The Price of Privacy and the Limits of LP Decoding

Cynthia Dwork
dwork@microsoft.com

Frank McSherry
mcsherry@microsoft.com

Kunal Talwar
kunal@microsoft.com

Microsoft Research Silicon Valley
Mountain View, CA 94043

ABSTRACT

This work is at the intersection of two lines of research. One line, initiated by Dinur and Nissim, investigates the price, in accuracy, of protecting privacy in a statistical database. The second, growing from an extensive literature on compressed sensing (see in particular the work of Donoho and collaborators [14, 7, 13, 11]) and explicitly connected to error-correcting codes by Candès and Tao ([4]; see also [5, 3]), is in the use of linear programming for error correction.

Our principal result is the discovery of a sharp threshhold $\rho^* \approx 0.239$, so that if $\rho < \rho^*$ and A is a random $m \times n$ encoding matrix of independently chosen standard Gaussians, where $m = O(n)$, then with overwhelming probability over choice of A, for all $x \in \mathbb{R}^n$, LP decoding corrects $\lfloor \rho m \rfloor$ arbitrary errors in the encoding Ax, while decoding can be made to fail if the error rate exceeds ρ^*. Our bound resolves an open question of Candès, Rudelson, Tao, and Vershyin [3] and (oddly, but explicably) refutes empirical conclusions of Donoho [11] and Candès et al [3]. By scaling and rounding we can easily transform these results to obtain polynomial-time decodable random linear codes with polynomial-sized alphabets tolerating any $\rho < \rho^* \approx 0.239$ fraction of arbitrary errors.

In the context of privacy-preserving datamining our results say that any privacy mechanism, interactive or non-interactive, providing reasonably accurate answers to a 0.761 fraction of randomly generated weighted subset sum queries, and arbitrary answers on the remaining 0.239 fraction, is blatantly non-private.

Categories and Subject Descriptors: E.4 Coding and Information Theory: Error Control Codes; H.2.8 Database Applications: Statistical Databases; G.3 Probability and Statistics: Probabilistic Algorithms

General Terms: Algorithms, Theory

Keywords: Privacy, LP Decoding, Compressed Sensing, Basis Pursuit

STOC'07, June 11–13, 2007, San Diego, California, USA.
Copyright 2007 ACM 978-1-59593-631-8/07/0006 ...$5.00.

1. INTRODUCTION

This work is at the intersection of two lines of research. One line, initiated by Dinur and Nissim and providing our original motivation, investigates the price, in accuracy, of protecting privacy in a statistical database. The conflict is between the *curator*, whose goal is to answer questions while preserving the privacy of individuals, and the *attacker*, whose goal is to compromise privacy. The second line, growing from an extensive literature on compressed sensing, is in the use of linear programming for error correction. Here, the conflict is between the *adversary*, who corrupts a signal, and the *decoder*, who attempts to reconstruct the message.

Recall the classical problem of transmitting a message in the presence of adversarially generated noise. Given a vector $x \in \mathbb{R}^n$, one approach is to encode x using an $m \times n$ encoding matrix A, and to transmit $Ax \in \mathbb{R}^m$. We show the existence of a sharp threshhold ρ^* such that for $\rho < \rho^*$, there is a fixed constant c such that if $m \geq cn$ and the entries of A are chosen independently from a standard Gaussian distribution, then with overwhelming probability, for all $x \in \mathbb{R}^n$, if the number of errors in the received word $Ax + e$ is at most ρm, the vector x is exactly retrieved using linear programming. Moreover, we show the bound ρ^* to be tight: once the fraction of errors exceeds ρ^*, (again with overwhelming probability) LP decoding can always be made to fail in a very strong sense. This resolves the "interesting challenge" posed in [4, 3].

We draw on a line of work with very deep roots in statistics that studies the use of ℓ_1 minimization via linear programming for recovering sparse signals from linear measurements. There is a rich literature in this field; the work of Donoho and collaborators contains several pivotal advances [14, 7, 13, 11]), while recent work of Candès and Tao and collaborators makes the existential more concrete [4, 3] by identifying specific conditions on the $m \times n$ coding matrix A for which LP decoding is guaranteed to yield the correct answer[1].

We extend our error-correction result in several ways:

1. We handle "mixed" errors – in addition to the ρ fraction of arbitrary errors, we tolerate any number of small errors, in the sense that if the magnitude of the small errors is bounded by α then reconstruction yields an x' of Euclidean distance at most $O(\alpha)$ from x. We may think of the error vector as $e + f$, where e has support ρm but its non-zero entries are arbitrary, while f may have support m but its non-zero entries

[1] The complexity of testing satisfaction of the conditions is not addressed.

are bounded by α. In this context the results in [3] yield an answer x' satisfying $||x-x'||_2 \leq O(||f||_1/\sqrt{m})$ (which is $O(\alpha\sqrt{m})$ in our setting), while we achieve $||x-x'||_2 \leq ||f||_\infty$ (which is $O(\alpha)$).

2. We obtain similar results for the case in which the coding matrix A has random, independently chosen, ± 1 entries. In this case, the error threshold $\rho^*_{\pm 1}$ is smaller. However, for any error rate up to $\rho^* \approx 0.239$ above, we get near perfect reconstruction – when no small errors are permitted, there may be a constant number (not fraction!) of entries in which the reconstructed x' differs from x

3. From both the results for Gaussians and for the ± 1 case, we obtain random linear error-correcting codes over finite fields, with polynomial-sized alphabets. See Section 6 for details.

4. We prove that ρ^* is optimal when the entries of A are i.i.d. Gaussians. Experiments by Donoho [11] and Candès et al [3] showed reconstruction in the presence of error rates "nearly" 0.3 when $m = 2n$ [11] and 35% when $m = 4n$ [3]. However, in the experiments the noise was not chosen adaptively, as a function of the query matrix A and the signal x, probably accounting for the discrepancy with our results.

5. We obtain a reconstruction gap for compressed sensing via linear programming – we explain the problem in Section 5. For the reader already familiar with compressed sensing, we note that our results show that LP decoding solves the compressed sensising problem for any density smaller than ρ^*, with the number of questions being $(1 - \frac{n}{m})m$ with n and m as above. Moreover, for *any* $f(m) = \omega(1)$, LP-based compressed sensing can be made to fail even with $m - f(m)$ measurements, even if, say, $m = 2^{2^n}$, if the sparseness of the input exceeds ρ^*.

We now turn to privacy-preserving data analysis, which was our original motivation for studying the results of [11, 3]. A statistic is a quantity computed from a sample. The goal of a statistical database is to provide statistics about a population while simultaneously protecting the privacy of the individuals in the database. A recent and highly fruitful direction is in a model of computation in which a trusted entity, the *curator*, sits between the possibly adversarial user of the database and the actual data. Queries are functions mapping the database to a (vector of) real(s). The curator computes the correct answer to the query and adds noise to the response. This natural approach allows highly accurate responses while maintaining strong privacy guarantees [18, 2, 17], but it suffers from a drawback: even for the simple case in which the database is a binary vector and the queries are subset sums of selected elements of the vector, the magnitude of noise added must increase with the total number of questions asked. A line of research initiated by Dinur and Nissim indicates that this increase is inherent [9]. They showed that if the database is a vector x of n bits and the curator provides relatively accurate (within $o(\sqrt{n})$) answers to $n \log^2 n$ random subset sum queries, then by using linear programming the attacker can reconstruct a database x' agreeing with x in all but $o(n)$ entries, ie, satisfying $support(x - x') \in o(n)$. We call this *blatant non-privacy.*

The Dinur-Nissim setting, while at first blush simplistic, is in fact sufficiently rich to capture many natural questions. For example, the rows of the database may be quite complex, but the attacker may know enough information about an individual in the database to uniquely identify his row. In this case the goal is to prevent any single *additional* bit of information to be learned from the database. In fact, careful use of hash functions can handle the "row-naming problem" even if the attacker does not know enough to uniquely identify individuals. Thus we can imagine a scenario in which an attacker reconstructs a close approximation to the database, where each row is identified with a set of hash values, and a "secret bit" is learned for many rows. At a later time the attacker may learn enough about an individual in the database to deduce the hash values for her record to identify the row corresponding to the individual, and thus obtain her "secret bit." Details appear in [8]. So naming a set of rows to specify a query is not just a theoretical possibility, and the assumption of only a single sensitive attribute per user still yields meaningful results.

