseteq [n]$, we define $p_S \in \mathbb{R}^n$ to equal p over indices in S and zero elsewhere.

We use $f \lesssim g$ to denote $f = O(g)$.

3   Tight lower bound for $p = 2, k = 1$

We may assume that the measurements are orthonormal, since this can be performed in post-processing of the output, by multiplying Ax on the left to orthogonalize A. We will give a lower bound for the following instance:

Alice chooses random $i^* \in [n]$ and i.i.d. Gaussian noise $w \in \mathbb{R}^n$ with $\mathbb{E}[\|w\|_2^2] = \sigma^2 = \Theta(1)$, then sets $x = e_{i^*} + w$. Bob performs R rounds of adaptive measurements on x, getting $y^r = A^r x = (y_1^r, \ldots, y_{m_r}^r)$ in each round r. Let I^* and Y^r denote the random variables from which i^* and y^r are drawn, respectively. We will bound $I(I^*; Y^1, Y^2, \ldots, Y^r)$.

We may assume Bob is deterministic, since we are giving a lower bound for a distribution over inputs – for any randomized Bob that succeeds with probability $1 - \delta$, there exists a choice of random seed such that the corresponding deterministic Bob also succeeds with probability $1 - \delta$.

First, we give a bound on the information received from any single measurement, depending on Bob's posterior distribution on I^* at that point:

Lemma 3.1. *Let I^* be a random variable over $[n]$ with probability distribution $p_i = \Pr[I^* = i]$, and define*

$$b = \sum_{i=1}^{n} p_i \log(np_i).$$

Define $X = e_{I^*} + N(0, I_n \sigma^2/n)$. Consider any fixed vector $v \in \mathbb{R}^n$ independent of X with $\|v\|_2 = 1$, and define $Y = v \cdot X$. Then

$$I(v_{I^*}; Y) \le C(b+1)$$

for some constant C.

Proof. Let $S_i = \{j \mid 2^i \le np_j < 2^{i+1}\}$ for $i > 0$ and $S_0 = \{i \mid np_i < 2\}$. Define $t_i = \sum_{j \in S_i} p_j = \Pr[I^* \in S_i]$. Then

$$\begin{aligned}
\sum_{i=0}^{\infty} it_i &= \sum_{i>0} \sum_{j \in S_i} p_j \cdot i \\
&\le \sum_{i>0} \sum_{j \in S_i} p_j \log(np_j) \\
&= b - \sum_{j \in S_0} p_j \log(np_j) \\
&\le b - t_0 \log(nt_0/|S_0|) \\
&\le b + |S_0|/(ne)
\end{aligned}$$

using convexity and minimizing $x \log ax$ at $x = 1/(ae)$. Hence

$$\sum_{i=0}^{\infty} it_i < b + 1 \tag{2}$$

Let $W = N(0, \sigma^2/n)$. For any measurement vector v, let $Y = v \cdot X \sim v_{I^*} + W$. Let $Y_i = (Y \mid I^* \in S_i)$. Because $\sum v_j^2 = 1$,

$$\mathbb{E}[Y_i^2] = \sigma^2/n + \sum_{j \in S_i} v_j^2 p_j/t_i \le \sigma^2/n + \|p_{S_i}\|_\infty /t_i \le \sigma^2/n + 2^{i+1}/(nt_i). \tag{3}$$

Let T be the (discrete) random variable denoting the i such that $I^* \in S_i$. Then Y is drawn from Y_T, and T has probability distribution t. Hence

$$\begin{aligned}
h(Y) &\le h((Y, T)) \\
&= H(T) + h(Y_T \mid T) \\
&= H(t) + \sum_{i \ge 0} t_i h(Y_i) \\
&\le H(t) + \sum_{i \ge 0} t_i h(N(0, \mathbb{E}[Y_i^2]))
\end{aligned}$$

because the Gaussian distribution maximizes entropy subject to a power constraint. Using the same technique as the Shannon-Hartley theorem,

$$\begin{aligned}
I(v_{I^*}, Y) = I(v_{I^*}; v_{I^*} + W) &= h(v_{I^*} + W) - h(v_{I^*} + W \mid v_{I^*}) \\
&= h(Y) - h(W) \\
&\le H(t) + \sum_{i \ge 0} t_i(h(N(0, \mathbb{E}[Y_i^2])) - h(W)) \\
&= H(t) + \frac{1}{2} \sum_{i \ge 0} t_i \ln(\frac{\mathbb{E}[Y_i^2]}{\mathbb{E}[W^2]})
\end{aligned}$$

4

and hence by Equation (3),

$$I(v_{I^*}; Y) \le H(t) + \frac{\ln 2}{2} \sum_{i \ge 0} t_i \log(1 + \frac{2^{i+1}}{t_i \sigma^2}). \tag{4}$$

All that requires is to show that this is $O(1 + b)$. Since $\sigma = \Theta(1)$, we have

$$\sum_i t_i \log(1 + \frac{2^i}{\sigma^2 t_i}) \le \log(1 + 1/\sigma^2) + \sum_i t_i \log(1 + \frac{2^i}{t_i})$$

$$\le O(1) + \sum_i t_i \log(1 + 2^i) + \sum_i t_i \log(1 + 1/t_i). \tag{5}$$

Now, $\log(1 + 2^i) \lesssim i$ for $i > 0$ and is $O(1)$ for $i = 0$, so by Equation (2),

$$\sum_i t_i \log(1 + 2^i) \lesssim 1 + \sum_{i > 0} i t_i < 2 + b.$$

Next, $\log(1 + 1/t_i) \lesssim \log(1/t_i)$ for $t_i \le 1/2$, so

$$\sum_i t_i \log(1 + 1/t_i) \lesssim \sum_{i | t_i \le 1/2} t_i \log(1/t_i) + \sum_{i | t_i > 1/2} 1 \le H(t) + 1.$$

Plugging into Equations (5) and (4),

$$I(v_{I^*}, Y) \lesssim 1 + b + H(t). \tag{6}$$

To bound $H(t)$, we consider the partition $T_+ = \{i \mid t_i > 1/2^i\}$ and $T_- = \{i \mid t_i \le 1/2^i\}$. Then

$$H(t) = \sum_i t_i \log(1/t_i)$$

$$\le \sum_{i \in T_+} i t_i + \sum_{t \in T_-} t_i \log(1/t_i)$$

$$\le 1 + b + \sum_{t \in T_-} t_i \log(1/t_i)$$

But $x \log(1/x)$ is increasing on $[0, 1/e]$, so

$$\sum_{t \in T_-} t_i \log(1/t_i) \le t_0 \log(1/t_0) + t_1 \log(1/t_1) + \sum_{i \ge 2} \frac{1}{2^i} \log(1/2^i) \le 2/e + 3/2 = O(1)$$

and hence $H(t) \le b + O(1)$. Combining with Equation (6) gives that

$$I(v_{I^*}; Y) \lesssim b + 1$$

as desired. $\qquad \square$

Theorem 3.2. *Any scheme using R rounds with number of measurements $m_1, m_2, \ldots, m_R > 0$ in each round has*

$$I(I^*; Y^1, \ldots, Y^R) \le C^R \prod_i m_i$$

for some constant $C > 1$.

