Basis pursuit and the Johnson–Lindenstrauss lemma

(Lecture Notes)

Jiří Matoušek
notes prepared with the help of Marek Krčál

May 2007

Abstract

We give a self-contained presentation of a recent interesting result (discovered independently by Donoho, by Candès and Tao, and by Rudelson and Vershynin), stating that if A is a random matrix of a suitable size, then sparse solutions to the system of linear equations $A\mathbf{x} = \mathbf{b}$ can be computed efficiently via linear programming. A part of the text is similar to a section in the book *J. Matoušek, B. Gärtner: Understanding and Using Linear programming*, but there the result is just explained without proof. Here we describe a proof by Baraniuk et al. based on the Johnson–Lindenstrauss lemma (time permitting, I want to include a full proof of the lemma as well in the future).

1 Preliminaries

Here we review tools needed in the sequel.

1.1 Dense sets in a sphere

Let $S^{n-1} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = 1\}$ denote the unit sphere in $\mathbb{R}^n$ (note that S^2 is the 2-dimensional sphere living in $\mathbb{R}^3$). We are given a number $\delta > 0$, and we want to place a reasonably small finite set N of points on S^{n-1} in such a way that each $\mathbf{x} \in S^{n-1}$ has some point of N at distance no larger than δ. Such an N is called ***δ*-dense** in S^{n-1}.

It is generally difficult to find good explicit constructions for arbitrary δ and n. The following simple but clever existential argument yields a δ-dense set whose size has essentially the best possible order of magnitude (for n large).

Lemma 1.1 (Small δ-dense sets in the sphere) *For each $\delta \in (0, 1]$, there exists a δ-dense set $N \subseteq S^{n-1}$ that satisfies*

$$|N| \leq \left(\frac{4}{\delta}\right)^n .$$

Proof. In order to construct a small δ-dense set, we start with the empty set and keep adding points one by one. The trick is that we do not worry about

δ-density along the way, but we always keep the current set *δ-separated*, which means that every two points have distance at least δ. Clearly, if no more points can be added, the resulting set N must be δ-dense.

For each $\mathbf{x} \in N$, consider the ball of radius $\frac{\delta}{2}$ centered at $\mathbf{x}$. Since N is δ-separated, these balls have disjoint interiors, and they are contained in the ball $B(0, 1 + \delta/2) \subseteq B(0, 2)$. Therefore, $\mathrm{vol}(B(0,2)) \geq |N| \mathrm{vol}(B(0, \frac{\delta}{2}))$, and since $\mathrm{vol}(B(0,r))$ in $\mathbb{R}^n$ is proportional to r^n, the lemma follows. $\square$

1.2 Almost isometry and dense sets

Definition 1.2 *Let $F: \mathbb{R}^n \to \mathbb{R}^m$ be a mapping, and let $\varepsilon \in (0,1)$ be a real number. We call F a* **(Euclidean) ε-almost isometry**[1] *if for every $\mathbf{x} \in \mathbb{R}^n$ we have*

$$(1-\varepsilon)\|\mathbf{x}\|_2 \leq \|F(\mathbf{x})\|_2 \leq (1+\varepsilon)\|\mathbf{x}\|_2.$$

We will need a result telling us that if a *linear* map behaves like an almost-isometry on a sufficiently dense set in the sphere, then it is already an almost-isometry on all of $\mathbb{R}^n$, although with a worse ε.

Lemma 1.3 *Let $\varepsilon \in (0, \frac{1}{3})$, let $N \subset S^{n-1}$ be ε-dense, and let $F: \mathbb{R}^n \to \mathbb{R}^m$ be a linear map satisfying $1 - \varepsilon \leq \|F(q)\|_2 \leq 1 + \varepsilon$ for all $\mathbf{q} \in N$. Then F is a 3ε-almost isometry.*

Proof. First we note that since F is a linear map, it suffices to prove the almost-isometry property for all $\mathbf{x} \in S^{n-1}$; that is, $1 - 3\varepsilon \leq \|F(\mathbf{x})\|_2 \leq 1 + 3\varepsilon$ for all $\mathbf{x} \in S^{n-1}$.