Research statisticians like to "look at the data." Indeed, conversations with experts in this field frequently involve pleas for a "noisy table" that will permit significantly accurate answers to be derived for computations that are not specified at the outset. For these people the implications of the Dinur-Nissim results are particularly significant: no "noisy table" can provide very accurate answers to too many questions; otherwise the table could be used to simulate the interactive mechanism, and a Dinur-Nissim style attack could be mounted against the table. Even worse, while in the interactive setting the noise can be adapted to the queries, in the non-interactive setting the curator does not have this freedom to aid in protecting privacy.

Our work extends the results of Dinur and Nissim to the case in which a $\rho < \rho^*$ fraction of the query responses are arbitrarily inaccurate and any number of the responses may additionally suffer from error $o(\sqrt{n})$. In retrospect, the confluence of error-correction and proving non-privacy of a statistical database is not surprising: the attacker's goal is to take several measurements and reconstruct the database, or an approximation thereto. In this sense the results of Donoho [11] and Candès et al. [3] were very timely, while the context for our work sharpened our interest in tolerating small errors.

Finally, we obtain additional blatant non-privacy results when the attacker is not restricted to run in time polynomial in the size of the database.

Related Work.

There is extensive literature on compressed sensing and on proving conditions under which ℓ_0/ℓ_1-equivalence holds (see e.g. [3, 7, 11, 10, 12, 15] and the citations therein). The work of Donoho [12] attempts to quantify the behavior of the permissible error rate for LP decoding to succeed, as a function of the redundancy of the code $\delta \overset{def}{=} \frac{m-n}{m}$. They look at the setting where the entries of A are i.i.d. Gaussians and establish a lower bound $\rho_N(\delta)$ (misleadingly referred to as a threshold by Donoho and Tanner [15]) on the permissible error rate. In the limit of δ approaching one, they prove that ℓ_0/ℓ_1 equivalence holds for any $\rho \leq 0.168$.

Independent of our work, Candès and Randall [6] look at a mixed-error model similar to ours. They use a different linear program, or a second-order cone program, and bound the error in recovery $\|x' - x\|$ in the mixed-error model in terms of the recovery error in absence of the wild noise.

We remark that the use of linear programming for correcting errors in linear encodings should not be confused with the recent work of Feldman and collaborators [19, 20, 21, 22] on LP decoding of turbo-codes, low-density parity check codes, and expander codes, all of which are binary codes.

2. NOTATION AND PRELIMINARIES

For a vector x, we shall denote by $|x|$ the ℓ_1 norm of x, and by $\|x\|$, its ℓ_2 norm.

We call a vector z (ρ,α)-small if all but a ρ fraction of its entries have magnitude no larger than α. In other words, for some set T, $|T| \leq \rho \cdot dim(z)$, it is the case that $|z_i| \leq \alpha$ for all $i \notin T$.

2.1 LP Decoding

Let $x \in \mathbb{R}^n$ and let A be an $m \times n$ matrix. In the simplest setting, A will be populated with random independent $N(0,1)$ Gaussian entries. But we shall also look at cases when the entries of A are chosen independently from $\{-1,+1\}$ or from $\{-1,0,+1\}$. We consider the encoding $y = Ax$ of x.

Our error model would allow for an arbitrary corruption of a ρ fraction of the entries of y, *and* a *small error* α in every entry of y. The limits on ρ and α will be discussed at a later point in the paper.

We look at the following decoding procedure: given the corrupted message y', we solve the following linear program that optimizes a linear objective function over variables x and y:

$$\begin{array}{ll} \text{minimize} & |y - y'| \\ \text{subject to:} & y = Ax \end{array}$$

We remark that while this is not strictly speaking a linear program as written, it can easily be converted to one. Let (x', y) be the optimal solution of the above linear program. We shall show that x' is "close" to x under suitable assumptions on ρ, α, and m (and in fact $x = x'$ when $\alpha = 0$ and $m = m(\rho)$ is a suitably large multiple of n).

2.2 Lower bounds for Privacy

One of our main motivations for studying the above question was to show that in order to prevent gross privacy violations, a curator must add a reasonably large amount (α) of error to a large fraction (ρ) of the queries asked of it.

Consider the set of queries given by the matrix A, i.e. for each row a_i of A, we ask the query $a_i \cdot x$, where x is (say) a binary database. Suppose that the answers returned by the curator have the following property:

- A fraction $(1 - \rho)$ of the answers are correct to within an error of α.
- A ρ fraction of the answers can be answered arbitrarily.

It is clear that one can use LP decoding in this setting; we shall add an additional set of constraints $0 \leq x_j \leq 1$ to the linear program. This gives us a vector $x' \in \mathbb{R}^n$ that is "close" to the true database x. We shall simply round each entry of x' to the nearer of $\{0, 1\}$ and use the rounded value $\hat{x}$ as our guess for the database.

Recall that blatant non-privacy is the approximate reconstruction of x, that is, the construction of and x' of Hamming distance at most $o(n)$ from x.

In the next two sections, we shall prove general results about the LP decoding procedure. We then apply those to the privacy settings to derive lower bounds for noise addition needed to prevent blatant non-privacy.

3. LP DECODING: GAUSSIAN ENSEMBLE

In this section, we shall prove that when the entries of A are i.i.d. Gaussian, for any $\rho < \rho^* = 0.239...$ (where the constant ρ^* is defined in Lemma 2) and any α, the reconstructed vector x' is such that $\|x' - x\| \leq O(\alpha)$.

THEOREM 1. *Given any $\rho < \rho^*$, there exist absolute constants $c_1, c_2, c_3 > 0$ such that the following holds. Whenever $m \geq c_1 n$ and n is large enough, with probability $(1 - exp(-c_2 n))$, an $m \times n$ matrix A with independent $N(0,1)$ entries has the following property: for every vector x and every error vector e that is (ρ,α)-small, the vector x' reconstructed by the LP decoding procedure is such that $\|x' - x\| \leq c_3 \alpha$.*

We first define the constant ρ^*.

LEMMA 2. *Let $X_1, \ldots, X_m$ be i.i.d. $N(0,1)$ random variables and let $Y_1, \ldots, Y_m$ be the sorted ordering (in decreasing order) of $|X_1|, \ldots, |X_m|$. For a $\rho > 0$, let S_ρ be the random variable $Y_1 + Y_2 + \ldots + Y_{\lceil \rho m \rceil}$. Let S denote $E[S_1]$. Then there exists a constant ρ^* such that $\lim_{m \to \infty} E[S_{\rho^*}]/S = \frac{1}{2}$.*

PROOF. Let X be distributed as $N(0,1)$ and let $Y = |X|$. Let $f(\cdot)$ denote the p.d.f. of Y and $F(\cdot)$ be its c.d.f. Define $g(x) = \int_x^\infty y f(y) dy$. Clearly g is continuous and decreasing in $[0, \infty]$ with $g(0) = E[Y] = S/m$ and $\lim_{x \to \infty} g(x) = 0$. Thus there exists x^* such that $g(x^*) = E[Y]/2$. Let $\rho^* = (1 - F^{-1}(x^*))$. It is easy to check that this ρ^* has the desired property. Indeed let $T_x = \sum_{i: Y_i \geq x} Y_i$. Then $E[T_{x^*}] = m \int_{x^*}^\infty y f(y) dy$ by linearity of expectation. Moreover, the expected value of $|T_{x^*} - S_{\rho^*}|$ is bounded by $O(\sqrt{m})$, so that $E[T_{x^*}]/m$ approaches $E[S_{\rho^*}]/m$. Since S grows linearly with m, the claim follows. □

It is easy to numerically evaluate the constant $\rho^* \approx 0.239$[2]. For $\rho < \rho^*$, S_ρ/S is bounded away from $\frac{1}{2}$, with high probability. The following lemma, that we shall prove in the next section, formalizes this.

