

Compressive Sensing

March 3, 2008

Chapter 1

Compressive Sampling

1.1 Introduction

Although it is convenient for conceptual and theoretical purposes to think of signals as general functions of time. In practice they are usually acquired, processed, stored and transmitted as discrete and finite time samples. We need to study this sampling process carefully to determine to what extent a sampling or discretization allows us to reconstruct the original information in the signal. Furthermore, real signals such as speech or images are not arbitrary functions. Depending on the type of signal, they have special structure. No one would confuse the output of a random number generator with human speech. It is also important to understand the extent to which we can compress the basic information in the signal to minimize storage space and maximize transmission time.

Shannon sampling is one approach to these issues. In this approach we model real signals as functions $f(t)$ in $L_2(-\infty, \infty)$ that are bandlimited. Thus if the frequency support of $\hat{f}(\omega)$ is contained in the interval $[-\Omega, \Omega]$ and we sample the signal at discrete time intervals with equal spacing less than $1/2\pi\Omega$, i.e., faster than the Nyquist rate, we can reconstruct the original signal exactly from the discrete samples. This method will work provided hardware exists to sample the signal at the required rate. Increasingly this is a problem because modern technologies can generate signals of higher bandwidth than existing hardware can sample.

There are other models for signals that exploit different properties of real signals and can be used as alternatives to Shannon sampling. In this chapter we introduce an alternative model that is based on the sparsity of many real signals. Intuitively we think of a signal as sparse if its expression with respect to some

chosen basis has coefficients that are mostly zero (or very small). The content of the signal is in the location and values of the spikes, the nonzero terms. For example the return from a radar signal at an airport is typically null, except for a few spikes locating the positions and velocities of nearby aircraft. A time trace of the sound from a musical instrument might not be sparse, whereas the Fourier transform of the the same signal would be sparse. This approach to the modeling, processing and storage of real signals is called compressive sampling.

We start with a simple model of a signal as a real n -tuple $x \in R_n$. We think of n as large. We are especially interested in the case were x is k -sparse. That is, at most k of the components x_1, \dots, x_n of x are nonzero. We think of k as small with respect to n . In order to obtain information about the signal x we sample it. Here, a sample is a linear functional f_c on R_n . That is, the sample $f_c(x) = c \cdot x$ where $c = (c_1, \dots, c_n)$ is a given sample vector and $c \cdot x$ is the dot product of c and x . Once c is chosen, the dot product is easy to implement in hardware. In the case where $c_i = \delta_{ij}$ for fixed j and $i = 1, \dots, n$ the sample would yield the value of the component x_j . Now suppose we take m different samples y_ℓ of x , i.e., we have m distinct sample vectors $c^{(\ell)}$, $\ell = 1, \dots, m$. We can describe the sampling process by the equation $y = \Phi x$ where the $m \times n$ sampling matrix Φ is defined by

$$(\Phi_{\ell,j}) = \begin{pmatrix} c_j^{(1)} \\ \vdots \\ c_j^{(m)} \end{pmatrix}, \quad (1.1)$$

and y is an m -tuple. For compressive sampling m is less than n , so that the system is underdetermined. The problem is to design the sampling matrix ϕ so that we can determine the signal x uniquely from m samples. Obviously, this is impossible for arbitrary signals x . It is possible if we know in advance that x has some special form. For compressive sampling at its simplest, the only requirement on the signal x is that it is k -sparse, i.e., that at most $k < n$ of the components x_j are nonzero. The only assumption is k -sparsity, not the possible locations of the nonzero components.