5

Proof. Let the signal in the absence of noise be $Z^r = A^r e_{I^*} \in \mathbb{R}^{m_r}$, and the signal in the presence of noise be $Y^r = A^r(e_{I^*} + N(0, \sigma^2 I_n/n)) = Z^r + W^r$ where $W^r = N(0, \sigma^2 I_{m_r}/n)$ independently. In round r, after observations $y^1, \ldots, y^{r-1}$ of $Y^1, \ldots, Y^{r-1}$, let p^r be the distribution on $(I^* \mid y^1, \ldots, y^{r-1})$. That is, p^r is Bob's posterior distribution on I^* at the beginning of round r.

We define

$$
\begin{aligned}
b_r &= H(I^*) - H(I^* \mid y^1, \ldots, y^{r-1}) \\
&= \log n - H(p^r) \\
&= \sum p_i^r \log(n p_i^r).
\end{aligned}
$$

Because the rows of A^r are deterministic given $y^1, \ldots, y^{r-1}$, Lemma 3.1 shows that any single measurement $j \in [m_r]$ satisfies

$$
I(Z_j^r; Y_j^r \mid y^1, \ldots, y^{r-1}) \le C(b_r + 1).
$$

for some constant C. Thus by Lemma B.1

$$
I(Z^r; Y^r \mid y^1, \ldots, y^{r-1}) \le C m_r(b_r + 1).
$$

There is a Markov chain $(I^* \mid y^1, \ldots, y^{r-1}) \to (Z^r \mid y^1, \ldots, y^{r-1}) \to (Y^r \mid y^1, \ldots, y^{r-1})$, so

$$
I(I^*; Y^r \mid y^1, \ldots, y^{r-1}) \le I(Z^r; Y^r \mid y^1, \ldots, y^{r-1}) \le C m_r(b_r + 1).
$$

We define $B_r = I(I^*; Y^1, \ldots, Y^{r-1}) = \mathbb{E}_y \, b_r$. Therefore

$$
\begin{aligned}
B_{r+1} &= I(I^*; Y^1, \ldots, Y^r) \\
&= I(I^*; Y^1, \ldots, Y^{r-1}) + I(I^*; Y^r \mid Y^1, \ldots, Y^{r-1}) \\
&= B_r + \mathop{\mathbb{E}}_{y^1, \ldots, y^{r-1}} I(I^*; Y^r \mid y^1, \ldots, y^{r-1}) \\
&\le B_r + C m_r \mathop{\mathbb{E}}_{y^1, \ldots, y^{r-1}} (b_r + 1) \\
&= (B_r + 1)(C m_r + 1) - 1 \\
&\le C' m_r(B_r + 1)
\end{aligned}
$$

for some constant C'. Then for some constant $D \ge C'$,

$$
I(I^*; Y^1, \ldots, Y^R) = B_{R+1} \le D^R \prod_i m_i
$$

as desired. $\qquad \square$

Corollary 3.3. *Any scheme using R rounds with m measurements has*

$$
I(I^*; Y^1, \ldots, Y^R) \le (Cm/R)^R
$$

for some constant C. Thus for sparse recovery, $m = \Omega(R \log^{1/R} n)$. Minimizing over R, we find that $m = \Omega(\log \log n)$ independent of R.

Proof. The equation follows from the AM-GM inequality. Furthermore, our setup is such that Bob can recover I^* from Y with large probability, so $I(I^*; Y) = \Omega(\log n)$; this was formally shown in Lemma 6.3 of [HIKP12] (modifying Lemma 4.3 of [PW11] to adaptive measurements and $\epsilon = \Theta(1)$). The result follows. $\qquad \square$

4 Lower bound for dependence on k and ϵ for ℓ_1/ℓ_1

In Section 4.1 we establish a new lower bound on the communication complexity of a two-party communication problem that we call $\mathsf{Multi}\ell_\infty$. In Section 4.2 we then show how to use an adaptive $(1+\epsilon)$-approximate ℓ_1/ℓ_1 sparse recovery scheme $\mathcal{A}$ to solve the communication problem $\mathsf{Multi}\ell_\infty$. By the communication lower bound in Section 4.1, we obtain a lower bound on the number of measurements required of $\mathcal{A}$.

4.1 Direct sum for distributional ℓ_∞

We assume basic familiarity with communication complexity; see the textbook of Kushilevitz and Nisan [KN97] for further background. Our reason for using communication complexity is to prove lower bounds, and we will do so by using information-theoretic arguments. We refer the reader to the thesis of Bar-Yossef [Bar02] for a comprehensive introduction to information-theoretic arguments used in communication complexity.

We consider two-party randomized communication complexity. There are two parties, Alice and Bob, with input vectors x and y respectively, and their goal is to solve a promise problem $f(x,y)$. The parties have private randomness. The communication cost of a protocol is its maximum transcript length, over all possible inputs and random coin tosses. The randomized communication complexity $R_\delta(f)$ is the minimum communication cost of a randomized protocol Π which for every input (x,y) outputs $f(x,y)$ with probability at least $1-\delta$ (over the random coin tosses of the parties). We also study the distributional complexity of f, in which the parties are deterministic and the inputs (x,y) are drawn from distribution μ, and a protocol is correct if it succeeds with probability at least $1-\delta$ in outputting $f(x,y)$, where the probability is now taken over $(x,y)\sim\mu$. We define $D_{\mu,\delta}(f)$ to be the minimum communication cost of a correct protocol Π.

We consider the following promise problem $\mathsf{Gap}\ell_\infty^B$, where B is a parameter, which was studied in [SS02, BJKS04]. The inputs are pairs (x,y) of m-dimensional vectors, with $x_i, y_i \in \{0,1,2,\ldots,B\}$ for all $i \in [m]$, with the promise that (x,y) is one of the following types of instance:

- NO instance: for all i, $|x_i - y_i| \in \{0,1\}$, or

- YES instance: there is a unique i for which $|x_i - y_i| = B$, and for all $j \neq i$, $|x_j - y_j| \in \{0,1\}$.

The goal of a protocol is to decide which of the two cases (NO or YES) the input is in.

Consider the distribution σ: for each $j \in [m]$, choose a random pair $(Z_j, P_j) \in \{0,1,2,\ldots,B\} \times \{0,1\} \setminus \{(0,1),(B,0)\}$. If $(Z_j, P_j) = (z,0)$, then $X_j = z$ and Y_j is uniformly distributed in $\{z, z+1\}$; if $(Z_j, P_j) = (z,1)$, then $Y_j = z$ and X_j is uniformly distributed on $\{z-1,z\}$. Let $Z = (Z_1, \ldots, Z_m)$ and $P = (P_1, \ldots, P_m)$. Next choose a random coordinate $S \in [m]$. For coordinate S, replace (X_S, Y_S) with a uniform element of $\{(0,0),(0,B)\}$. Let $X = (X_1, \ldots, X_m)$ and $Y = (Y_1, \ldots, Y_m)$.