LEMMA 3. *Let $X_1, \ldots, X_m$ be i.i.d. $N(0,1)$ random variables and let $Y_1, \ldots, Y_m$ be the sorted ordering (in decreasing order) of $|X_1|, \ldots, |X_m|$. Let S_ρ and ρ^* be as above. Then for any $\rho < \rho^*$, there is a $\delta > 0$ and a $c > 0$ such that*

$$Pr[S_\rho > S(\frac{1}{2} - \delta)] \leq \exp(-cm)$$

To get some intuition on why this lemma is useful, we sketch a proof of Theorem 1 for the zero small-error case using it. Let $x \in \mathbb{R}^n$ and let A be an $m \times n$ matrix of i.i.d. Gaussians. Let $y' = Ax + e$ such that e has support at most ρm for a constant $\rho < \rho^*$. Then we wish to show that x is indeed the optimum solution to the linear program, i.e. for any $x' \in \mathbb{R}^n, x' \neq x$, $|y' - Ax'| > |y' - Ax|$.

$y' - Ax$ is precisely the vector e, which has support at most ρm. On the other hand, $y' - Ax' = y' - Ax - A(x' - x) =$

[2] Or to be slightly more precise, 0.239031891449516803895...

$e - Az$ where $z = x' - x$. Let T be the support of e and let $|w|_S$ denote $\sum_{i \in S} |w_i|$ for any vector w, any subset $S \subseteq [m]$.

Suppose that LP decoding fails and that there is an x, x', e such that $|y' - Ax'| \leq |y' - Ax|$. Rewriting, we get $|e - Az|_T + |e - Az|_{T^c} \leq |e|_T$. Using the triangle inequality, we have $|e|_T \leq |e - Az|_T + |Az|_T$. Adding the two inequalities and noting that e is zero on T^c, we get $|Az|_{T^c} \leq |Az|_T$. Now note that since z is non-zero, and each entry of A is an i.i.d. Gaussian, each coordinate of the the vector (Az) is an i.i.d. Gaussian. The inequality $|Az|_{T^c} \leq |Az|_T$ says that the sum of some ρm of m i.i.d. Gaussians is larger than the sum of the remaining $(1 - \rho)m$ of them. The lemma above says that for suitable ρ, this happens with probability no larger than e^{-cm}. Thus any one z is exponentially unlikely to be a candidate for $x' - x$. We shall show later that in fact the probability that any z is a candidate is exponentially small. The result would then follow.

In the next subsection, we prove Lemma 3. We then use a net argument to show that with high probability, no large z is a candidate for $x' - x$ if the error vector e is (ρ,α)-small. We put things together to prove Theorem 1 in section 3.3.

3.1 Concentration for a single point

In this section, we prove the following concentration result for S_ρ.

LEMMA 4. *Let $X_1, \ldots, X_m$, $Y_1, \ldots, Y_m$, S_ρ and S be as above. Then for any ρ, and any $\delta > 0$, there is a constant $c_4 > 0$ such that with probability $(1 - exp(-c_4 m))$, $|S_\rho - ES_\rho| \leq \delta S$.*

PROOF. Let M_ρ denote the median value of $S_\rho = S_\rho(\vec{X})$. Note that with respect to the ℓ_1 norm, the function S_ρ is a Lipschitz with Lipschitz constant 1. In other words, if X and X' differ only in co-ordinate i, $|S_\rho(X) - S_\rho(X')| \leq |X_i - X'_i|$. Thus with respect to ℓ_2 norm, S_ρ is Lipschitz with Lipschitz constant $\sqrt{m}$.

Now we use the isoperimetric inequality for the Gaussian measure (see e.g. [24]). This says that for any set A with measure at least a half, $(A_t)^c$ has measure at most $e^{-t^2/2}$, where $A_t = \{x \in \mathbb{R}^m : d(x, A) \leq t\}$, where $d(x, A)$ is defined naturally as $\inf_{y \in A} \|x - y\|$. Taking A to be the set $\{x \in \mathbb{R}^m : S_\rho(x) \geq M_\rho\}$, we get

$$Pr[d(x, A) \leq t] \geq 1 - e^{-t^2/2}$$

By the Lipschitz condition, $[d(x, A) \leq t]$ implies that $[S_\rho(x) \geq M_\rho - t\sqrt{m}]$. Thus with probability $1 - e^{-t^2/2}$, $S_\rho(x) \geq M_\rho - t\sqrt{m}$. Similarly, we get that with probability at least $1 - e^{-t^2/2}$, $S_\rho(x) \leq M_\rho + t\sqrt{m}$.

We now use this fact in two different ways. First we use this to bound the difference between the mean and the median of S_ρ:

$$\begin{aligned}
|M_\rho - E[S_\rho]| &\leq E[|M_\rho - S_\rho|] \\
&= \int_y Pr[|M_\rho - S_\rho| \geq y] dy \\
&\leq \int_y 2e^{-y^2/2m} dy \\
&\leq c_5/\sqrt{m}
\end{aligned}$$

for a constant c_5.

Moreover, setting t to be $\delta S/2\sqrt{m}$, we get that with probability at least $(1 - 2exp(-t^2/2))$, $|S_\rho(x) - M_\rho| \leq S\delta/2$. Noting that $S = c_6 m$ for a constant c_6, we get the failure probability above to be $exp(-c_4 m)$. The result follows. □

COROLLARY 5. *For any $\rho < \rho^*$, there is a $\delta > 0$ and a $c_7 > 0$ such that with probability $(1 - exp(-c_7 m))$, $S_\rho \leq (\frac{1}{2} - \delta)S$.*

PROOF. Note that there is a constant $\delta > 0$ such that for large enough m, $E[S_\rho]/S \leq \frac{1}{2} - 2\delta$. Indeed $E[S_\rho] = E[S_{\rho^*}] - \sum_{i=\lceil \rho m \rceil + 1}^{\lceil \rho^* m \rceil} E[Y_i]$. Each of the Y_i's in the summation has expectation at least as large as EY. Thus $E[S_\rho]/S \leq \frac{1}{2} - \frac{(\rho^* - \rho) m E[Y]}{m E[Y]} \leq \frac{1}{2} - 2\delta$ for a suitable constant δ. The result then follows from the lemma. □

To facilitate the net argument in the next section, we also note that

COROLLARY 6. *For any $\varepsilon > 0$, there exists a constant $c_8 > 0$ such that with probability $(1 - exp(-c_8 m))$, it holds that $(1 - \varepsilon)S \leq S_1 \leq (1 + \varepsilon)S$.*

PROOF. Follows immediately by taking $\rho = 1$. □

3.2 The net argument

We now show how to extend the argument to all of $\mathbb{R}^n$. For a given $\gamma > 0$, let $v_1, \ldots, v_K$ be a set of points, $\|v_i\| = 1$ such that for any z, $\|z\| = 1$, there is an $i \in [K]$ such that $\|z - v_i\| \leq \gamma$. Such sets of points exist with $K = (\frac{1}{\gamma})^{O(n)}$; this can be easily shown by taking a maximal set of points in S^{n-1} such that any two of them are at distance at least γ.

For $z \in S^{n-1}$, if A is an $m \times n$ matrix with $N(0, 1)$ entries, each entry in Az is distributed as $N(0, 1)$. From the two lemmas in the last section, with probability $1 - exp(-cm)$, we have

- $S_\rho(Az) \leq S(\frac{1}{2} - \delta)$, and
- $(1 - \varepsilon)S \leq S_1(Az) \leq (1 + \varepsilon)S$.

Taking $m = c_9 n$ for large enough c_9, we can use a union bound over the net points $v_1, \ldots, v_K$ to show that with probability at least $(1 - exp(-c_{10} n))$, each of the net points satisfies the above two properties.

We now show that whenever this is true for a small enough γ, a similar claim is true for all points $z \in \mathbb{R}^n$.

LEMMA 7. *Given any $\rho < \rho^*$, there exist absolute constants $c_{11}, c_{12}, \delta > 0$ such that the following holds. Whenever $m \geq c_{11} n$ and n is large enough, with probability $(1 - exp(-c_{12} n))$, an $m \times n$ matrix A with independent $N(0, 1)$ entries has the following property: for every vector z and every subset $I \subseteq [m]$ with $|I| \leq \rho m$, $|Az| - 2|Az|_I \geq \delta S\|z\|$.*

PROOF. Let $z \in S^{n-1}$ be arbitrary. Since N is a γ-net, there exists $v_0 \in N$ such that $\|z - v_0\| \leq \gamma$. Let z_1 denote $z - v_0$ and let $\gamma_1 < \gamma$ denote $\|z_1\|$. Using the net property again, we get a $v_1 \in N$ such that $\|z_1 - \gamma_1 v_1\| \leq \gamma\gamma_1$. Again we can denote by z_2 the vector $z_1 - \gamma_1 v_1$ and its norm by γ_2; clearly $\gamma_2 \leq \gamma^2$. Repeating this argument, we get a sequence of net points $v_0, v_1, \ldots$ and real numbers $\gamma_1, \gamma_2, \ldots$ so that

$$z = \sum_{i \geq 0} \gamma_i v_i$$

where $\gamma_0 = 1$ and $\gamma_i \leq \gamma^i$.