We begin with the upper bound—this is where the trick lies. Let $M = \max\{\|F(\mathbf{x})\|_2 : \mathbf{x} \in S^{n-1}\}$ and let $\mathbf{x}_0 \in S^{n-1}$ be a point where M is attained. Let $\mathbf{q}_0$ be a point of N with $\|\mathbf{x}_0 - \mathbf{q}_0\|_2 \leq \varepsilon$. Then by the linearity of F and by triangle inequality $M = \|F(\mathbf{x}_0)\|_2 \leq \|F(\mathbf{q}_0)\|_2 + \|F(\mathbf{x}_0 - \mathbf{q}_0)\|_2 \leq 1 + \varepsilon + M\|\mathbf{x}_0 - \mathbf{q}_0\|_2 \leq 1 + \varepsilon + M\varepsilon$, and thus $M \leq (1+\varepsilon)/(1-\varepsilon) \leq 1 + 3\varepsilon$ (the last inequality is valid for all $\varepsilon \in (0, \frac{1}{3})$). Hence $\|F(\mathbf{x})\|_2 \leq 1 + 3\varepsilon$ for all $\mathbf{x} \in S^{n-1}$.

It remains to bound $\|F(\mathbf{x})\|_2$ from below, which is routine. We choose $\mathbf{q} \in N$ with $\|\mathbf{x} - \mathbf{q}\|_2 \leq \varepsilon$ and we calculate $\|F(\mathbf{x})\|_2 \geq \|F(\mathbf{q})\|_2 - \|F(\mathbf{x} - \mathbf{q})\|_2 \geq 1 - \varepsilon - (1 + 3\varepsilon)\varepsilon = 1 - 2\varepsilon - 3\varepsilon^2 \geq 1 - 3\varepsilon$. $\square$

1.3 A result from the proof of the Johnson–Lindenstrauss lemma

Let A be a random matrix with m rows and t columns, where each entry a_{ij} has the standard normal distribution $N(0,1)$ and all entries are mutually independent. Let $B = \frac{1}{\sqrt{m}} A$, and let us regard B as a linear map $B: \mathbb{R}^t \to \mathbb{R}^m$.

[1] This notion is closely related to distortion. An ε-almost isometry has distortion at most $\frac{1+\varepsilon}{1-\varepsilon}$, and on the other hand, any mapping F between Euclidean spaces with distortion at most $\frac{1+\varepsilon}{1-\varepsilon}$ can be re-scaled so that the resulting map is an ε-almost isometry.

Proposition 1.4 *Let $\varepsilon \in [0, \frac{1}{2})$. If B is the random linear map as above and $\mathbf{x} \in S^{t-1}$ is an arbitrary unit vector in $\mathbb{R}^t$, then*

$$\Pr\Big[1 - \varepsilon \le \|B\mathbf{x}\|_2 \le 1 + \varepsilon\Big] \ge 1 - \mathrm{e}^{-c\varepsilon^2 m},$$

where c is a positive constant.

Let us note that this proposition does *not* claim that B is an ε-isometry with high probability; indeed, for $t > m$ it cannot be, since it is never injective.

Similar results can be established for many other models of random matrices, for instance, matrices with independent random ± 1 entries (suitably scaled). The main theorem (Theorem 3.2) discussed below then transfers accordingly.

2 Sparse solutions of linear systems

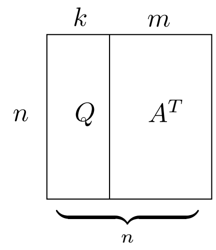
2.1 Motivation I: A coding problem

A cosmic probe wants to send the results of its measurements, represented by a vector $\mathbf{w} \in \mathbb{R}^k$, back to Earth. A fraction of coordinates can be corrupted during the transmission. We admit *gross errors*; that is, if the number 3.1415 is sent and it gets corrupted, it can be received as 2152.66, or 3.1425, or -10^{11}, or any other real number.