A simple example will illustrate the utility of compressive sampling. Suppose we have a sequence of long n -component signals $x^{(h)}$ that are 1-sparse. Thus, each signal consists of a single spike at some location $j^{(h)}$ with all other components 0. To determine the signal uniquely we need only find the location and the value of the spike. How can we do this from a minimum number of samples? This problem has an elegant solution, particularly simple in the case $n = 2^\ell$ for ℓ a positive integer. For each 1-sparse signal x we number the components x_j from

$j = 0$ to $j = 2^\ell - 1$. Similarly we will number the rows and columns of Φ , starting from 0. Recall that each j can be written uniquely in binary numbers as $[j_{\ell-1}, j_{\ell-2}, \dots, j_1, j_0]$ where $j = j_0 2^0 + j_1 2^1 + \dots + j_{\ell-1} 2^{\ell-1}$ where each $j_s = 0, 1$. We design the $m \times n$ sampling matrix Φ where $m = \ell + 1$ by first filling the top (0th) row with ones: $\Phi_{0,j} = 1$. For the remaining m terms in the 0th column of Φ we enter the binary number for 0, for the remaining m terms in the 1st column we enter the binary number for 1, and so forth. Thus the sampling matrix is given by $\Phi_{ij} = j_{\ell-i}$ for $i = 1, \dots, \ell$. Computing $y = \Phi x$ we have $y_0 = x_{j(h)}$ the magnitude of the spike, whereas if $y_0 \neq 0$ $[y_m/y_0, \dots, y_1/y_0]$ is the location of the spike, written in binary. For example, suppose $\ell = 3$, $m = 4$, $n = 8$ and the signal x has the spike $x_6 = 6.1$. Then $y = \Phi x$ becomes

$$\begin{pmatrix} 6.1 \\ 0 \\ 6.1 \\ 6.1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 6.1 \\ 0 \end{pmatrix},$$

and from y we see that the value of the spike is $y_0 = 6.1$ and the location is $[1, 1, 0] = 6$. We see from this construction that for 1-sparse signals of large length n we can identify the signal uniquely with only about $\log_2 n$ samples; it isn't necessary to sample all n components individually. For storage or transmission of the information, the compression is impressive. Rather than store n numbers we need only store $m \approx \log_2 n$ numbers.

In this chapter we will study the problem, for given k, n with $k < n$, of how to design $m \times n$ sampling matrices (or encoders) to uniquely determine k -sparse signals x such that the number of samples m is as small as possible. Later we shall treat the more important case where noise and measuring errors are allowed and we want to find a k -sparse approximation \tilde{x} of x with approximation error and m as small as possible.

1.2 Algebraic theory

The sampling problem is closely associated with the structure of the null space $N(\Phi)$ of the $m \times n$ sampling/encoding matrix. Clearly, the rank of Φ is $\leq m$,

and $m < n$ since our system is underdetermined. There will be many signals x that will yield the same sample y . Recall that $N(\Phi) = \{z \in R_n : \Phi z = \theta\}$ and $\dim N(\Phi) \geq n - m$. The hyperplane $F(y)$ of signals giving the sample y takes the form

$$F(y) = \{x \in R_n : \Phi x = y\} = x_0 + N(\Phi)$$

for any $x_0 \in F(y)$. Note that every signal x lies on one and only one of these hyperplanes. Once we have sampled or encoded the signal we will need to try to reconstruct the signal from the sample. We define a decoder as a mapping $\Delta : R_m \rightarrow R_n$, which could be nonlinear. Given a signal x and a decoder Δ we think of $\bar{x} = \Delta(\Phi(x)) = \Delta(y)$ as our approximation to x , based on the sample. Ideally we would like an encoding-decoding pair Φ, Δ such that $\bar{x} = x$. However, this is not possible for an underdetermined system, unless we restrict the set of signals. Clearly, ensuring that at most one allowable signal x gives a particular sample y (so that we can recover x from y via a suitable decoder) is equivalent to requiring that each hyperplane $F(y)$ contains at most one allowable signal. Initially we restrict to signals that are sparse with respect to the standard basis e_j for R_n , without any other requirements. More precisely, given a positive integer $k < n$ we will restrict x to the subset Σ_k of k -sparse signals. Here,

$$\Sigma_k = \{x \in R_n : \#\text{supp}(x) \leq k\}, \text{ where } \text{supp}(x) = \{i : x_i \neq 0\} \quad (1.2)$$

and $\#\text{supp}$ is the cardinality of the support of x . In other words, Σ_k is the subset of R_n consisting of all signals with at most k nonzero components. Note that Σ_k is not a subspace of R_n because the sum of two k -sparse signals may not be k -sparse. Indeed, if $x^{(1)}, x^{(2)} \in \Sigma_k$, the best that we can guarantee is that $x^{(1)} + x^{(2)} \in \Sigma_{2k}$.