By scaling, we get that for any $z \in \mathbb{R}^n$, we can write

$$z = \|z\| \sum_{i \geq 0} \gamma_i v_i.$$

Next consider a particular index set $I \subset [m]$ with $|I| \leq \rho m$. Thus we have that:

$$\begin{aligned} |Az|_I &\leq \|z\| \sum_{i \geq 0} \gamma_i |Av_i|_I \\ &\leq \|z\| \sum_{i \geq 0} \gamma_i S(\frac{1}{2} - \delta) \\ &\leq \|z\| \sum_{i \geq 0} \gamma^i S(\frac{1}{2} - \delta) \\ &= S\|z\| \frac{\frac{1}{2} - \delta}{(1 - \gamma)} \end{aligned}$$

Also,

$$\begin{aligned} |Az|_1 &\geq \|z\|(|Av_0|_1 - \sum_{i \geq 1} \gamma_i |Av_i|_1) \\ &\geq (1 - \varepsilon)S\|z\| - \sum_{i \geq i} \gamma^i (1 + \varepsilon) S\|z\| \\ &\geq (1 - \varepsilon - \frac{\gamma(1 + \varepsilon)}{(1 - \gamma)}) S\|z\| \end{aligned}$$

Thus it follows that $|Az| - 2|Az|_I \geq S\|z\|(2\delta - \varepsilon - \frac{\gamma(1+\varepsilon)}{(1-\gamma)})$. For a given $\delta = \delta_\rho$, we can pick ε and γ small enough so that $|Az| - 2|Az|_I \geq \delta S\|z\|$. □

3.3 Putting it together

Let A be the encoding matrix so that a vector x is encoded as $y = Ax$. Let the vector of answers by $y' = y + e$ such that e is (ρ,α)-small. Suppose that the ℓ_1-minimizer is a vector x'. Let T be the set of indices where e is large, i.e. $T = \{i : e_i \geq \alpha\}$.

Since x' is the ℓ_1-minimizer for $y' = y + e$,

$$|Ax - y'| \geq |Ax' - y'|$$

Rewriting $y' = Ax + e$ and setting $z = (x' - x)$, we get

$$|e| \geq |Az - e|$$

Now note that

$$\begin{aligned} |e| &= |e|_T + |e|_{T^c} \\ &\leq |Az - e|_T + |Az|_T + \alpha|T^c| \end{aligned}$$

and

$$\begin{aligned} |Az - e| &= |Az - e|_T + |Az - e|_{T^c} \\ &\geq |Az - e|_T + |Az|_{T^c} - |e|_{T^c} \\ &\geq |Az - e|_T + |Az|_{T^c} - \alpha|T^c| \end{aligned}$$

Thus we conclude

$$|Az|_T + \alpha|T^c| \geq |Az|_{T^c} - \alpha|T^c| \qquad (1)$$

or

$$|Az| - 2|Az|_T \leq 2\alpha|T^c| \qquad (2)$$

Combining with lemma 7, we get that $\|z\| \leq 2\alpha|T^c|/\delta S = c_{14}\alpha$. Thus we have proved Theorem 1.

Theorem 1. *Given any $\rho < \rho^*$, there exist absolute constants $c_1, c_2, c_3 > 0$ such that the following holds. Whenever $m \geq c_1 n$ and n is large enough, with probability $(1 - exp(-c_2 n))$, an $m \times n$ matrix A with independent $N(0,1)$ entries has the following property: for every vector x and every error vector e that is (ρ,α)-small, the vector x' reconstructed by the LP decoding procedure is such that $\|x' - x\| \leq c_3\alpha$.*

We also record the $\alpha = 0$ case of the above theorem, which allows for perfect reconstruction.

Corollary 8. *Given any $\rho < \rho^*$, there exist absolute constants $c_1, c_2, c_3 > 0$ such that the following holds. Whenever $m \geq c_1 n$ and n is large enough, with probability $(1 - exp(-c_2 n))$, an $m \times n$ matrix A with independent $N(0,1)$ entries has the following property: for every vector x and every error vector e with support at most ρm, the LP decoding procedure reconstructs the vector x.*

We note that the constant c_1 grows approximately as $2\pi/(\rho^* - \rho)^3$ in our proof.

3.4 Lower Bounds for Privacy

Note that if the rounded vector $\hat{x}$ differs from x in bit i, then $(x' - x)_i$ is at least $\frac{1}{2}$, so that $\|x' - x\|^2 \geq \frac{1}{4}|\{i : x_i \neq \hat{x}_i\}|$. From Theorem 1, $\|x' - x\| \leq c_3\alpha$. Thus $|\{i : x_i \neq \hat{x}_i\}| \leq (2c_3\alpha)^2$. Thus we have shown that:

Theorem 9 (Blatant non-privacy: Gaussian queries). *Given any $\rho < \rho^*$, there exist absolute constants $c_1, c_2, c_3 > 0$ such that the following holds. There exists an efficient attacker that asks $m \leq c_1 n$ queries of the form $\sum_j a_{ij} x_j$, and given answers that are within an error of α for all but a ρ fraction of the questions, with probability $(1 - exp(-c_2 n))$, reconstructs a database $\hat{x}$ such that $\hat{x}$ agrees with x on all but $(2c_3\alpha)^2$ of the entries.*

3.5 Semi-oblivious noise: fixed support

Suppose that the support of the wild noise T is chosen non-adversarially. In this case, we argue that $|T|$ can be as large as $(\frac{1}{2} - \epsilon)m$. Indeed, in the above proof, if T is chosen without looking at A, then for a fixed z, the best error vector has $e|_T = Az|_T$. From corollary 6, $|Az| \geq (1 - \delta)S$ and $|Az|_T \leq (\frac{1}{2} - \delta)S$ with probability $(1 - exp(-cm))$ for suitable constants δ, c. The rest of the argument goes through unchanged and we conclude that:

Theorem 10. *Given any $\rho < \frac{1}{2}$, there exist absolute constants $c_1', c_2', c_3' > 0$ such that for any $T \subseteq [m]$ with $|T| \leq \rho m$, the following holds. Whenever $m \geq c_1' n$ and n is large enough, with probability $(1 - exp(-c_2' n))$, an $m \times n$ matrix A with independent $N(0,1)$ entries has the following property: for every vector x and every error vector e such that $|e_i| \leq \alpha$ for any $i \notin T$, the vector x' reconstructed by the LP decoding procedure is such that $\|x' - x\| \leq c_3'\alpha$.*

3.6 Oblivious Symmetric noise

We next consider a non-adversarial noise model where each entry of $e|_T$ is chosen from a symmetric zero-mean probability distribution. In this setting, we shall show that T as large as $(1 - \varepsilon)m$ still allows for recovery of x. In fact, we shall only use the fact that $Pr[e_i > 0] \le \frac{1}{2}$ and $Pr[e_i < 0] \le \frac{1}{2}$.

Given a candidate bad z and an error vector e, we construct another error vector e' such that if $|Az - e| < |e|$ then $|Az - e'| < |e'|$, and $e'_i \in \{0, (Az)_i\}$ for $i \in T$. Indeed for any $i \in T$, we set e'_i to 0 if $|e_i| \ge |e_i - (Az)_i|$, and to $(Az)_i$ otherwise. For $i \notin T$, e'_i is set to e_i. It is easy to verify that e' satisfies the desired condition and that $Pr[e'_i = 0] \ge \frac{1}{2}$, where the probability is taken over the random choice of e_i.