Using similar arguments to those in [BJKS04], we can show that there are positive, sufficiently small constants δ_0 and C so that for any randomized protocol Π which succeeds with probability at least $1-\delta_0$ on distribution σ,

$$I(X,Y;\Pi|Z,P) \geq \frac{Cm}{B^2}, \tag{7}$$

where, with some abuse of notation, Π is also used to denote the transcript of the corresponding randomized protocol, and here the input (X,Y) is drawn from σ conditioned on (X,Y) being a NO instance. Here, Π is randomized, and succeeds with probability at least $1-\delta_0$, where the probability is over the joint space of the random coins of Π and the input distribution.

Our starting point for proving (7) is Jayram's lower bound for the conditional mutual information when the inputs are drawn from a related distribution (reference [70] on p.182 of [Bar02]), but we require several non-trivial modifications to his argument in order to apply it to bound the conditional mutual information for our input distribution, which is σ conditioned on (X,Y) being a NO instance. Essentially, we are able to show that the variation distance between our distribution and his distribution is small, and use this to bound the difference in the conditional mutual information between the two distributions. The proof is rather technical, and we postpone it to Appendix C.

We make a few simple refinements to (7). Define the random variable W which is 1 if (X,Y) is a YES instance, and 0 if (X,Y) is a NO instance. Then by definition of the mutual information, if (X,Y) is drawn from σ without conditioning on (X,Y) being a NO instance, then we have

$$\begin{aligned} I(X,Y;\Pi|W,Z,P) &\geq \frac{1}{2}I(X,Y;\Pi|Z,P,W=0) \\ &= \Omega(m/B^2). \end{aligned}$$

Observe that

$$I(X,Y;\Pi|S,W,Z,P) \geq I(X,Y;\Pi|W,Z,P) - H(S) = \Omega(m/B^2), \tag{8}$$

where we assume that $\Omega(m/B^2) - \log m = \Omega(m/B^2)$. Define the constant $\delta_1 = \delta_0/4$. We now define a problem which involves solving r copies of $\mathsf{Gap}\ell_\infty^B$.

Definition 4.1 ($\mathsf{Multi}\ell_\infty^{r,B}$ Problem). *There are r pairs of inputs $(x^1,y^1),(x^2,y^2),\ldots,(x^r,y^r)$ such that each pair (x^i,y^i) is a legal instance of the $\mathsf{Gap}\ell_\infty^B$ problem. Alice is given $x^1,\ldots,x^r$. Bob is given $y^1,\ldots,y^r$. The goal is to output a vector $v \in \{NO,YES\}^r$, so that for at least a $1-\delta_1$ fraction of the entries i, $v_i = \mathsf{Gap}\ell_\infty^B(x^i,y^i)$.*

Remark 4.2. *Notice that Definition 4.1 is defining a promise problem. We will study the distributional complexity of this problem under the distribution σ^r, which is a product distribution on the r instances $(x^1,y^1),(x^2,y^2),\ldots,(x^r,y^r)$.*

Theorem 4.3. $D_{\sigma^r,\delta_1}(\mathsf{Multi}\ell_\infty^{r,B}) = \Omega(rm/B^2)$.

Proof. Let Π be any deterministic protocol for $\mathsf{Multi}\ell_\infty^{r,B}$ which succeeds with probability at least $1-\delta_1$ in solving $\mathsf{Multi}\ell_\infty^{r,B}$ when the inputs are drawn from σ^r, where the probability is taken over the input distribution. We show that Π has communication cost $\Omega(rm/B^2)$.

Let $X^1,Y^1,S^1,W^1,Z^1,P^1\ldots,X^r,Y^r,S^r,W^r,Z^r$, and P^r be the random variables associated with σ^r, i.e., X^j,Y^j,S^j,W^j,P^j and Z^j correspond to the random variables X,Y,S,W,Z,P associated with the j-th independent instance drawn according to σ, defined above. We let $X = (X^1,\ldots,X^r)$, $X^{<j} = (X^1,\ldots,X^{j-1})$, and X^{-j} equal X without X^j. Similarly we define these vectors for Y,S,W,Z and P.

By the chain rule for mutual information, $I(X^1,\ldots,X^r,Y^1,\ldots,Y^r;\Pi|S,W,Z,P)$ is equal to $\sum_{j=1}^r I(X^j,Y^j;\Pi|X^{<j},Y^{<j},S,W,Z,P)$. Let V be the output of Π, and V_j be its j-th coordinate. For a value $j \in [r]$, we say that j is *good* if $\Pr_{X,Y}[V_j = \mathsf{Gap}\ell_\infty^B(X^j,Y^j)] \geq 1 - \frac{2\delta_0}{3}$. Since Π succeeds with probability at least $1-\delta_1 = 1-\delta_0/4$ in outputting a vector with at least a $1-\delta_0/4$ fraction of correct entries, the expected probability of success over a random $j \in [r]$ is at least $1-\delta_0/2$, and so by a Markov argument, there are $\Omega(r)$ good indices j.

Fix a value of $j \in [r]$ that is good, and consider $I(X^j,Y^j;\Pi|X^{<j},Y^{<j},S,W,Z,P)$. By expanding the conditioning, $I(X^j,Y^j;\Pi|X^{<j},Y^{<j},S,W,Z,P)$ is equal to

$$\mathbf{E}_{x,y,s,w,z,p}[I(X^j,Y^j;\Pi \mid (X^{<j},Y^{<j},S^{-j},W^{-j},Z^{-j},P^{-j}) = (x,y,s,w,z,p),S^j,W^j,Z^j,P^j)]. \tag{9}$$

For each x, y, s, w, z, p, define a randomized protocol $\Pi_{x,y,s,w,z,p}$ for $\mathsf{Gap}\ell_\infty^B$ under distribution σ. Suppose that Alice is given a and Bob is given b, where $(a, b) \sim \sigma$. Alice sets $X^j = a$, while Bob sets $Y^j = b$. Alice and Bob use x, y, s, w, z and p to set their remaining inputs as follows. Alice sets $X^{<j} = x$ and Bob sets $Y^{<j} = y$. Alice and Bob can randomly set their remaining inputs without any communication, since for $j' > j$, conditioned on $S^{j'}, W^{j'}, Z^{j'}$, and $P^{j'}$, Alice and Bob's inputs are independent. Alice and Bob run Π on inputs X, Y, and define $\Pi_{x,y,s,w,z,p}(a, b) = V_j$. We say a tuple (x, y, s, w, z, p) is *good* if

$$\Pr_{X,Y}[V_j = \mathsf{Gap}\ell_\infty^B(X^j, Y^j) \mid X^{<j} = x, Y^{<j} = y, S^{-j} = s, W^{-j} = w, Z^{-j} = z, P^{-j} = p] \geq 1 - \delta_0.$$