Exercise 1 Show that even though Σ_k is not a vector space, it can be expressed as the set-theoretical union of subspaces $X_{T_k} = \{x : \text{supp}(x) \subset T_k\}$ where T_k runs over all k -element subsets of the integers $\{1, 2, \dots, n\}$, i.e.,

$$\Sigma_k = \cup_{T_k} X_{T_k}.$$

If we choose Σ_k for some fixed $k < n$ as our set of allowable signals, in order to recover each k -sparse signal x from its sample y , we need to find the possible $m \times n$ sample matrices Φ such that $y = \Phi x$ uniquely determines x for all $x \in \Sigma_k$. To characterize the sampling matrices we first recall the representation (1.1) of Φ in terms of its column vectors and introduce notation to describe submatrices of Φ . Let T be a set of $\#(T) = \ell$ column indices $1 \leq i_1 < i_2 < \dots < i_\ell \leq n$. Then

$\Phi_T = (c_{i_1}, \dots, c_{i_\ell})$ is the $\#(T) \times n$ submatrix of Φ formed from the columns indexed by T . Then, $\Phi_T^{\text{tr}} \Phi_T$ is the $\#(T) \times \#(T)$ matrix with (s, t) matrix element given by the dot product $c_{i_s} \cdot c_{i_t}$, $1 \leq s, t \leq \#(T)$.

Theorem 1 *Let Φ be a $m \times n$ matrix and k a positive integer $\leq n$. The following are equivalent*

- (1) *For all $y \in R_m$, each $F(y)$ contains at most one element of Σ_k .*
- (2) $\Sigma_{2k} \cap N(\Phi) = \{\theta\}$.
- (3) *For any set of column indices T with $\#(T) = 2k$, the matrix Φ_T has rank $2k$.*
- (4) *For any set of column indices T with $\#(T) = 2k$, the $2k \times 2k$ matrix $\Phi_T^{\text{tr}} \Phi_T$ has all eigenvalues $\lambda_j > 0$ and is invertible.*

PROOF:

- (1) \rightarrow (2). Suppose $x \in \Sigma_{2k}$ and $\Phi x = \theta$. Since x has at most $2k$ nonzero components, we can always find vectors $x^{(1)}, x^{(2)}$ each with at most k nonzero components, and such that $x = x^{(1)} - x^{(2)}$. Now set $y = \Phi x^{(1)}$. Since $\Phi x = \Phi x^{(1)} - \Phi x^{(2)} = \theta$ we have that both $x^{(1)}, x^{(2)} \in F(y)$. By (1), $x^{(1)} = x^{(2)}$, so $x = \theta$.
- (2) \rightarrow (3). Let T be a set of indices $1 \leq i_1 < i_2 < \dots < i_{2k} \leq n$ with $\#(T) = 2k$ and let $x \in R_n$ such that $\text{supp}(x) \subseteq T$, so $x \in \Sigma_{2k}$. If $\sum_{\ell=1}^{2k} x_{i_\ell} c_{i_\ell} = \Phi x = \theta$ then $x \in N(\Phi)$, so $x = \theta$ by (2). Thus the $2k$ column vectors indexed by T must be linearly independent.
- (3) \rightarrow (4). Let T and x be chosen as in the preceding proof and consider the quadratic form

$$\begin{aligned} \langle x, \Phi^{\text{tr}} \Phi x \rangle &= \sum_{s,t=1}^{2k} x_{i_s} (c_{i_s} \cdot c_{i_t}) x_{i_t} = \left(\sum_{s=1}^{2k} x_{i_s} c_{i_s} \right) \cdot \left(\sum_{t=1}^{2k} x_{i_t} c_{i_t} \right) \\ &= \left\| \sum_{t=1}^{2k} x_{i_t} c_{i_t} \right\|_2^2. \end{aligned}$$