Each of the terms $|e'_i| - |(Az)_i - e'_i|$, for $i \in T$ has mean at most zero and variance at most c for some constant c. Thus the probability that $\sum_{i \in T} |e'_i| - |(Az)_i - e'_i|$ exceeds $\delta|T|$ for any $\delta > 0$ is, by Chernoff bounds, at most $exp(-\delta^2 |T|^2 / 2c|T|)$ $\le exp(-c'|T|)$ for some constant c'. Moreover, for any $i \notin T$, $|e'_i| - |(Az)_i - e'_i| \le 2\alpha - |(Az)_i|$. Thus $\sum_{i \notin T} |e'_i| - |(Az)_i - e'_i|$ is at most $2\alpha|T^c| - c''\|z\||T^c|$ with probability $(1 - exp(-c''|T^c|))$ for constant c'', c'''. For LP-decoding to fail, $\sum_{i \in [m]} |e'_i| - |(Az)_i - e'_i|$ must exceed zero. For $|T| < (1 - \varepsilon)m$ and $\|z\| \ge c''_3 \alpha$, this happens with probability exponentially small in m. The rest of the argument is similar to the adversarial-error case. Thus we have proved that

THEOREM 11. *Given any $\rho < 1$ and any $T \subseteq [m]$ with $|T| \le \rho m$, there exist absolute constants $c''_1, c''_2, c''_3 > 0$ such that the following holds. Whenever $m \ge c''_1 n$ and n is large enough, with probability $(1 - exp(-c''_2 n))$, an $m \times n$ matrix A with independent $N(0,1)$ entries has the following property: for every vector x and every error vector e such that $|e_i| \le \alpha$ for any $i \notin T$, and e_i is drawn from a symmetric zero-mean distribution for $i \in T$, the vector x' reconstructed by the LP decoding procedure is such that $\|x' - x\| \le c''_3 \alpha$.*

We note the proof above also establishes that if *each* entry of the error vector e were drawn independently from a probability distribution $\mathcal{D}_i$ such that (a) $Pr_{\mathcal{D}_i}[E > \alpha] \le \frac{1}{2}$,(b) $Pr_{\mathcal{D}_i}[E < -\alpha] \le \frac{1}{2}$, and (c) $Pr_{\mathcal{D}_i}[|E| \ge \alpha] \le 1 - \epsilon$ for some fixed $\epsilon > 0$. Then LP decoding reconstructs a vector x' such that $\|x' - x\| \le O(\alpha)$. Thus in the privacy setting, if the curator was adding independent near-symmetric noise to every answer, the noise added to *almost every* answer should be $\Omega(\sqrt{n})$ to avoid blatant non-privacy.

3.7 LP decoding fails beyond ρ^*

In this section, we show that for any error rate larger than ρ^*, LP decoding can almost surely be made to fail in a very strong sense when A comes from the Gaussian ensemble. With overwhelming probability, the matrix A has the following property: given an original vector x, and an arbitrary vector x' of the adversary's choosing, the adversary can choose a (ρ,0)-small error vector e such that LP decoding would reconstruct x' when given $y' = Ax + e$.

The adversary sets $z = x' - x$ and computes the vector Az. Note that each entry of Az is an independent $N(0, \|z\|^2)$ Gaussian random variable. From lemma 4, with high probability, the sum of the highest ρm of $(Az)_i$'s is more than $S\|z\|/2$, and hence the error vector e which agrees with Az on these largest ρm entries, and is zero elsewhere has a smaller ℓ_1 distance to z than to 0. In particular, this means that in the case when a $(1 - \rho)$ fraction of the answers are given accurately, LP decoding does not recover the true signal. Thus LP decoding can be made to fail, with high probability, when the matrix A has independent $N(0,1)$ entries, beyond any error rate larger than ρ^*.

In other words, the value ρ^* captures exactly the permissible error rate when an adversary is allowed to arbitrarily change a small fraction of the entries.

4. LP DECODING: THE ± 1'S CASE

In this section, we show that if each entry of A is chosen from one of the following two probability distributions, the results of the previous section continue to hold:

$$a_{ij} = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases}$$

$$a_{ij} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}$$

We first show the measure concentration result analogous to Lemma 4.

4.1 Concentration for a single point

Let $z \in S^{n-1}$ be arbitrary and let $X_i = \sum_j a_{ij} z_j$. Since z is fixed in this section, various quantities such as M_ρ that depend on z will be used without an explicit parametrization. Note that the X_i's are independent and identically distributed. We first show some properties of this distribution. The following lemma is due to Achlioptas [1]

LEMMA 12 ([1]). *Let X_i be as above. Then $E[X_i^2] = 1$. Moreover, let $Q = \sum_{j=1}^{k} X_i^2$. Then for any $\varepsilon > 0$,*

$$Pr[Q > (1 + \varepsilon)k] < ((1 + \varepsilon)exp(-\varepsilon))^{\frac{k}{2}}$$

$$Pr[Q < (1 - \varepsilon)k] < exp(-\frac{k}{2}(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}))$$

COROLLARY 13. *There is a universal constant τ such that for any z and n, if X_i as above, $E[|X_i|] \ge \tau$.*

PROOF. For $k = 1$, we get that $Pr[X_i^2 < \frac{1}{4}]$ is bounded away from 1. Thus with constant probability, $|X_i| > \frac{1}{2}$. □

As before, given $\vec{X} = (X_1, \ldots, X_m)$, let $Y_1, \ldots, Y_m$ be sorted ordering of $|X_i|$'s and let S_ρ be as before. We shall use Talagrand's inequality that we state next.

Let (Ω, μ) be a probability distribution and let $\vec{X} \in \Omega^m$ be a sequence of independent samples from X. Let $A \subseteq \Omega^m$ be a set such that $Pr(A) \ge \frac{1}{2}$. Talagrand's inequality [25] says that

$$Pr[d(\vec{X}, A) > t] \le 2e^{-t^2/4}$$

where $d(\vec{X}, A)$ denotes the *convex distance* of the point $\vec{X}$ from the set A. Here the convex distance is defined as follows: we say that $d(\vec{X}, A) \le t$, if for every vector $\vec{\alpha} = (\alpha_1, \ldots, \alpha_m)$, there exists an $\vec{X}' \in A$ such that

$$\sum_{i : X_i \neq X'_i} \alpha_i \le t\|\alpha\|$$

We now use this to show that S_ρ is well concentrated around its median.

LEMMA 14. *Let X, Y and S_ρ be as above. Let M_ρ be the median value of S_ρ. Then for any $0 < s < m$, with probability $(1 - 5exp(-s^2/8m))$, $|S_\rho - M_\rho| < s$. Moreover, for $s \ge m$, $Pr[|S_\rho - M_\rho| > s] \le 5exp(-s/24)$.*

PROOF. Let $\vec{X}$ denote $X_1, \ldots, X_m$. Let $A = \{\vec{X} \in \Omega^m : S_\rho(\vec{X}) \leq M_\rho\}$. Talagrand's inequality says that

$$Pr[d(\vec{X}, A) > t] \leq 2e^{-t^2/4}$$

Let $\vec{X}$ be such that $d(\vec{X}, A) \leq t$ and $\|\vec{X}\| \leq \sqrt{2m}$. Consider the vector $\vec{\alpha} = |X_1|, \ldots, |X_m|$, i.e. $\alpha_i = |X_i|$. Let $\vec{X}'$ be the corresponding point in A guaranteed by the convex distance definition. Then

$$\begin{aligned} S_\rho(\vec{X}) - S_\rho(\vec{X}') &= \sum_{i:|X_i| \neq |X_i'|} (|X_i| - |X_i'|) \\ &\leq \sum_{i:|X_i| > |X_i'|} |X_i| \\ &\leq \sum_{i:X_i \neq X_i'} |X_i| \\ &\leq t\|\vec{X}\| \end{aligned}$$

Setting $t = s/\sqrt{2m}$, we get that with probability $1 - 2exp(-s^2/8m)$, $S_\rho \leq M_\rho + s$. Using a similar argument with $A' = \{\vec{X} : S_\rho(\vec{X}) \geq M_\rho\}$, we can conclude that with probability $1 - 2exp(-s^2/8m)$, $S_\rho > M_\rho - s$.

Note that we assumed that $\|\vec{X}\| \leq \sqrt{2m}$. However, by lemma 12, this is true with probability $(1 - exp(-m/8))$ This implies the first part of the claim.