By a Markov argument, and using that j is good, we have $\Pr_{x,y,s,w,z,p}[(x, y, s, w, z, p)$ is good $] = \Omega(1)$. Plugging into (9), $I(X^j, Y^j; \Pi | X^{<j}, Y^{<j}, S, W, Z, P)$ is at least a constant times

$$\mathbf{E}_{x,y,s,w,z,p}[I(X^j Y^j; \Pi | (X^{<j}, Y^{<j}, S^{-j}, W^{-j}, Z^{-j}, P^{-j}) = (x, y, s, w, z, p),$$
$$S^j, W^j, Z^j, P^j, (x, y, s, w, z, p) \text{ is good})].$$

For any (x, y, s, w, z, p) that is good, $\Pi_{x,y,s,w,z,p}(a, b) = V_j$ with probability at least $1 - \delta_0$, over the joint distribution of the randomness of $\Pi_{x,y,s,w,z,p}$ and $(a, b) \sim \sigma$. By (8),

$$\mathbf{E}_{x,y,s,w,z,p}[I(X^j, Y^j; \Pi | (X^{<j}, Y^{<j}, S^{-j}, W^{-j}, Z^{-j}, P^{-j}) = (x, y, s, w, z, p),$$
$$S^j, W^j, Z^j, P^j, (x, y, s, w, z, p) \text{ is good}] = \Omega\left(\frac{m}{B^2}\right).$$

Since there are $\Omega(r)$ good indices j, we have $I(X^1, \ldots, X^r; \Pi | S, W, Z, P) = \Omega(mr/B^2)$. Since the distributional complexity $D_{\sigma^r, \delta_1}(\mathsf{Multi}\ell_\infty^{r,B})$ is at least the minimum of $I(X^1, \ldots, X^r; \Pi | S, W, Z, P)$ over deterministic protocols Π which succeed with probability at least $1 - \delta_1$ on input distribution σ^r, it follows that $D_{\sigma^r, \delta_1}(\mathsf{Multi}\ell_\infty^{r,B}) = \Omega(mr/B^2)$. $\qquad\square$

4.2 The overall lower bound

We use the theorem in the previous subsection with an extension of the method of section 6.3 of [PW11].

Let $X \subset \mathbb{R}^n$ be a distribution with $x_i \in \{-n^d, \ldots, n^d\}$ for all $i \in [n]$ and $x \in X$. Here $d = \Theta(1)$ is a parameter. Given an adaptive compressed sensing scheme $\mathcal{A}$, we define a $(1 + \epsilon)$-approximate ℓ_1/ℓ_1 sparse recovery *multiround bit scheme* on X as follows.

Let A^i be the i-th (adaptively chosen) measurement matrix of the compressed sensing scheme. We may assume that the union of rows in matrices $A^1, \ldots, A^r$ generated by $\mathcal{A}$ is an orthonormal system, since the rows can be orthogonalized in a post-processing step. We can assume that $r \leq n$.

Choose a random $u \in \mathbb{R}^n$ from distribution $\mathcal{N}(0, \frac{1}{n^c} \cdot I_{n \times n})$, where $c = \Theta(1)$ is a parameter. We require that the compressed sensing scheme outputs a valid result of $(1 + \epsilon)$-approximate recovery on $x + u$ with probability at least $1 - \delta$, over the choice of u and its random coins. By Yao's minimax principle, we can fix the randomness of the compressed sensing scheme and assume that the scheme is deterministic.

Let B^1 be the matrix A^1 with entries rounded to $t \log n$ bits for a parameter $t = \Theta(1)$. We compute $B^1 x$. Then, we compute $B^1 x + A^1 u$. From this, we compute A^2, using the algorithm specified by $\mathcal{A}$ as if $B^1 x + A^1 u$ were equal to $A^1 x'$ for some x'. For this, we use the following lemma, which is Lemma 5.1 of [DIPW10].

Lemma 4.4. *Consider any $m \times n$ matrix A with orthonormal rows. Let B be the result of rounding A to b bits per entry. Then for any $v \in \mathbb{R}^n$ there exists an $s \in \mathbb{R}^n$ with $Bv = A(v - s)$ and $\|s\|_1 < n^2 2^{-b} \|v\|_1$.*

In general for $i \geq 2$, given $B^1 x + A^1 u, B^2 x + A^2 u, \ldots, B^{i-1} x + A^{i-1} u$ we compute A^i, and round to $t \log n$ bits per entry to get B^i. The output of the multiround bit scheme is the same as that of the compressed sensing scheme. If the compressed sensing scheme uses r rounds, then the multiround bit scheme uses r rounds. Let b denote the total number of bits in the concatenation of discrete vectors $B^1 x, B^2 x, \ldots, B^r x$.

We give a generalization of Lemma 5.2 of [PW11] which relates bit schemes to sparse recovery schemes. Here we need to generalize the relation from non-adaptive schemes to adaptive schemes, using Gaussian noise instead of uniform noise, and arguing about multiple rounds of the algorithm.

Lemma 4.5. *For $t = O(1 + c + d)$, a lower bound of $\Omega(b)$ bits for a multiround bit scheme with error probability at most $\delta + 1/n$ implies a lower bound of $\Omega(b/((1 + c + d) \log n))$ measurements for $(1 + \epsilon)$-approximate sparse recovery schemes with failure probability at most δ.*

Proof. Let $\mathcal{A}$ be a $(1+\epsilon)$-approximate adaptive compressed sensing scheme with failure probability δ. We will show that the associated multiround bit scheme has failure probability $\delta + 1/n$.

By Lemma 4.4, for any vector $x \in \{-n^d, \ldots, n^d\}$ we have $B^1 x = A^1(x + s)$ for a vector s with $\|s\|_1 \leq n^2 2^{-t \log n} \|x\|_1$, so $\|s\|_2 \leq n^{2.5-t} \|x\|_2 \leq n^{3.5+d-t}$. Notice that $u + s \sim \mathcal{N}(s, \frac{1}{n^c} \cdot I_{n \times n})$. We use the following quick suboptimal upper bound on the statistical distance between two univariate normal distributions, which suffices for our purposes.

Fact 4.6. *(see section 3 of [Pol05]) The variation distance between $\mathcal{N}(\theta_1, 1)$ and $\mathcal{N}(\theta_2, 1)$ is $\frac{4\tau}{\sqrt{2\pi}} + O(\tau^2)$, where $\tau = |\theta_1 - \theta_2|/2$.*

It follows by Fact 4.6 and independence across coordinates, that the variation distance between $\mathcal{N}(0, \frac{1}{n^c} \cdot I_{n \times n})$ and $\mathcal{N}(s, \frac{1}{n^c} \cdot I_{n \times n})$ is the same as that between $\mathcal{N}(0, I_{n \times n})$ and $\mathcal{N}(s \cdot n^{c/2}, I_{n \times n})$, which can be upper-bounded as

$$
\begin{aligned}
\sum_{i=1}^{n} \cdot \frac{2n^{c/2}|s_i|}{\sqrt{2\pi}} + O(n^c s_i^2) &= O(n^{c/2}\|s\|_1 + n^c \|s\|_2^2) \\
&= O(n^{c/2} \cdot \sqrt{n}\|s\|_2 + n^c \|s\|_2^2) \\
&= O(n^{c/2+4+d-t} + n^{c+7+2d-2t}).
\end{aligned}
$$

It follows that for $t = O(1 + c + d)$, the variation distance is at most $1/n^2$.