By (3), the $2k$ column vectors c_{i_t} are linearly independent, so the quadratic form $\langle x, \Phi^{\text{tr}} \Phi x \rangle$ is > 0 for all nonzero x with $\text{supp}(x) \subseteq T$. Note that

the matrix $\Phi_T^{\text{tr}}\Phi_T$ is symmetric. It is a well known fact from linear algebra that any $N \times N$ real symmetric matrix A has N real eigenvalues and that these eigenvalues are all positive if and only if the quadratic form $\sum_{i,j=1}^N y_i A_{ij} y_j > 0$ for all real nonzero vectors y . Furthermore a real symmetric matrix with all positive eigenvalues is invertible. This establishes (4).

(4) \rightarrow (1). Suppose $x^{(1)}, x^{(2)} \in F(y) \cap \Sigma_k$ for some $y \in R_m$. Setting $x = x^{(1)} - x^{(2)}$ we see that $\Phi x = \Phi x^{(1)} - \Phi x^{(2)} = y - y = \theta$ and $x \in \Sigma_{2k}$. Let T be the index set $\{i_t\}$ for the support of x . Then the fact that x is in the null space of Φ can be expressed in terms of the column vectors of Φ as $\sum_{t=1}^{2k} x_{i_t} c_{i_t} = \Phi x = \theta$. Thus

$$\langle x, \Phi^{\text{tr}}\Phi x \rangle = \sum_{s,t=1}^{2k} x_{i_s} (c_{i_s} \cdot c_{i_t}) x_{i_t} = \left\| \sum_{t=1}^{2k} x_{i_t} c_{i_t} \right\|_2^2 = 0$$

By (4), the eigenvalues of $\Phi_T^{\text{tr}}\Phi_T$ are all positive so the only way for this last sum to vanish is if $x = \theta$. This implies $x^{(1)} = x^{(2)}$.

Q.E.D.

If the $m \times n$ sample matrix Φ satisfies the requirements of the theorem for some k with $2k \leq m < n$ then we can show the existence of an encoder/decoder system that reproduces any k -sparse signal x . Necessary and sufficient conditions for unique reproduction are that every subset of $2k$ column vectors must be linearly independent. If this is true we have $m \geq 2k$. Now let T be any $2k$ -index set that contains the support of x . Then the encoder gives the sample $y = \Phi x = \Phi_T x$, so $\Phi_T^{\text{tr}} y = (\Phi_T^{\text{tr}} \Phi_T) x$. Since the $2k \times 2k$ matrix $\Phi_T^{\text{tr}} \Phi_T$ is invertible, we can recover x from y via the decoding operation

$$x = \Delta(y) = (\Phi_T^{\text{tr}} \Phi_T)^{-1} \Phi_T^{\text{tr}} y.$$

Although this construction requires us to find some T containing the support of x , we can avoid this by using the ℓ_1 (or any ℓ_p) norm to find x from a minimization problem:

$$x = \Delta(y) = \text{Argmin}_{z \in \Sigma_k} \|y - \Phi z\|_1,$$

where by Argmin we mean the vector z that achieves the minimum value of $\|y - \Phi z\|_1$. Indeed the minimum value of 0 is achieved for exactly one $x \in \Sigma_k$, as the theorem demonstrates.

Example 1 Consider the $2k \times n$ Vandermonde matrix

$$\Phi = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ a_1 & a_2 & \cdots & a_n \\ a_1^2 & a_2^2 & \cdots & a_n^2 \\ \vdots & \vdots & \vdots & \vdots \\ a_1^{k-1} & a_2^{k-1} & \cdots & a_n^{k-1} \end{pmatrix},$$

where $a_1 < a_2 < \cdots < a_n$. It is an exercise in linear algebra to show that the determinant of a square Vandermonde matrix is equal to $\pm \prod_{i>j} (a_i - a_j) \neq 0$, i.e., a square Vandermonde matrix is invertible. Thus any $2k \times 2k$ submatrix of Φ has rank $2k$ and Φ satisfies the conditions of Theorem 1 with $m = 2k$.