This problem belongs to the theory of *error-correcting codes*. Most of the results of this theory deal with encoding messages over finite alphabets. Here we will discuss a solution that deals directly with real numbers.

We choose a suitable integer $n > k$ and a suitable $n \times k$ *encoding matrix* Q of rank k, and instead of $\mathbf{w}$ we send the vector $\mathbf{z} = Q\mathbf{w} \in \mathbb{R}^n$. Because of the errors, the received vector is not $\mathbf{z}$ but $\tilde{\mathbf{z}} = \mathbf{z} + \mathbf{x}$, where $\mathbf{x} \in \mathbb{R}^n$ is a vector with a small number of nonzero coordinates—we assume $|\mathrm{supp}(\mathbf{x})| \le r$, where r is the allowed number of errors and where we denote $\mathrm{supp}(\mathbf{x}) = \{i : x_i \ne 0\}$. We ask, under what conditions can $\mathbf{w}$ be recovered from $\tilde{\mathbf{z}}$?

Somewhat counterintuitively, we concentrate on the task of finding the "error vector" $\mathbf{x}$ (then $\mathbf{w}$ can be computed by solving a system of linear equations). Let $m = n - k$ and let A be an $m \times n$ matrix such that $AQ = (0)_{m \times k}$. That is, considering the k-dimensional linear subspace of $\mathbb{R}^n$ generated by the columns of Q, the rows of A all lie in its orthogonal complement. The following picture illustrates the dimensions of the matrices:



$AQ = 0$

We have

$$A\tilde{\mathbf{z}} = A(Q\mathbf{w} + \mathbf{x}) = 0\mathbf{w} + A\mathbf{x},$$

and thus $\mathbf{x}$ is a solution of the system of linear equations $A\mathbf{x} = \mathbf{b}$ with $\mathbf{b} = A\tilde{\mathbf{z}}$. The system $A\mathbf{x} = \mathbf{b}$ has infinitely many solutions in general, but we are searching for one with a small support. As we will see, under suitable conditions relating n, m, r and A, such a *sparse* solution of $A\mathbf{x} = \mathbf{b}$ turns out to be unique (and thus it has to be the desired error vector), and it can be computed efficiently by linear programming!

This brings us to the following computational problem:

Sparse solution of a system of linear equations

Input: $A \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $r \in \mathbb{N}$.
Task: Find an $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{b}$ and $|\mathrm{supp}(\mathbf{x})| \leq r$.

2.2 Motivation II: Removing noise from a signal

We have a screenshot from a TV with bad signal and the image is noisy. We can represent the image in the form of a vector of reals $\mathbf{f} + \mathbf{g}$, where $\mathbf{f}$ stands for the original image and $\mathbf{g}$ stands for the noise.

The task of extracting the original image is not quite well-defined; we must make some assumption on how the image differs from the noise. One way is to assume that

$$\mathbf{f} = \sum_{i\in I} \beta_i \phi_i \quad \text{and} \quad \mathbf{g} = \sum_{j\in J} \alpha_j \sigma_j,$$

where the ϕ_i are from some convenient fixed set of image generators, e.g., suitable smooth wavelets, while the σ_j are from some convenient fixed set of noise generators, say some kind of "spike" functions. We want to find the coefficients β_i and α_j. The trouble is that the set $\{\phi_i : i \in I\} \cup \{\sigma_j : j \in J\}$ need not be linear independent. A hope for solution: find the β_i and α_j such that most of them are zeros. Thus we again arrive at the quest for sparse solution.

2.3 A linear algebra view

Given a matrix A and a natural number r, we consider the question: When does the system $A\mathbf{x} = \mathbf{b}$ have at most one solution $\mathbf{x}$ with $|\mathrm{supp}(\mathbf{x})| \leq r$ for all $\mathbf{b}$?