Next we prove the second part. First note that lemma 12 implies that $Pr[\|\vec{X}\| > \sqrt{rm}] \leq (r/e^{r-1})^{\frac{m}{2}} \leq exp(-mr/4)$ for $r \geq 6$. Moreover, assuming that $\|\vec{X}\| \leq \sqrt{rm}$, we can use $t = s/\sqrt{rm}$ in the above argument. Taking $r = \frac{6s}{m}$ gives an overall failure probability no larger than $5exp(-s/24)$. □

COROLLARY 15. *The mean and the median of S_ρ converge.*

PROOF. Observe that

$$\begin{aligned} |M_\rho - E[S_\rho]| &\leq E[|M_\rho - S_\rho(x)|] \\ &= \int_y Pr[|M_\rho - S_\rho| \geq y]dy \\ &\leq \int_0^m 5exp(-y^2/8m)dy + \int_m^\infty 5exp(-y/24)dy \\ &\leq c_{15}/\sqrt{m} + c_{16}exp(-m/24) \end{aligned}$$

Hence the claim. □

COROLLARY 16. *For any $\varepsilon > 0$, there exists a $c_{16} > 0$ such that with probability $(1 - exp(-c_{16}m))$, $(1 - \varepsilon)E[S_1] \leq S_1 \leq (1 + \varepsilon)E[S_1]$.*

PROOF. Follows immediately from lemma 14 and corollaries 13 and 15. □

4.2 Lower bounds on $\rho^*_{\pm 1}$

Given a $z \in S^{n-1}$, the distribution of $X_i = \sum_j a_{ij}z_j$ is well defined, and there is a value $\rho^*(z)$ such that whenever $\rho < \rho^*(z)$, $E[S_\rho] < \frac{1}{2}E[S_1]$. In this section, we address the question of how small $\rho^*(z)$ can be. We shall give two bounds: the first one holds for all z and shows that $\rho^*_{\pm 1} \overset{def}{=} \inf_{z \in \mathbb{R}^n} \rho^*(z)$ is bounded below by a small positive constant independent of z. The second bound gives much better lower bounds assuming that $\|z\|_3^3/\|z\|_2^3$ is small.

THEOREM 17. *There is a universal constant $\hat{\rho}$ such that the following holds. Let X be any random variable with $E[|X|^2] = 1$ that satisfies for any $\varepsilon > 0$:*

$$Pr[X^2 > (1 + \varepsilon)] < ((1 + \varepsilon)exp(-\varepsilon))^{\frac{1}{2}}, \textit{and}$$

$$Pr[X^2 < (1 - \varepsilon)] < (exp(-\frac{\varepsilon^2}{4} + \frac{\varepsilon^3}{6})).$$

Then for any x such that $Pr[|X| \geq x] \leq \hat{\rho}$, it holds that $\int_x^\infty yf(y)dy \leq \frac{1}{2}E[|X|]$, where $f(\cdot)$ denotes the p.d.f. of $|X|$.

PROOF. Recall that from corollary 13, under the assumptions of the theorem, $E[|X|]$ is bounded below by a universal constant τ. Let x be such that $Pr[|X| \geq x] \leq \hat{\rho}$. Now observe that for any $y > x$, $Pr[|X| > y] \leq \min(\hat{\rho}, (ey/e^y)^{\frac{1}{2}})$. Thus

$$\begin{aligned} \int_x^\infty yf(y)dy &= \int_x^\infty Pr[|X| > y]dy \\ &\leq \int_x^\infty \min(\hat{\rho}, (ey/e^y)^{\frac{1}{2}})dy \\ &\leq \int_0^{x'} \hat{\rho}dy + \int_{x'}^\infty (ey/e^y)^{\frac{1}{2}}dy \end{aligned}$$

for and $x' \geq x$. Note that the second integral goes to zero as x' approaches infinity. Let x' be such that second integral is at most $\frac{\tau}{4}$. Taking $\rho < \frac{\tau}{4x'}$ then suffices. □

For the second bound, we use a quantitative version of the central limit theorem. Goldstein [23, Prop 3.2] shows that for X_i as above (with $\|z\| = 1$), and G a gaussian with variance one,

$$d_{EM}(X_i, G) \leq 3\sum_j |z_j|^3,$$

where d_{EM} denotes the *earthmover* or *Kantarovich* distance between two distributions. Recall that the Kantarovich distance between X and Y is $\inf E|X - Y|$, where the infimum is over all couplings of X and Y that have the correct marginals. This implies that there is a coupling function $f : \mathbb{R} \rightarrow \mathbb{R}$ such that $f(G) \sim X_i$, and $E_{x \sim G}[|x - f(x)|] = d_{EM}(X_i, G)$. Also note that both X_i and G are symmetric distributions so one can assume that f is symmetric and $sgn(f(x)) = sgn(x)$. Thus $\int_0^\infty |x - f(x)|\mu_G(x) \leq \frac{1}{2}d_{EM}(X_i, G)$.

Let $a \geq 0$ be arbitrary and let us consider the integral $\int_a^\infty f(x)\mu_G(x)dx$. Clearly $|\int_a^\infty f(x)\mu_G(x)dx - \int_a^\infty x\mu_G(x)dx|$ is at most $\int_a^\infty |x - f(x)|\mu_G(x)dx \leq \frac{1}{2}d_{EM}(X_i, G)$.

Suppose that ρ is such that $\int_{\Phi^{-1}(1-\rho)}^\infty x\mu_G(x)dx \leq (\frac{1}{2} - \delta)\int_0^\infty x\mu_G(x)dx$ (Φ here is the c.d.f. of a Gaussian with mean zero and variance one). Then we have that

$$\begin{aligned}
& \int_{\Phi^{-1}(1-\rho)}^{\infty} f(x)\mu_G(x)dx \\
\leq \; & \int_{\Phi^{-1}(1-\rho)}^{\infty} x\mu_G(x)dx + \frac{1}{2}d_{EM}(X_i, G) \\
\leq \; & (\frac{1}{2} - \delta)\int_{0}^{\infty} x\mu_G(x)dx + \frac{1}{2}d_{EM}(X_i, G) \\
\leq \; & \frac{1}{2}\int_{0}^{\infty} x\mu_G(x)dx - \delta\sqrt{\frac{1}{2\pi}} + \frac{1}{2}d_{EM}(X_i, G) \\
\leq \; & \frac{1}{2}\int_{0}^{\infty} f(x)\mu_G(x)dx + d_{EM}(X_i, G) - \delta\sqrt{\frac{1}{2\pi}} \\
\leq \; & \frac{1}{2}\int_{0}^{\infty} f(x)\mu_G(x)dx + 3\sum_j |z_j|^3 - \delta\sqrt{\frac{1}{2\pi}}
\end{aligned}$$

Thus whenever, $3\sum_j |z_j|^3 \leq \delta\sqrt{\frac{1}{2\pi}}$, the value ρ is less than $\rho^*(z)$. We record this as

Lemma 18. *There exists a constant c_{17} such that for any $z \in S^{n-1}$, we have $\rho^*(z) \geq \rho^* - c_{17}\sum_j |z_j|^3$.*

4.3 Net Argument

We note that the net argument of section 3.2 only used corollaries 5 and 6. Since we proved their analogues for the ± 1 case in the previous subsection, we can conclude that, for a constant $\rho^*_{\pm 1}$, the following holds.

Lemma 19. *Given any $\rho < \rho^*_{\pm 1}$, there exist absolute constants $c_{18}, c_{19}, \delta' > 0$ such that the following holds. Whenever $m \geq c_{18}n$ and n is large enough, with probability $(1 - exp(-c_{19}n))$, an $m \times n$ matrix A with independent entries from one of the distributions above has the following property: for every vector z and every subset $I \subseteq [m]$ with $|I| \leq \rho m$, $|Az| - 2|Az|_I \geq \delta' S\|z\|$.*

When we restrict ourselves to sets of the form $\{z : \|z\|_3^3/\|z\|_2^3 < \varepsilon_1\}$ for some $\varepsilon_1 > 0$, the argument of the previous section does not work any more; since when we write

$$z = \|z\| \sum_{i \geq 0} \gamma_i v_i$$

where $\gamma_0 = 1$ and $\gamma_i \leq \gamma^i$, the net points $\{v_i : i \geq 1\}$ do not satisfy the upper bound on the three norm. However for any index set $I \subset [m]$, $|Av_i|_I$ is still bounded above by $S(1+\varepsilon)$, which suffices for the argument (with a slightly worse constant). The rest of the argument goes through unchanged. Thus it follows that