Exercise 2 Work out explicitly the action of the Vandermonde coder/decoder for a) $k = 1$, and b) $k = 2$.

1.3 Analytic theory

The algebraic solution to the compressed sampling problem presented in the preceding section is, by itself, of only modest practical significance. Firstly, there is a significant practical problem of actually computing x for large values of n . Also, the solution assumes that the signals are precisely k -sparse and that the inverse of the matrix $(\Phi_T^{\text{tr}} \Phi_T)^{-1}$ can be computed with no error. Real signals typically have a few spikes with all other components small but not zero. The numerical computation of the matrix inverse may be unstable for large n and k . (This is the case with the Vandermonde matrix.) The signals may be partially corrupted by noise. We need to develop analytic estimates that enable us to determine how well the encoding/decoding procedure approximates the initial signal. We also need to be concerned with the design of the decoder, so that it can efficiently compute the approximation.

1.3.1 Recovering a sparse solution of $y = \Phi x$ via ℓ_1 minimization

Theorem 1 gave necessary and sufficient conditions that there was at most one signal $\hat{x} \in \Sigma_k \cap F(y)$ that produced a given sample y . (Recall that if $x = \tilde{x}$ is one solution of the equation $y = \Phi x$ then all solutions are of the form $x \in F(y) =$

$\{\tilde{x} + h : h \in N(\Phi)\}$ where $N(\Phi)$ is the null space of the $m \times n$ sampling matrix Φ .) However, the theorem did not lead directly to an efficient method for explicit computation of the k -sparse solution from the sample y . Here we will study the feasibility of finding \hat{x} by solving the ℓ_1 minimization problem

$$\hat{x} = \operatorname{Argmin}_{x \in F(y)} \|x\|_1 \quad (1.3)$$

Already in Section ?? we have seen examples of the special ability of the ℓ_1 norm to produce solutions x of $y = \Phi x$ with maximal sparsity. Now we will find necessary and sufficient conditions that this ℓ_1 minimization will lead to a unique k -sparse solution.

Suppose that the equation $y = \Phi x$ has a k -sparse solution \hat{x} and that this solution also satisfies (1.3), i.e., it has minimal ℓ_1 norm. This means that

$$\|\hat{x} + h\|_1 \geq \|\hat{x}\|_1$$

for all $h \in N(\Phi)$. If $T \subset \{1, 2, \dots, n\}$ is the index set for the support of \hat{x} and T^c is the set of remaining indices, then $\#(T) = k$, $\#(T^c) = n - k$ and the minimization condition can be written in terms of absolute values as

$$\sum_{i \in T} |\hat{x}_i + h_i| + \sum_{i \in T^c} |h_i| \geq \sum_{i \in T} |\hat{x}_i|.$$

Exercise 3 Show that for any two real numbers a, b it is always true that $|a + b| - |a| \geq \operatorname{sgn}(a)b$, where $\operatorname{sgn}(a) = 1$ if $a \geq 0$ and $\operatorname{sgn}(a) = -1$ if $a < 0$. In particular, show that this is a strict equality unless $ab < 0$ and $|b| > |a|$.

Using the result of Exercise 3 we can obtain a minimization condition

$$\|\hat{x} + h\|_1 - \|\hat{x}\|_1 \geq \sum_{i \in T} \operatorname{sgn}(\hat{x}_i)h_i + \sum_{i \in T^c} |h_i|, \quad (1.4)$$

where the left-hand side is ≥ 0 for all $h \in N(\Phi)$. Similarly, since $N(\Phi)$ is a vector space we can replace h by $-h$ in (1.4) and obtain

$$\|\hat{x} - h\|_1 - \|\hat{x}\|_1 \geq -\sum_{i \in T} \operatorname{sgn}(\hat{x}_i)h_i + \sum_{i \in T^c} |h_i|, \quad (1.5)$$

where again the left-hand side is ≥ 0 for all $h \in N(\Phi)$. It follows from (1.4) and (1.5) that the strict inequality

$$\sum_{i \in T} \operatorname{sgn}(\hat{x}_i)h_i < \sum_{i \in T^c} |h_i| \quad (1.6)$$

for all nonzero $h \in N(\Phi)$ is a sufficient condition that $\|\hat{x} + h\|_1 - \|\hat{x}\|_1 > 0$, hence that \hat{x} is the unique ℓ_1 minimum. If we replace $<$ by \leq in (1.6) then we have a sufficient condition that \hat{x} is a minimum, but uniqueness isn't guaranteed.