Observation 2.1 *The answer is positive if and only if every at most $2r$ columns of A are linearly independent.*

Proof. We will prove only one direction: If every $2r$ columns are linearly independent, then there is at most one sparse solution. Suppose that $\mathbf{x}$ and $\mathbf{x}'$ are two different sparse solutions. Then $\mathbf{y} = \mathbf{x} - \mathbf{x}'$ has at most $2r$ nonzero components and satisfies $A\mathbf{y} = A\mathbf{x}' - A\mathbf{x}'' = \mathbf{0}$, and hence it defines a linear dependence of at most $2r$ columns of A. □

Let us note that Observation 2.1 gives, in particular, $m \geq 2r$. On the other hand, if we choose a "random" $2r\times n$ matrix A, we almost surely have every $2r$ columns linearly independent. (We don't want to assume or introduce the knowledge required to state and prove this claim rigorously.) So in the coding problem, if we set n so that $n = k + 2r$, choose A randomly, and let the columns of Q form a basis of the orthogonal complement of the row space of A, we seem to be done—a random A has almost surely every $2r$ columns linearly independent, and in such case, assuming that no more than r errors occurred, the sparse error vector $\mathbf{x}$ is always determined uniquely, and so is the original message $\mathbf{w}$.

Efficiency? But a major question remains—how can we *find* the unknown sparse solution $\mathbf{x}$? Unfortunately, it turns out that the problem of computing a sparse solution of $A\mathbf{x} = \mathbf{b}$ is difficult (NP-hard) in general, even for A satisfying the conditions of Observation 2.1.

Since the problem of finding a sparse solution of $A\mathbf{x} = \mathbf{b}$ is important and computationally difficult, several heuristic methods have been proposed for solving it at least approximately and at least in some cases. One of them, described next, turned out to be considerably more powerful than the others.

3 Basis pursuit

A sparse solution $\mathbf{x}$ is "small" in the sense of having few nonzero components. The idea is to look for $\mathbf{x}$ that is "small" in another sense that is easier to deal with, namely, with small $|x_1| + |x_2| + \cdots + |x_n|$. The last quantity is commonly denoted by $\|\mathbf{x}\|_1$ and called the ***L_1*-norm** of $\mathbf{x}$ (while $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$ is the usual Euclidean norm, which can also be called the L_2-norm). We thus arrive at the following optimization problem (usually called *basis pursuit* in the literature):

$$\text{Minimize } \|\mathbf{x}\|_1 \text{ subject to } \mathbf{x} \in \mathbb{R}^n \text{ and } A\mathbf{x} = \mathbf{b}. \tag{BP}$$

This problem can be reformulated as a linear program:

$$\begin{array}{ll} \text{Minimize} & u_1 + u_2 + \cdots + u_n \\ \text{subject to} & A\mathbf{x} = \mathbf{b} \\ & -\mathbf{u} \leq \mathbf{x} \leq \mathbf{u} \\ & \mathbf{x}, \mathbf{u} \in \mathbb{R}^n, \ \mathbf{u} \geq \mathbf{0}. \end{array} \tag{BP$'$}$$

To check the equivalence of (BP) and (BP$'$), we just note that in an optimal solution of (BP$'$) we have $u_i = |x_i|$ for every i.

The basis pursuit approach to finding a sparse solution of $A\mathbf{x} = \mathbf{b}$ thus consists in computing an optimal solution $\mathbf{x}^*$ of (BP) by linear programming, and hoping that, with some luck, this $\mathbf{x}^*$ might also be the sparse solution or at least close to it.

At first sight it is not clear why basis pursuit should have any chance of finding a sparse solution. After all, the desired sparse solution might have a few huge components, while $\mathbf{x}^*$, a minimizer of the ℓ_1-norm, might have all components nonzero but tiny.

Surprisingly, experiments have revealed that basis pursuit actually performs excellently, and it usually finds the sparse solution exactly even in conditions that don't look very favorable. (In contrast to this, while minimizing the Euclidean norm of $\mathbf{x}$ instead of the L_1 norm is also computationally feasible, it almost never finds a sparse solution.)