Therefore, if $\mathcal{T}^1$ is the algorithm which takes $A^1(x + u)$ and produces A^2, then $\mathcal{T}^1(A^1(x + u)) = \mathcal{T}^1(B^1 x + A^1 u)$ with probability at least $1 - 1/n^2$. This follows since $B^1 x + A^1 u = A^1(x + u + s)$ and $u + s$ and u have variation distance at most $1/n^2$.

In the second round, $B^2 x + A^2 u$ is obtained, and importantly we have for the algorithm $\mathcal{T}^2$ in the second round, $\mathcal{T}^2(A^2(x + u)) = \mathcal{T}^2(B^2 x + A^2 u)$ with probability at least $1 - 1/n^2$. This follows since A^2 is a deterministic function of $A^1 u$, and $A^1 u$ and $A^2 u$ are independent since A^1 and A^2 are orthonormal while u is a vector of i.i.d. Gaussians (here we use the rotational invariance / symmetry of Gaussian space). It follows by induction that with probability at least $1 - r/n^2 \geq 1 - 1/n$, the output of the multiround bit scheme agrees with that of $\mathcal{A}$ on input $x + u$.

Hence, if m_i is the number of measurements in round i, and $m = \sum_{i=1}^{r} m_i$, then we have a multiround bit scheme using a total of $b = mt \log n = O(m(1 + c + d) \log n)$ bits and with failure probability $\delta + 1/n$. $\qquad \square$

The rest of the proof is similar to the proof of the non-adaptive lower bound for ℓ_1/ℓ_1 sparse recovery given in [PW11]. We sketch the proof, referring the reader to [PW11] for some of the details. Fix parameters $B = \Theta(1/\epsilon^{1/2})$, $r = k$, $m = 1/\epsilon^{3/2}$, and $n = k/\epsilon^3$. Given an instance

$(x^1, y^1), \ldots, (x^r, y^r)$ of $\mathsf{Multi}\ell_\infty^{r,B}$ we define the input signal z to a sparse recovery problem. We allocate a set S^i of m disjoint coordinates in a universe of size n for each pair (x^i, y^i), and on these coordinates place the vector $y^i - x^i$. The locations turn out to be essential for the proof of Lemma 4.8 below, and are placed uniformly at random among the n total coordinates (subject to the constraint that the S^i are disjoint). Let ρ be the induced distribution on z.

Fix a $(1 + \epsilon)$-approximate k-sparse recovery multiround bit scheme Alg that uses b bits and succeeds with probability at least $1 - \delta_1/2$ over $z \sim \rho$. Let S be the set of top k coordinates in z. As shown in equation (14) of [PW11], Alg has the guarantee that if $w = Alg(z)$, then

$$\|(w - z)_S\|_1 + \|(w - z)_{[n]\setminus S}\|_1 \le (1 + 2\epsilon)\|z_{[n]\setminus S}\|_1. \tag{10}$$

(the $1 + 2\epsilon$ instead of the $1 + \epsilon$ factor is to handle the rounding of entries of the A^i and the noise vector u). Next is our generalization of Lemma 6.8 of [PW11].

Lemma 4.7. *For $B = \Theta(1/\epsilon^{1/2})$ sufficiently large, suppose that*

$$\Pr_{z \sim \rho}[\|(w - z)_S\|_1 \le 10\epsilon \cdot \|z_{[n]\setminus S}\|_1] \ge 1 - \frac{\delta_1}{2}.$$

Then Alg requires $b = \Omega(k/\epsilon^{1/2})$.

Proof. We show how to use Alg to solve instances of $\mathsf{Multi}\ell_\infty^{r,B}$ with probability at least $1 - \delta_1$, where the probability is over input instances to $\mathsf{Multi}\ell_\infty^{r,B}$ distributed according to σ^r, inducing the distribution ρ on z. The lower bound will follow by Theorem 4.3. Let w be the output of Alg.

Given $x^1, \ldots, x^r$, Alice places $-x^i$ on the appropriate coordinates in the set S^i used in defining z, obtaining a vector z_{Alice}. Given $y^1, \ldots, y^r$, Bob places the y^i on the appropriate coordinates in S^i. He thus creates a vector z_{Bob} for which $z_{Alice} + z_{Bob} = z$. In round i, Alice transmits $B^i z_{Alice}$ to Bob, who computes $B^i(z_{Alice} + z_{Bob})$ and transmits it back to Alice. Alice can then compute $B^i(z) + A^i(u)$ for a random $u \sim \mathcal{N}(0, \frac{1}{n^c} \cdot I_{n \times n})$. We can assume all coordinates of the output vector w are in the real interval $[0, B]$, since rounding the coordinates to this interval can only decrease the error.

To continue the analysis, we use a proof technique of [PW11] (see the proof of Lemma 6.8 of [PW11] for a comparison). For each i we say that S^i is *bad* if either

- there is no coordinate j in S^i for which $|w_j| \ge \frac{B}{2}$ yet (x^i, y^i) is a YES instance of $\mathsf{Gap}\ell_\infty^B$, or

- there is a coordinate j in S^i for which $|w_j| \ge \frac{B}{2}$ yet either (x^i, y^i) is a NO instance of $\mathsf{Gap}\ell_\infty^B$ or j is not the unique j^* for which $y_{j^*}^i - x_{j^*}^i = B$.

The proof of Lemma 6.8 of [PW11] shows that the fraction $C > 0$ of bad S^i can be made an arbitrarily small constant by appropriately choosing an appropriate $B = \Theta(1/\epsilon^{1/2})$. Here we choose $C = \delta_1$. We also condition on $\|u\|_2 \le n^{-c}$ for a sufficiently large constant $c > 0$, which occurs with probability at least $1 - 1/n$. Hence, with probability at least $1 - \delta_1/2 - 1/n > 1 - \delta_1$, we have a $1 - \delta_1$ fraction of indices i for which the following algorithm correctly outputs $\mathsf{Gap}\ell_\infty(x^i, y^i)$: if there is a $j \in S^i$ for which $|w_j| \ge B/2$, output YES, otherwise output NO. It follows by Theorem 4.3 that Alg requires $b = \Omega(k/\epsilon^{1/2})$, independent of the number of rounds. $\square$

The next lemma is the same as Lemma 6.9 of [PW11], replacing δ in the lemma statement there with the constant δ_1 and observing that the lemma holds for compressed sensing schemes with an arbitrary number of rounds.