Theorem 2 *Suppose $x = \hat{x}$ is a k -sparse solution of the equation $y = \Phi x$ and T is the index set of \hat{x} . A necessary and sufficient condition that \hat{x} is a minimal ℓ_1 -norm solution is*

$$\sum_{i \in T} \text{sgn}(\hat{x}_i) h_i \leq \sum_{i \in T^c} |h_i|, \forall h \in N(\Phi). \quad (1.7)$$

A necessary and sufficient condition that \hat{x} is the unique minimal ℓ_1 -norm solution is

$$\sum_{i \in T} \text{sgn}(\hat{x}_i) h_i < \sum_{i \in T^c} |h_i|, \forall h \in N(\Phi), h \neq \theta. \quad (1.8)$$

PROOF: We have already shown the sufficiency. To show that the conditions are necessary, suppose there is a $\tilde{h} \in N(\Phi)$ such that $\sum_{i \in T} \text{sgn}(\hat{x}_i) \tilde{h}_i > \sum_{i \in T^c} |\tilde{h}_i|$. Then the same strict inequality will hold for all $h = \alpha \tilde{h} \in N(\Phi)$ where $\alpha > 0$. Now choose α so small that $|h_i| < |\hat{x}_i|$ for all $i \in T$. Then from one of the results of Exercise 3 we have strict equality $|\hat{x}_i + h_i| - |\hat{x}_i| = \text{sgn}(\hat{x}_i) h_i$ for all $i \in T$. Thus

$$\|\hat{x} - h\|_1 - \|h\|_1 = - \sum_{i \in T} \text{sgn}(\hat{x}_i) h_i + \sum_{i \in T^c} |h_i| < 0,$$

so \hat{x} is not an ℓ_1 minimum. Q.E.D.

Note that recovery of a k -sparse signal \hat{x} depends not on the magnitude of the nonzero components of the signal, but just on its pattern of signs. A key insight provided by Theorem 2 is that sparse signal recovery via sampling is closely related to the structure of the null space $N(\Phi)$.

Corollary 1 *A necessary and sufficient condition that **every** k -sparse signal \hat{x} can be obtained as the unique minimal ℓ_1 -norm solution to its sample equation $y = \Phi x$ is that*

$$\sum_{i \in T} |h_i| < \sum_{i \in T^c} |h_i|, \forall h \in N(\Phi), h \neq \theta, \quad (1.9)$$

for all k -index sets T or, equivalently,

$$\|h\|_1 < 2 \sum_{i \in T^c} |h_i|. \quad (1.10)$$

Condition (1.9), while difficult to verify in practice, points to an important property that the null space must possess. In order that the sample matrix be able to reconstruct sparse signals supported on small index sets T , the null space must counterbalance by being distributed through all indices. This is often referred to as an uncertainty principle for compressive sampling. It suggests that the components of the sampling matrix, rather than being highly structured to fit a particular type of signal, should be chosen at random. We shall see that this insight is correct.

Exercise 4 *Prove Corollary 1.*

If we can show that the conditions of Theorem 2 are satisfied then we can recover uniquely the k -sparse signal \hat{x} from the sample y by solving the minimization problem

$$\hat{x} = \operatorname{Argmin}_{x \in F(y)} \|x\|_1. \quad (1.11)$$

Although, unlike ℓ_2 (least squares) minimization, this problem has no analytic solution, it can be solved by numerically efficient linear programming routines, that are widely available.