Let us try to formalize this situation.

Definition 3.1 *A matrix A is* **BP-exact for sparsity r** *if for all $\mathbf{b} \in \mathbb{R}^m$ such that $A\mathbf{x} = \mathbf{b}$ has a unique sparse solution $\tilde{\mathbf{x}}$ with $|\mathrm{supp}(\tilde{\mathbf{x}})| \le r$, the problem (BP) has $\tilde{\mathbf{x}}$ as the unique minimum.*

Here is the main result we want to prove in this text:

Theorem 3.2 (Donoho; Candès and Tao; Rudelson and Vershynin) *There are real constants C and $c_1 > 0$ such that if n, m, r are integers with $1 \le r \le n/C$ and $m \ge Cr \log \frac{n}{r}$ and if $A \in \mathbb{R}^{m \times n}$ is a random matrix with entries drawn independently from the standard normal distribution $N(0,1)$, then*

$$\Pr\Big[A \text{ is BP-exact for sparsity } r\Big] \ge 1 - \mathrm{e}^{-c_1 m}.$$

Remark. The theorem is asymptotically optimal in the following sense: For $m = o(r \log \frac{n}{r})$, *no* $m \times n$ matrix at all can be BP-exact for sparsity r. This follows from a result of Linial and Novik, concerning certain convex polytopes, by a reduction found by Donoho.

4 Restricted almost-isometry

Here, following Candès and Tao (with some simplification), we connect the property of BP-exactness of a matrix A to another property of A:

Definition 4.1 *A matrix A has the property of* **t-restricted ε-almost isometry** *if the corresponding linear mapping satisfies the condition of ε-almost isometry for every sparse $\mathbf{x}$; that is, if*

$$(1-\varepsilon)\|\mathbf{x}\|_2 \le \|A\mathbf{x}\|_2 \le (1+\varepsilon)\|\mathbf{x}\|_2$$

for all $\mathbf{x} \in \mathbb{R}^n$ with $|\mathrm{supp}(\mathbf{x})| \le t$.

Lemma 4.2 *There is a constant $\varepsilon_0 > 0$ such that if a matrix A has the property of $3r$-restricted ε_0-almost isometry, then it is BP-exact for sparsity r.*

Let us recall that Observation 2.1 shows that if every $2r$ columns of A are linearly independent, then $A\mathbf{x} = \mathbf{b}$ has at most one sparse solution for every $\mathbf{b}$. The condition of $3r$-restricted ε_0-almost isometry in the lemma can be viewed as a strengthening of the assumption of Observation 2.1: Instead of every $2r$ columns, we need to deal with every $3r$ columns, and more significantly, instead of wanting the columns merely linearly independent, we want them almost orthogonal (this is an alternative view of restricted almost-isometry).

Proof of Lemma 4.2. Let us suppose that A has the property of $3r$-restricted ε_0-almost isometry, and that $\tilde{\mathbf{x}}$ is a solution of $A\mathbf{x} = \mathbf{b}$ for some $\mathbf{b}$ with $|\mathrm{supp}(\tilde{\mathbf{x}})| \le r$.

For contradiction, we assume that $\tilde{\mathbf{x}}$ is not the unique minimum of (BP), and so there is another solution of $A\mathbf{x} = \mathbf{b}$ with smaller or equal L_1-norm. We write this solution in the form $\tilde{\mathbf{x}} + \mathbf{\Delta}$; so

$$A\mathbf{\Delta} = \mathbf{0}, \quad \|\tilde{\mathbf{x}} + \mathbf{\Delta}\|_1 \leq \|\tilde{\mathbf{x}}\|_1.$$

We want to reach a contradiction assuming $\mathbf{\Delta} \neq \mathbf{0}$.