Lemma 20. *Given any $\rho < \rho^*$, there exist absolute constants $c_{20}, c_{21}, c_{22}, c_{23} > 0$ such that the following holds. Whenever $m \geq c_{20}n$ and n is large enough, with probability $(1 - exp(-c_{21}n))$, an $m \times n$ matrix A with entries drawn from one of the distribution described above has the following property: for every vector z and every subset $I \subseteq [m]$ with $|I| \leq \rho m$, either $\|z\|_3^3 \geq c_{22}\|z\|_2^3$ or $|Az| - 2|Az|_I \geq c_{23}S\|z\|$.*

4.4 Putting it together

Analogous to the Gaussian case, we have proved that

Theorem 21. *Given any $\rho < \rho^*_{\pm 1}$, there exist absolute constants $c_{24}, c_{25}, c_{26} > 0$ such that the following holds. Whenever $m \geq c_{24}n$ and n is large enough, with probability $(1 - exp(-c_{25}n))$, an $m \times n$ matrix A with entries drawn from one of the distribution described above has the following property: for every vector x and every error vector e that is $(\rho, \alpha) -$ small, the vector x' reconstructed by the LP decoding procedure is such that $\|x' - x\| \leq c_{26}\alpha$.*

Moreover, from lemma 20, it follows that

Theorem 22. *Given any $\rho < \rho^*$, there exist absolute constants $c_{27}, c_{28}, c_{29}, c_{30} > 0$ such that the following holds. Whenever $m \geq c_{27}n$ and n is large enough, with probability $(1 - exp(-c_{28}n))$, an $m \times n$ matrix A with entries drawn from one of the distribution described above has the following property: for every vector x and every error vector e that is $(\rho, \alpha) -$ small, the vector x' reconstructed by the LP decoding procedure is such that either $\|x' - x\|_3^3 \geq c_{29}\|x' - x\|_2^3$ or $\|x' - x\| \leq c_{30}\alpha$.*

4.5 Privacy Implications

In the setting when we round the vector x' to a binary database $\hat{x}$, we have that $|x'_j - x_j| \geq \frac{1}{2}$ whenever $\hat{x}_j \neq x_j$. Further, we can assume without loss of generality that $|x'_j - x_j| \leq 1$, since we can add the constraint $0 \leq x'_j \leq 1$ in the linear program. Suppose that $|\{j : \hat{x}_j \neq x_j\}| \geq B$. Then $\|x' - x\| \geq B/4$. Let $z = (x' - x)/\|x' - x\|$ so that $z_j \leq 4/B$ for all j. Thus $\|z\|_3^3 \leq \frac{4}{B}\|z\| = 4/B$. Thus, if $B \geq 4/c_{29}$, then $\|x' - x\| \leq c_{30}\alpha$ so that $B \leq (2c_{30}\alpha)^2$. Thus we have shown that

Theorem 23 (Blatant non-privacy for ± 1 queries). *Given any $\rho < \rho^*$, there exist absolute constants $c_{31}, c_{32}, c_{33}, c_{34} > 0$ such that the following holds. There exists an efficient attacker that asks $m \leq c_{31}n$ queries of the form $\sum_j a_{ij}x_j$ where each a_{ij} is ± 1, and given answers that are within an error of α for all but a ρ fraction of the questions, with probability $(1 - exp(-c_{32}n))$, reconstructs a database $\hat{x}$ such that $\hat{x}$ agree with x on all but* $\max\{c_{33}, (c_{34}\alpha)^2\}$ *of the entries.*

5. COMPRESSED SENSING IMPLICATIONS

Donoho [16], provides the following motivation:

> “As our modern technology-driven civilization acquires and exploits ever-increasing amounts of data, ’everyone’ now knows that most of the data we acquire ’can be thrown away’ with almost no perceptual loss – witness the broad success of lossy compression formats for sounds, images, and specialized technical data. The phenomenon of ubiquitous compressibility aises very natural questions: *why go to so much effort to acquire* **all** *the data when* **most** *of what we get will be thrown away? Can’t we just* **directly measure** *the part that won’t end up being thrown away?*”

In compressed sensing there is a sparse signal of length n and the goal is to reconstruct the signal using few linear measurements. The signal corresponds to the error vector in the error-correction paradigm. When the coding matrix A is $m \times n$, the $(m - n) \times m$ matrix A^* annhilates the range of A: $A^*(Ax + e) = A^*e$. Thus, the quantity of interest is the number of measurements, which is $m - n$.

Previous results handle signals having support t with $m - n \in O(t \log^c n)$ measurements [11, 3], and support up to $m/3000$ when $m - n \approx 3m/4$. Our proof shows that LP decoding based compressed sensing works for any density smaller than $\rho^* \approx 0.239$, with the number of questions $m - n = (1 - \varepsilon)m$, where $\varepsilon = n/m$ in the proof above. Thus any vector z whose support is smaller than $\rho^* m$ can be reconstructed using $(1 - \varepsilon)m$ Gaussian measurements.

Moreover, even with $m - \omega(1)$ questions, the LP decoding procedure can be made to fail if the sparseness of the input is allowed to be larger than ρ^*. In other words, for almost every measurement matrix A with $m - \omega(1)$ rows and every $\rho > \rho^*$, there is a vector with support size at most ρm that will not be reconstructed Of course for m questions, the sparseness can be as large as 1. Thus there is a phase transition at ρ^*.

6. ERROR CORRECTING CODES

One consequence of the results above is that given a vector in $x \in \{0, 1\}^n$, one can encode it by picking a random ± 1 matrix A, and encoding x as Ax. We note that this computation is over integers so that this is not a random linear code in the sense usually used in coding theory. Each symbol in a codeword would be $O(\log n)$ bits long, and given a $(1 - \rho)$ fraction of codewords for any $\rho < \rho^*_{\pm 1}$, the decoding procedure above would reconstruct the original vector x. Moreover, if the error rate $\rho < \rho^*$, the reconstructed vector is correct except possibly at a constant number of bits.

This can be fixed in two ways. An easy fix is to concatenate this code with any standard error-correcting code that can handle a constant number of errors. Alternately, we can take A to be a Gaussian matrix and use the results of Section 3. Moreover, the coefficients can be made integral by scaling up the Gaussians by a factor of $c_{35}n$ and rounding to the nearest integer. It is easy to see that this gives an integer matrix with $O(\log n)$-bit coefficients. Each symbol in a codeword is $O(\log n)$ bits long, and we get perfect reconstruction up to an error rate $\rho < \rho^*$. For $x \in \{1, \ldots, p\}^n$, we get an alphabet size of $O(p^2 n^{\frac{3}{2}})$. A similar result appears in [3]. While they use arithmetic operations over arbitrary-precision reals during the encoding process, a finer discretization of A would work in their setting as well.

7. INEFFICIENT ATTACKS

7.1 Unique Reconstruction

In this section, we show that using a linear number of ± 1 questions, an attacker can reconstruct almost the whole database if the curator is constrained to answer at least $\frac{1}{2} + \varepsilon$ of the questions within an absolute error of $o(\sqrt{n})$. We however give up on computational efficiency.

THEOREM 24. *For any $\varepsilon > 0$ and any function $\alpha = \alpha(n)$, there is constant c and an attack using cn ± 1 questions that reconstructs a database that agrees with the real database in all but at most $(\frac{4\alpha}{\varepsilon})^2$ entries, if the curator answers at least $\frac{1}{2} + \varepsilon$ of the questions within an absolute error of α.*

The proof will rely on the following lemma.

LEMMA 25. *Let $Y = \sum_{i=1}^{k} X_i$ where each X_i is a ± 1 random variable with mean zero. Then for any y and any l, $Pr[Y \in [y, y + l]] \leq \frac{l}{\sqrt{k}}$.*

PROOF. Note that $Pr[Y = y] = \binom{k}{(k+y)/2}(\frac{1}{2})^k$. This expression is at most $\binom{k}{\lceil k/2 \rceil}(\frac{1}{2})^k$. Using Stirling's approximation, this is bounded by $\sqrt{\frac{2}{\pi k}}$. The claim follows. □

PROOF. (of Theorem 24) We shall show that picking the questions at random works. More precisely, the attacker picks the $cn \times n$ matrix A with ± 1 entries uniformly at random, and then outputs any database x' such that $|y_i - (Ax')_i| \leq \alpha$ for at least $\frac{1}{2} + \varepsilon$ of the questions.