This text is not meant to cover linear programming, but it is worth pointing out that the ℓ_1 minimization problem is equivalent to a linear programming problem. To see this we first express the $m \times n$ sampling matrix in terms of its row vectors r_i :

$$\Phi = \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix}, \quad r_i \in R_n, \quad i = 1, \dots, m.$$

Then we can express the ℓ_1 minimization problem (1.11) as the linear programming problem

$$\min \sum_{j=1}^n u_j, \text{ such that } y_i = r_i \cdot x, \quad -u_j \leq x_j \leq u_j. \quad (1.12)$$

Similarly the (possibly overdetermined) ℓ_1 problem

$$\min_{x \in F(y)} \|y - \Phi x\|_1 = \min \sum_{i=1}^m |y_i - r_i \cdot x| \quad (1.13)$$

can be expressed as the linear programming problem

$$\min \sum_{i=1}^m u_i, \text{ such that } -u_i \leq y_i - r_i \cdot x \leq u_i. \quad (1.14)$$

In general, a linear programming problem is a maximization or minimization problem that can be expressed entirely through linear equations and linear inequalities. Such problems can be solved, for example, by the simplex method [].

Exercise 5 Show that the ℓ_∞ minimization problems analogous to (1.11) and (1.13) can be expressed as problems in linear programming.

1.3.2 A theoretical structure for deterministic compressive sampling

Instead of restricting to signals $x \in R_n$ that are strictly k -sparse we will consider the larger class of compressible signals, those that can be approximated by elements of Σ_k . As a measure of the approximation of x by k -sparse signals we adopt

$$\sigma_k(x) = \inf_{z \in \Sigma_k} \|x - z\|_1. \quad (1.15)$$

Exercise 6 We order the components of $x \in R_n$ in terms of the magnitude of the absolute value, so that

$$|x_{i_1}| \geq |x_{i_2}| \geq \dots \geq |x_{i_n}|.$$

Show that $\sigma_k(x) = \sum_{j=k+1}^n |x_{i_j}|$, i.e., $\sigma_k(x)$ is the sum of the absolute values of the $n - k$ smallest components of x . Show that

$$\text{Argmin}_{z \in \Sigma_k} \|x - z\|_1 = x^k$$

where x^k has index set $T = \{i_1, i_2, \dots, i_k\}$. This shows that x^k is the closest k -sparse approximation of x with respect to the ℓ_1 norm.

Exercise 7 If

$$x = (3, -2, 6, 0, 3, -1, 2, -5) \in R_8,$$

find x^k and $\sigma_k(x)$ for $k = 1, 2, \dots, 7$.

Clearly, $\sigma_k(x) = 0 \iff x \in \Sigma_k$. We could consider x as a good candidate for compression provided $\sigma_k(x) < \epsilon$ for some suitably small ϵ . If x is compressible but not exactly k -sparse we cannot expect to reproduce it exactly, but we can try to approximate it by a k -sparse signal. (We assume that our decoders Δ are

designed to give k -sparse reconstructions of x .) Let K be a compact set in the signal space R_n , say a ball $\|x\|_p \leq C$ for some constant C . We can measure the performance of a given encoder/decoder pair Φ, Δ on K by determining the maximum error in estimating any $x \in K$ by $\Delta(\Phi x) = \Delta(y) \in \Sigma_k$. We will use the measure

$$E_m(K, p) = \sup_{x \in K} \|x - \Delta(\Phi x)\|_p. \quad (1.16)$$

(Recall that Φ is $m \times n$.) Here, $E_m(K, p) = 0$ means that the error is 0.

At this point we can give a precise description of the mathematical properties that we want an encoder/decoder pair to satisfy. We say that (Φ, Δ) is **Instance-Optimal** of order k in the ℓ_1 norm if there is a constant $C > 0$ such that

$$\|x - \Delta(\Phi x)\|_1 \leq C \sigma_k(x) = C \|x - x^k\|_1, \quad \forall x \in R_n. \quad (1.17)$$

Clearly, this is an important requirement for an encoder/decoder since it implies that as the signals x are increasingly k -compressible, i.e., as $x - x^k \rightarrow 0$, the k -sparse output of the decoder is a better and better approximation of the signal. In particular, if $x \in \Sigma_k$ then the decoder recovers x . Note that $\sigma_k(h) = \|h - h^k\|_1 = \sum_{i \in T^c} |h_i|$ where T is the index set of h^k . possible encoder/decoder pairs that achieve the optimum.