Let us note that *if* A were an almost-isometry, then $\mathbf{\Delta} \neq \mathbf{0}$ would imply $A\mathbf{\Delta} \neq \mathbf{0}$ and we would have a contradiction immediately. Of course, we cannot expect the whole A to be an almost-isometry—we have control only over small blocks of A.

First we set $S := \mathrm{supp}(\tilde{\mathbf{x}})$ and we show that a substantial part of $\mathbf{\Delta}$, in terms of the L_1-norm, has to live on S. This is where we use the condition $\|\tilde{\mathbf{x}} + \mathbf{\Delta}\|_1 \leq \|\tilde{\mathbf{x}}\|_1$.

Claim. *We have*

$$\|\mathbf{\Delta}_S\|_1 \geq \|\mathbf{\Delta}_{\overline{S}}\|_1,$$

where $\mathbf{\Delta}_S$ denotes the vector consisting of the components of $\mathbf{\Delta}$ indexed by S, and $\overline{S} = \{1, 2, \ldots, n\} \setminus S$.

Proof.

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_1 &\geq \|\tilde{\mathbf{x}} + \mathbf{\Delta}\|_1 \\ &= \|(\tilde{\mathbf{x}} + \mathbf{\Delta})_S\|_1 + \|(\tilde{\mathbf{x}} + \mathbf{\Delta})_{\overline{S}}\|_1 \\ &= \|\tilde{\mathbf{x}}_S + \mathbf{\Delta}_S\|_1 + \|\mathbf{\Delta}_{\overline{S}}\|_1 \\ &\geq \|\tilde{\mathbf{x}}\|_1 - \|\mathbf{\Delta}_S\|_1 + \|\mathbf{\Delta}_{\overline{S}}\|_1. \end{aligned}$$

Comparing the first and last terms in this chain yields the inequality in the claim.

Next, we want to show, roughly speaking, that S also accounts for most of the Euclidean norm of $\mathbf{\Delta}$. To this end, we partition the index set $\overline{S}$ into blocks $B_1, B_2, \ldots$ of size $2r$ each (except for the last block which may have fewer elements). Namely, B_1 are the indices of the $2r$ largest coordinates of $\mathbf{\Delta}_{\overline{S}}$ in absolute value, B_2 are the indices of the next $2r$ largest coordinates, etc.

This choice of the blocks implies that for every $i \in B_{j+1}$ we have

$$|\Delta_i| \leq \frac{\|\mathbf{\Delta}_{B_j}\|_1}{2r},$$

and consequently,

$$\|\mathbf{\Delta}_{B_{j+1}}\|_2 \leq \frac{\|\mathbf{\Delta}_{B_j}\|_1}{\sqrt{2r}}$$

(notice that this inequality contains both the Euclidean norm and the L_1-norm). Then we can bound

$$\begin{aligned} \sum_{j \geq 1} \|\mathbf{\Delta}_{B_{j+1}}\|_2 &\leq \sum_{j \geq 1} \frac{\|\mathbf{\Delta}_{B_j}\|_1}{\sqrt{2r}} = \frac{1}{\sqrt{2r}} \|\mathbf{\Delta}_{\overline{S}}\|_1 \\ &\leq \frac{1}{\sqrt{2r}} \|\mathbf{\Delta}_S\|_1 \leq \frac{1}{\sqrt{2}} \|\mathbf{\Delta}_S\|_2, \end{aligned}$$

where the last inequality uses that $\|\mathbf{v}\|_1 \leq \sqrt{s}\|\mathbf{v}\|_2$ for every s-element vector $\mathbf{v}$, and the previous one is the claim above.