Lemma 4.8. *Suppose* $\Pr_{z \sim \rho}[\|(w - z)_{[n] \setminus S}\|_1 \leq (1 - 8\epsilon) \cdot \|z_{[n] \setminus S}\|_1] \geq \delta_1$. *Then Alg requires* $b = \Omega(k \log(1/\epsilon)/\epsilon^{1/2})$.

Proof. As argued in Lemma 6.9 of [PW11], we have $I(w; z) = \Omega(\epsilon m r \log(n/(mr)))$, which implies that $b = \Omega(\epsilon m r \log(n/(mr)))$, independent of the number r of rounds used by *Alg*, since the only information about the signal is in the concatenation of $B^1 z, \ldots, B^r z$. $\square$

Finally, we combine our Lemma 4.7 and Lemma 4.8 to prove the analogue of Theorem 6.10 of [PW11], which completes this section.

Theorem 4.9. *Any* $(1 + \epsilon)$-*approximate* ℓ_1/ℓ_1 *recovery scheme with success probability at least* $1 - \delta_1/2 - 1/n$ *must make* $\Omega(k/(\epsilon^{1/2} \cdot \log(k/\epsilon)))$ *measurements.*

Proof. We will lower bound the number of bits used by any ℓ_1/ℓ_1 multiround bit scheme *Alg*. If *Alg* succeeds with probability at least $1 - \delta_1/2$, then in order to satisfy (10), we must either have $\|(w - z)_S\|_1 \leq 10\epsilon \cdot \|z_{[n] \setminus S}\|_1$ or $\|(w - z)_{[n] \setminus S}\|_1 \leq (1 - 8\epsilon)\|z_{[n] \setminus S}\|_1$. Since *Alg* succeeds with probability at least $1 - \delta_1/2$, it must either satisfy the hypothesis of Lemma 4.7 or Lemma 4.8. But by these two lemmas, it follows that $b = \Omega(k/\epsilon^{1/2})$. Therefore by Lemma 4.5, any $(1 + \epsilon)$-approximate ℓ_1/ℓ_1 sparse recovery algorithm succeeding with probability at least $1 - \delta_1/2 - 1/n$ requires $\Omega(k/(\epsilon^{1/2} \cdot \log(k/\epsilon)))$ measurements. $\square$

5 Acknowledgements

References

[ACD11]   E. Arias-Castro, E.J. Candes, and M. Davenport. On the fundamental limits of adaptive sensing. *Arxiv preprint arXiv:1111.4646*, 2011.

[AWZ08]   A. Aldroubi, H. Wang, and K. Zarringhalam. Sequential adaptive compressed sampling via huffman codes. *Preprint*, 2008.

[Bar02]   Ziv Bar-Yossef. *The Complexity of Massive Data Set Computations*. PhD thesis, UC Berkeley, 2002.

[BJKS04] Ziv Bar-Yossef, T. S. Jayram, Ravi Kumar, and D. Sivakumar. An information statistics approach to data stream and communication complexity. *J. Comput. Syst. Sci.*, 68(4):702–732, 2004.

[BR11]   Mark Braverman and Anup Rao. Information equals amortized communication. In *STOC*, 2011.

[CD11]   E.J. Candès and M.A. Davenport. How well can we estimate a sparse vector? *Arxiv preprint arXiv:1104.5246*, 2011.

[CHNR08] R. Castro, J. Haupt, R. Nowak, and G. Raz. Finding needles in noisy haystacks. *Proc. IEEE Conf. Acoustics, Speech, and Signal Proc.*, page 5133¡D0¿5136, 2008.

[CRT06]   E. J. Candès, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(8):1208–1223, 2006.

[DIPW10] K. Do Ba, P. Indyk, E. Price, and D. Woodruff. Lower bounds for sparse recovery. *SODA*, 2010.

[Don06]   D. L. Donoho. Compressed Sensing. *IEEE Trans. Info. Theory*, 52(4):1289–1306, Apr. 2006.

[GLPS10] Anna C. Gilbert, Yi Li, Ely Porat, and Martin J. Strauss. Approximate sparse recovery: optimizing time and measurements. In *STOC*, pages 475–484, 2010.

[GMRZ11] Parikshit Gopalan, Raghu Meka, Omer Reingold, and David Zuckerman. Pseudorandom generators for combinatorial shapes. In *STOC*, pages 253–262, 2011.

[HBCN09] J. Haupt, R. Baraniuk, R. Castro, and R. Nowak. Compressive distilled sensing. *Asilomar*, 2009.

[HCN09]   J. Haupt, R. Castro, and R. Nowak. Adaptive sensing for sparse signal recovery. *Proc. IEEE 13th Digital Sig. Proc./5th Sig. Proc. Education Workshop*, page 702¡D0¿707, 2009.

[HIKP12] H. Hassanieh, P. Indyk, D. Katabi, and E. Price. Nearly optimal sparse Fourier transform. *STOC*, 2012.

[IPW11]   P. Indyk, E. Price, and D.P. Woodruff. On the power of adaptivity in sparse recovery. *FOCS*, 2011.

[Jay02]   T.S. Jayram. Unpublished manuscript, 2002.

[JXC08]   S. Ji, Y. Xue, and L. Carin. Bayesian compressive sensing. *Signal Processing, IEEE Transactions on*, 56(6):2346–2356, 2008.

[KN97]   Eyal Kushilevitz and Noam Nisan. *Communication complexity.* Cambridge University Press, 1997.

[MSW08]   D. M. Malioutov, S. Sanghavi, and A. S. Willsky. Compressed sensing with sequential observations. *ICASSP*, 2008.

[Mut05]   S. Muthukrishnan. Data streams: Algorithms and applications). *FTTCS*, 2005.

[Pol05]   Pollard. Total variation distance between measures. 2005. http://www.stat.yale.edu/ pollard/Courses/607.spring05/handouts/Totalvar.pdf.

[PW11]   Eric Price and David P. Woodruff. (1+eps)-approximate sparse recovery. *CoRR*, abs/1110.4414, 2011.

[SAZ10]   N. Shental, A. Amir, and Or Zuk. Identification of rare alleles and their carriers using compressed se(que)nsing. *Nucleic Acids Research*, 38(19):1–22, 2010.

[SS02]   Michael E. Saks and Xiaodong Sun. Space lower bounds for distance approximation in the data stream model. In *Proceedings of the 34th Annual ACM Symposium on Theory of Computing (STOC)*, pages 360–369, 2002.

A Relationship between Post-Measurement and Pre-Measurement noise

In the setting of [ACD11], the goal is to recover a k-sparse x from observations of the form $Ax + w$, where A has unit norm rows and w is i.i.d. Gaussian with variance $\|x\|_2^2/\epsilon^2$. By ignoring the (irrelevant) component of w orthogonal to A, this is equivalent to recovering x from observations of the form $A(x + w)$. By contrast, our goal is to recover $x + w$ from observations of the form $A(x + w)$, and for general w rather than only for Gaussian w.