Let the true database be x and let x' be the reconstructed database. We wish to argue that x' agrees with x in all but $(\frac{4\alpha}{\varepsilon})^2$ entries. Assume the contrary, so that the vector $z = x' - x$ has at least $(\frac{4\alpha}{\varepsilon})^2$ entries that have absolute value 1. By assumption, $|(Az)_i| \leq 2\alpha$ for at least 2ε fraction of the questions, since any two sets of measure $(\frac{1}{2} + \varepsilon)$ must have an intersection of measure 2ε. We shall call such a z *bad*. We wish to argue that with high probability, A is such that no z is bad with respect to A.

Consider a z with at least $(\frac{4\alpha}{\varepsilon})^2$ ± 1 entries. For any i, $a_i z$ is the sum of at least $(\frac{4\alpha}{\varepsilon})^2$ ± 1 r.v.'s. Thus the above lemma implies that the probability that $a_i z$ lies in an interval of size 4α is at most ε. The expected number of questions for which $|a_i z| \leq 2\alpha$, is then at most εcn. Chernoff bounds therefore imply that the probability that this number exceeds 2ε is at most $exp(-\frac{\varepsilon cn}{2})$. Thus the probability of a particular z being bad is at most $exp(-\frac{\varepsilon cn}{2})$.

Taking a union bound over the at most 3^n possible z's, we get that with probability $exp(-n(\frac{\varepsilon c}{2} - \ln 3))$, no bad z exists. Taking $c \geq 2 \ln 3/\varepsilon$, this probability is exponentially small. □

7.2 List decoding

In this section, we argue that if the curator is constrained to answer at least a δ fraction of the answers approximately correctly, for any $\delta > 0$, the attacker can construct a list of at most $k = \lceil \frac{1}{\delta} \rceil$ candidate vectors $x_1, \ldots, x_k$ such that one of them agrees with the true database on all but $(\frac{8\alpha}{\delta^3})^2$ entries.

THEOREM 26. *Let $\delta > 0$ and $\alpha = \alpha(n)$ be arbitrary. Suppose that the curator is constrained to answer at least a δ fraction of the questions to within an absolute error or α. Then there is an attack consisting of cn ± 1 questions that constructs a list of $k = \lceil \frac{1}{\delta} \rceil$ candidate vectors $x_1, \ldots, x_k$ such that at least one of them agrees with the true database x on all but $(\frac{8\alpha}{\delta^3})^2$ entries.*

PROOF. We shall once again show that choosing random questions followed by a greedy decoding procedure works with high probability. More precisely, the attacker picks the $cn \times n$ matrix A with ± 1 entries uniformly at random, and then greedily finds vectors $x_1, \ldots$ so that each x_i agrees with the provided answers on at least a δ fraction of the answers, and differs from the previous x_j's on at least $(\frac{8\alpha}{\delta^3})^2$ entries.

We first establish the following claim. With high probability, the matrix A has the following property. If x and x' are two 0-1 vectors that differ in at least $(\frac{8\alpha}{\delta^3})^2$ entries, then at most a δ^3 fraction of the questions are such that $|a_i x - a_i x'| \leq 2\alpha$. Indeed, let $z = x - x'$ so that z has at least $(\frac{8\alpha}{\delta^3})^2$ ± 1's. By lemma 25, the probability that for a fixed z and i, $|a_i z| \leq 2\alpha$ is at most $\frac{\delta^3}{2}$. Thus by Chernoff bounds, the fraction of answers where $|a_i z|$ is small, is at most δ^3, with probability $1 - exp(-\delta^3 cn/2)$. Taking a union bound over at most 3^n possible z's, we get the desired result, provided that $c > 2 \ln 3/\delta^3$.

Assume that A indeed has the above property. First note that each x_s agrees with at least δ fraction of the answers. If we were to output $k+1 \geq 1+\frac{1}{\delta}$ x_s's, the total fraction of distinct answers covered by the x_s's will have to exceed $\delta(1+\frac{1}{\delta})-\delta^3\binom{k+1}{2}$, which is more than 1. Thus the algorithm outputs a list no longer than k.

Finally, note that if none of the outputs of the algorithm agree with x on all but $(\frac{8\alpha}{\delta^3})^2$ entries, then the algorithm would consider x as a candidate output. $\square$

We note that the δ^3 above can be improved to δ^2 if we allow the list size to be $2/\delta$.

8. CONCLUSIONS

A natural question is to determine the true value of the permissible error rate $\rho^*_{\pm 1} = \inf_z \rho^*(z)$. It is easy to see that for $z_n = (\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}})$, the distribution $\sum_j a_{ij} z_j$ approaches a Gaussian, and thus $\rho^*_{\pm 1}$ is bounded above by $\rho^* \approx 0.239$. However, the value of $\rho_{\pm 1}$ is strictly below 0.239. In fact, taking $z = (1.1, 1, 1, 1, 1, 1, 1)$, implies an upper bound of 0.226 on $\rho^*_{\pm 1}$!

While we compute the threshold in the limit when m/n is a large constant, the allowable error rate as a function of the ratio m/n is worth investigating. As mentioned earlier, lower bounds on this function are established in [12].

While we have proved high-probability bounds for random Gaussian matrices, the problem of explicitly constructing matrices A with allowable error threshold approaching, or even exceeding ρ^* remains open.

Comparing the results of section 7 with those of section 4, we note that while near-perfect decoding is information-theoretically possible for error rates up to a half, LP decoding fails at much lower error rates. It is thus natural to look for other efficient decoding procedures for the case of adversarial error.

Acknowledgements.

We would like to thank Sergey Yekhanin and Anupam Gupta for useful discussions.

9. REFERENCES

[1] D. Achlioptas. Database-friendly random projections: Johnson-Lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

[2] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the sulq framework. In C. Li, editor, *PODS*, pages 128–138. ACM, 2005.

[3] E. J. Candès, M. Rudelson, T. Tao, and R. Vershynin. Error correction via linear programming. In *FOCS*, pages 295–308. IEEE Computer Society, 2005.

[4] E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[5] E. J. Candès and T. Tao. Error correction via linear programming, 2005.

[6] E. J. Cands and P. Randall. Highly robust error correction by convex programming. Submitted, 2006.

[7] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci Comp*, 48(1):33–61, 1999.

[8] I. Dinur, C. Dwork, and K. Nissim. Privacy in public databases: A foundational approach, 2005. Manuscript.

[9] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *PODS*, pages 202–210. ACM, 2003.

[10] D. Donoho. For most large underdetermined systems of linear equations, the minimal 11-norm near-solution approximates the sparsest near-solution. *Communications on Pure and Applied Mathematics*, 59(7):907–934, 2006.

[11] D. Donoho. For most large underdetermined systems of linear equations, the minimal l1-norm solution is also the sparsest solution. *Communications on Pure and Applied Mathematics*, 59(6):797–829, 2006.

[12] D. Donoho. High-dimensional centrally symmetric polytopes with neighborliness proportional to dimension. *Discrete and Computational Geometry*, 35(4):617–652, 2006.

[13] D. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 48:2845–2862, 2001.

[14] D. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(3):879–921, 1998.

[15] D. Donoho and J. Tanner. Thresholds for the recovery of sparse solutions via l1 minimization. In *Proceedings of the Conference on Information Sciences and Systems*, 2006.

[16] D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

[17] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In S. Halevi and T. Rabin, editors, *TCC*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006.

[18] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In M. K. Franklin, editor, *CRYPTO*, volume 3152 of *Lecture Notes in Computer Science*, pages 528–544. Springer, 2004.

[19] J. Feldman. Decoding error-correcting codes via linear programming. Master's thesis, Massachusetts Institute of Technology, 2003.

[20] J. Feldman and D. Karger. Decoding turbo-like codes via linear programming. *Journal of Computer and System Sciences*, 68(4):733–752, 2004.

[21] J. Feldman, T. Malkin, C. Stein, R. Servedio, and M. Wainwright. LP decoding corrects a constant fraction of errors (an expanded version). In *ISIT*, pages 68–. IEEE, 2004.

[22] J. Feldman and C. Stein. LP decoding achieves capacity. In *SODA*, pages 460–469. SIAM, 2005.

[23] L. Goldstein. l^1 bounds in normal approximation. *Annals of Probability*, 2006. To Appear.

[24] M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2001.

[25] M. Talagrand. Concentration of measure and isoperimetric inequalities in product space. *Publ. Math. I.H.E.S.*, 81:73–205, 1995.