Altogether we have proved

$$\sum_{j\geq 1} \|\mathbf{\Delta}_{B_{j+1}}\|_2 \leq \frac{1}{\sqrt{2}}\|\mathbf{\Delta}_S\|_2. \tag{1}$$

Now we are ready for the final calculation leading to a contradiction, in which we use the restricted almost-isometry property of A for $T = S \cup B_1$ and for $T = B_2, B_3, \ldots$:

$$\begin{array}{rcll}
0 &=& \|A\mathbf{\Delta}\|_2 & \\
&\geq& \|A_{S\cup B_1}\mathbf{\Delta}_{S\cup B_1}\|_2 - \sum_{j\geq 2}\|A_{B_j}\mathbf{\Delta}_{B_j}\|_2 & \text{(triangle inequality)} \\
&\geq& (1-\varepsilon_0)\|\mathbf{\Delta}_{S\cup B_1}\|_2 - (1+\varepsilon_0)\sum_{j\geq 2}\|\mathbf{\Delta}_{B_j}\|_2 & \text{(almost-isometry)} \\
&\geq& (1-\varepsilon_0)\|\mathbf{\Delta}_S\|_2 - \frac{1+\varepsilon_0}{\sqrt{2}}\|\mathbf{\Delta}_S\|_2 & \text{(by (1))} \\
&=& \|\mathbf{\Delta}_S\|_2\left(1-\varepsilon_0-\frac{1+\varepsilon_0}{\sqrt{2}}\right). &
\end{array}$$

For ε_0 small, the part of the last expression in parentheses is positive, and since the whole expression is nonpositive, we get $\|\mathbf{\Delta}_S\|_2 = 0$. This is the desired contradiction establishing the lemma. $\square$

5 Proof of the main theorem

This part of the proof follows an idea of Baraniuk, Davenport, DeVore, and Wakin; the original proofs were different.

In view of Lemma 4.2, it suffices to prove that if A is a random matrix as in the main theorem and $B := \frac{1}{\sqrt{m}}A$ is the appropriate re-scaling (re-scaling obviously doesn't affect BP-exactness), then B has the property of $3r$-restricted ε_0-almost isometry with probability at least $1 - e^{-c_1 m}$ for a suitable positive constant c_1.

Let us write $t := 3r$ and suppose, as we may, that $t \leq n$. If B doesn't have the property of $3r$-restricted ε_0-almost isometry, then there exists a t-element set $T \subseteq \{1, 2, \ldots, n\}$ such that B_T is not an ε_0-almost isometry in the sense of Definition 1.2, where B_T denotes the matrix consisting of the columns of B indexed by T, as well as the corresponding linear map $\mathbb{R}^t \to \mathbb{R}^m$.

Let us fix an $(\varepsilon_0/3)$-dense set N_T in S^{t-1}. By Lemma 1.3, if B_T is not an ε_0-almost isometry, then there exists $\mathbf{q} \in N_T$ such that

$$1 - \varepsilon_0/3 \leq \|B_T\mathbf{q}\|_2 \leq 1 + \varepsilon_0/3 \tag{2}$$

does not hold. Since B_T is a random $t \times m$ matrix, Proposition 1.4 tells us that for any fixed $\mathbf{q} \in S^{t-1}$, the condition (2) fails with probability at most $e^{-c_2 m}$, where $c_2 > 0$ is a constant depending on ε_0.

By Lemma 1.1, we may assume $|N_T| \le K^t$ for a suitable constant K (again depending on ε_0), and hence the probability of (2) failing for *any* $\mathbf{q} \in N_T$ is at most $K^t e^{-c_2 m}$. Finally, there are $\binom{n}{t}$ possible choices of T, and so we calculate

$$\Pr\Big[\text{B does not have the } t\text{-restricted } \varepsilon_0\text{-almost isometry property}\Big] \le$$

$$\binom{n}{t} K^t \mathrm{e}^{-c_2 m} \le \left(\frac{n\mathrm{e}}{t}\right)^t K^t \mathrm{e}^{-c_2 m} = \exp\Big(3r(\ln \frac{n\mathrm{e}}{3r} + \ln K) - c_2 m\Big).$$

Using the assumption $m \ge Cr \log(n/r)$ from the theorem, it is easy to check that the last expression is bounded above by $e^{-c_2 m/2}$, say, provided that C is sufficiently large. This concludes the proof of the main theorem.