By arguments in [PW11, HIKP12], for Gaussian w the difference between recovering x and recovering $x + w$ is minor, so any lower bound of m in the [ACD11] setting implies a lower bound of $\min(m, \epsilon n)$ in our setting. The converse is only true for proofs that use Gaussian w, but our proof fits this category.

B Information Chain Rule with Linear Observations

Lemma B.1. *Suppose $a_i = b_i + w_i$ for $i \in [s]$ and the w_i are independent of each other and the b_i. Then*

$$I(a; b) \leq \sum_i I(a_i; b_i)$$

Proof. Note that $h(a \mid b) = h(a - b \mid b) = h(w \mid b) = h(w)$. Thus

$$I(a; b) = h(a) - h(a \mid b) = h(a) - h(w)$$
$$\leq \sum_i h(a_i) - h(w_i)$$
$$= \sum_i h(a_i) - h(a_i \mid b_i) = \sum_i I(a_i; b_i)$$

14

$\square$

C Switching Distributions from Jayram's Distributional Bound

We first sketch a proof of Jayram's lower bound on the distributional complexity of $\mathsf{Gap}\ell_\infty^B$ [Jay02], then change it to a different distribution that we need for our sparse recovery lower bounds in Subsection C.1. Let $X, Y \in \{0, 1, \ldots, B\}^m$. Define distribution $\mu^{m,B}$ as follows: for each $j \in [m]$, choose a random pair $(Z_j, P_j) \in \{0, 1, 2, \ldots, B\} \times \{0, 1\} \setminus \{(0, 1), (B, 0)\}$. If $(Z_j, P_j) = (z, 0)$, then $X_j = z$ and Y_j is uniformly distributed in $\{z, z + 1\}$; if $(Z_j, P_j) = (z, 1)$, then $Y_j = z$ and X_j is uniformly distributed on $\{z - 1, z\}$. Let $X = (X_1, \ldots, X_m)$, $Y = (Y_1, \ldots, Y_m)$, $Z = (Z_1, \ldots, Z_m)$ and $P = (P_1, \ldots, P_m)$.

The other distribution we define is $\sigma^{m,B}$, which is the same as distribution σ in Section 4 (we include m and B in the notation here for clarity). This is defined by first drawing X and Y according to distribution $\mu^{m,B}$. Then, we pick a random coordinate $S \in [m]$ and replace (X_S, Y_S) with a uniformly random element in the set $\{(0, 0), (0, B)\}$.

Let Π be a deterministic protocol that errs with probability at most δ on input distribution $\sigma^{m,B}$.

By the chain rule for mutual information, when X and Y are distributed according to $\mu^{m,B}$,

$$I(X, Y; \Pi | Z, P) \;=\; \sum_{j=1}^m I(X_j, Y_j; \Pi | X^{<j}, Y^{<j}, Z, P),$$

which is equal to

$$\sum_{j=1}^m \mathbf{E}_{x,y,z,p}[I(X_j, Y_j; \Pi \,| Z_j, P_j, X^{<j} = x, Y^{<j} = y, Z^{-j} = z, P^{-j} = p)].$$

Say that an index $j \in [m]$ is *good* if conditioned on $S = j$, Π succeeds on $\sigma^{m,B}$ with probability at least $1 - 2\delta$. By a Markov argument, at least $m/2$ of the indices j are good. Fix a good index j.

We say that the tuple (x, y, z, p) is *good* if conditioned on $S = j$, $X^{<j} = x$, $Y^{<j} = y$, $Z^{-j} = z$, and $P^{-j} = p$, Π succeeds on $\sigma^{m,B}$ with probability at least $1 - 4\delta$. By a Markov bound, with probability at least $1/2$, (x, y, z, p) is good. Fix a good (x, y, z, p).

We can define a single-coordinate protocol $\Pi_{x,y,z,p,j}$ as follows. The parties use x and y to fill in their input vectors X and Y for coordinates $j' < j$. They also use $Z^{-j} = z$, $P^{-j} = p$, and private randomness to fill in their inputs without any communication on the remaining coordinates $j' > j$. They place their single-coordinate input (U, V) on their j-th coordinate. The parties then output whatever Π outputs.

It follows that $\Pi_{x,y,z,p,j}$ is a single-coordinate protocol Π' which distinguishes $(0, 0)$ from $(0, B)$ under the uniform distribution with probability at least $1 - 4\delta$. For the single-coordinate problem, we need to bound $I(X_j, Y_j; \Pi' | Z_j, P_j)$ when (X_j, Y_j) is uniformly random from the set $\{(Z_j, Z_j), (Z_j, Z_j + 1)\}$ if $P_j = 0$, and (X_j, Y_j) is uniformly random from the set $\{(Z_j, Z_j), (Z_j - 1, Z_j)\}$ if $P_j = 1$. By the same argument as in the proof of Lemma 8.2 of [BJKS04], if $\Pi'_{u,v}$ denotes the distribution on transcripts induced by inputs u and v and private coins, then we have

$$I(X_j, Y_j; \Pi' | Z_j, P_j) \geq \Omega(1/B^2) \cdot (h^2(\Pi'_{0,0}, \Pi'_{0,B}) + h^2(\Pi'_{B,0}, \Pi'_{B,B})), \tag{11}$$

where

$$h(\alpha, \beta) = \sqrt{\frac{1}{2} \sum_{\omega \in \Omega} (\sqrt{\alpha(\omega)} - \sqrt{\beta(\omega)})^2}$$

is the Hellinger distance between distributions α and β on support Ω. For any two distributions α and β, if we define

$$D_{TV}(\alpha, \beta) = \frac{1}{2} \sum_{\omega \in \Omega} |\alpha(\omega) - \beta(\omega)|$$

to be the variation distance between them, then $\sqrt{2} \cdot h(\alpha, \beta) \geq D_{TV}(\alpha, \beta)$ (see Proposition 2.38 of [Bar02]).

Finally, since Π' succeeds with probability at least $1 - 4\delta$ on the uniform distribution on input pair in $\{(0,0), (0,B)\}$, we have

$$\sqrt{2} \cdot h(\Pi'_{0,0}, \Pi'_{0,B}) \geq D_{TV}(\Pi'_{0,0}, \Pi'_{0,B}) = \Omega(1).$$

Hence,

$$I(X_j, Y_j; \Pi | Z_j, P_j, X^{<j} = x, Y^{<j} = y, Z^{-j} = z, P^{-j} = p)$$
$$= \Omega(1/B^2)$$

for each of the $\Omega(m)$ good j. Thus $I(X, Y; \Pi | Z, P) = \Omega(m/B^2)$ when inputs X and Y are distributed according to $\mu^{m,B}$, and Π succeeds with probability at least $1 - \delta$ on X and Y distributed according to $\sigma^{m,B}$.

C.1   Changing the distribution

Consider the distribution

$$\zeta^{m,B} = (\sigma^{m,B} \mid (X_S, Y_S) = (0,0)).$$

We show $I(X, Y; \Pi | Z) = \Omega(m/B^2)$ when X and Y are distributed according to $\zeta^{m,B}$ rather than according to $\mu^{m,B}$.

For X and Y distributed according to $\zeta^{m,B}$, by the chain rule we again have that $I(X, Y; \Pi | Z, P)$ is equal to

$$\sum_{j=1}^{m} \mathbf{E}_{x,y,z,p}[I(X_j, Y_j; \Pi | Z_j, P_j, X^{<j} = x, Y^{<j} = y, Z^{-j} = z, P^{-j} = p)].$$

Again, say that an index $j \in [m]$ is good if conditioned on $S = j$, Π succeeds on $\sigma^{m,B}$ with probability at least $1 - 2\delta$. By a Markov argument, at least $m/2$ of the indices j are good. Fix a good index j.

Again, we say that the tuple (x, y, z, p) is good if conditioned on $S = j$, $X^{<j} = x$, $Y^{<j} = y$, $Z^{-j} = z$ and $P^{-j} = p$, Π succeeds on $\sigma^{m,B}$ with probability at least $1 - 4\delta$. By a Markov bound, with probability at least $1/2$, (x, y, z, p) is good. Fix a good (x, y, z, p).

As before, we can define a single-coordinate protocol $\Pi_{x,y,z,p,j}$. The parties use x and y to fill in their input vectors X and Y for coordinates $j' < j$. They can also use $Z^{-j} = z$, $P^{-j} = p$, and private randomness to fill in their inputs without any communication on the remaining coordinates $j' > j$. They place their single-coordinate input (U, V), uniformly drawn from $\{(0,0), (0,B)\}$, on their j-th coordinate. The parties output whatever Π outputs. Let Π' denote $\Pi_{x,y,z,p,j}$ for notational convenience.

The first issue is that unlike before Π' is not guaranteed to have success probability at least $1 - 4\delta$ since Π is not being run on input distribution $\sigma^{m,B}$ in this reduction. The second issue is in bounding $I(X_j, Y_j; \Pi' | Z_j, P_j)$ since (X_j, Y_j) is now drawn from the marginal distribution of $\zeta^{m,B}$ on coordinate j.

Notice that $S \neq j$ with probability $1 - 1/m$, which we condition on. This immediately resolves the second issue since now the marginal distribution on (X_j, Y_j) is the same under $\zeta^{m,B}$ as it was under $\sigma^{m,B}$; namely it is the following distribution: (X_j, Y_j) is uniformly random from the set $\{(Z_j, Z_j), (Z_j, Z_j + 1)\}$ if $P_j = 0$, and (X_j, Y_j) is uniformly random from the set $\{(Z_j, Z_j), (Z_j - 1, Z_j)\}$ if $P_j = 1$.

We now address the first issue. After conditioning on $S \neq j$, we have that (X^{-j}, Y^{-j}) is drawn from $\zeta^{m-1,B}$. If instead (X^{-j}, Y^{-j}) were drawn from $\mu^{m-1,B}$, then after placing (U, V) the input to Π would be drawn from $\sigma^{m,B}$ conditioned on a good tuple. Hence in that case, Π' would succeed with probability $1 - 4\delta$. Thus for our actual distribution on (X^{-j}, Y^{-j}), after conditioning on $S \neq j$, the success probability of Π' is at least

$$1 - 4\delta - D_{TV}(\mu^{m-1,B}, \zeta^{m-1,B}).$$

Let $C^{\mu,m-1,B}$ be the random variable which counts the number of coordinates i for which $(X_i, Y_i) = (0, 0)$ when X and Y are drawn from $\mu^{m-1,B}$. Let $C^{\zeta,m-1,B}$ be a random variable which counts the number of coordinates i for which $(X_i, Y_i) = (0, 0)$ when X and Y are drawn from $\zeta^{m-1,B}$. Observe that $(X_i, Y_i) = (0, 0)$ in μ only if $P_i = 0$ and $Z_i = 0$, which happens with probability $1/(2B)$. Hence, $C^{\mu,m-1,B}$ is distributed as $\text{Binomial}(m - 1, 1/(2B))$, while $C^{\zeta,m-1,B}$ is distributed as $\text{Binomial}(m - 2, 1/(2B)) + 1$. We use μ' to denote the distribution of $C^{\mu,m-1,B}$ and ζ' to denote the distribution of $C^{\zeta,m-1,B}$. Also, let ι denote the $\text{Binomial}(m - 2, 1/(2B))$ distribution. Conditioned on $C^{\mu,m-1,B} = C^{\zeta,m-1,B}$, we have that $\mu^{m-1,B}$ and $\zeta^{m-1,B}$ are equal as distributions, and so

$$D_{TV}(\mu^{m-1,B}, \zeta^{m-1,B}) \leq D_{TV}(\mu', \zeta').$$

We use the following fact:

Fact C.1. *(see, e.g., Fact 2.4 of [GMRZ11]). Any binomial distribution X with variance equal to σ^2 satisfies $D_{TV}(X, X + 1) \leq 2/\sigma$.*

By definition,

$$\mu' = (1 - 1/(2B)) \cdot \iota + 1/(2B) \cdot \zeta'.$$

Since the variance of the $\text{Binomial}(m - 2, 1/(2B))$ distribution is

$$(m - 2)/(2B) \cdot (1 - 1/(2B)) = m/(2B)(1 - o(1)),$$

applying Fact C.1 we have

$$
\begin{aligned}
D_{TV}(\mu', \zeta') &= D_{TV}((1 - 1/(2B)) \cdot \iota + (1/(2B)) \cdot \zeta', \zeta') \\
&= \frac{1}{2} \cdot \|(1 - 1/(2B)) \cdot \iota + (1/(2B)) \cdot \zeta' - \zeta'\|_1 \\
&= (1 - 1/(2B)) \cdot D_{TV}(\iota, \zeta') \\
&\leq \frac{2\sqrt{2B}}{\sqrt{m}} \cdot (1 + o(1)) \\
&= O\left(\sqrt{\frac{B}{m}}\right).
\end{aligned}
$$

It follows that the success probability of Π' is at least

$$1 - 4\delta - O\left(\sqrt{\frac{B}{m}}\right) \geq 1 - 5\delta.$$

17

Let E be an indicator random variable for the event that $S \neq j$. Then $H(E) = O((\log m)/m)$. Hence,

$$
\begin{aligned}
I(X_j, Y_j; \Pi'|Z_j, P_j) &\geq I(X_j, Y_j; \Pi'|Z_j, P_j, E) - O((\log m)/m) \\
&\geq (1 - 1/m) \cdot I(X_j, Y_j; \Pi'|Z_j, P_j, S \neq j) - O((\log m)/m) \\
&= \Omega(1/B^2),
\end{aligned}
$$

where we assume that $\Omega(1/B^2) - O((\log m)/m) = \Omega(1/B^2)$.

Hence, $I(X, Y; \Pi|Z, P) = \Omega(m/B^2)$ when inputs X and Y are distributed according to $\zeta^{m,B}$, and Π succeeds with probability at least $1 - \delta$ on X and Y distributed according to $\sigma^{m,B}$.