Definition 1 Given a sampling matrix Φ , for each $k \leq m$ let $\rho_k > 0$ be the smallest constant such that $\|h\|_1 \leq (1 + \rho_k) \sigma_k(h)$ for every $h \in N(\Phi)$. (Recall that $\sigma_k(h) = \|h - h^k\|_1 = \sum_{i \in T^c} |h_i| = \|h_{T^c}\|_1$ and $\|h\|_1 = \|h_T\|_1 + \|h_{T^c}\|_1$ where T is the index set of h^k .) If $\rho_k < 1$ we say that Φ satisfies the **null space property** for k -sparse signals.

It is possible that $\rho_k = \infty$. This happens when there is a nonzero k -sparse $h \in N(\Phi)$, a very undesirable feature for compressive sampling.

Exercise 8 Show that the null space property is equivalent to $\|h_T\|_1 \leq \rho_k \|h_{T^c}\|_1$.

Exercise 9 Show that $\rho_1 \leq \rho_2 \leq \dots \leq \rho_k \leq \dots$.

Exercise 10 Show that $\rho_k = 1/\lambda_k - 1$ where $\lambda_k = \max_{\|h\|_1=1, S} \sum_{i \in S^c} |h_i|$ and the maximum is taken over all k -element index sets S . Verify that $\rho_k > 0$.

It follows from (1.10) that if $\rho_k < 1$ then we satisfy the condition for uniquely capturing a k -sparse solution of $\Phi x = y$.

Now we come to the essence of our problem. Suppose $x \in R_n$ is a signal that we consider is “nearly” k -sparse and let $y = \Phi x \in R_m$ be its corresponding sample. We define a decoder

$$\Delta(y) = \min_{z, \Phi z = y} \|z\|_1 = \hat{x}.$$

We want our decoder to capture the closest k -sparse approximation x^k to x and ignore the smallest $n - k$ components.. Thus we want to guarantee that the decoder will yield a unique k -sparse solution \hat{x} that is as close as possible to x . In particular, if $x \in \Sigma_k$ we want $\hat{x} = x$.

To clarify the situation we follow an approach similar to the proof of Theorem 2. Since $y = \Phi \hat{x}$, there is a unique $h \in N(\Phi)$ such that $\hat{x} = x + h$. Since \hat{x} is a ℓ_1 minimum solution we must have

$$\|\hat{x}\|_1 = \|x + h\|_1 \leq \|x\|_1. \quad (1.18)$$

Now let T_0 be the set of indices corresponding to the k largest components of x in absolute value. (Thus T_0 is the index set of x^k .) Recall from the triangle inequality $|a| = |(a + b) + (-b)| \leq |a + b| + |-b|$, so $|a + b| \geq |a| - |b|$ and, by symmetry, $|a + b| \geq |b| - |a|$. Also, for any index set T and n -tuple z , let z_T be the vector with components $(z_T)_i = z_i$ for $i \in T$ and all other components zero. Then

$$\begin{aligned} \|x + h\|_1 &= \sum_{i \in T_0} (|x_i + h_i|) + \sum_{i \in T_0^c} (|x_i + h_i|) \geq \sum_{i \in T_0} (|x_i| - |h_i|) + \sum_{i \in T_0^c} (|h_i| - |x_i|) \\ &= (\|x\|_1 - \|x_{T_0}\|_1 - \|h_{T_0}\|_1) + (\|h_{T_0^c}\|_1 - \|x_{T_0^c}\|_1) \\ &= (\|x\|_1 + \|h_{T_0^c}\|_1) - (2\|x_{T_0^c}\|_1 + \|h_{T_0}\|_1). \end{aligned} \quad (1.19)$$

Comparing the inequalities (1.18) and (1.19) and noting that $\|x_{T_0^c}\|_1 = \|x - x^k\|_1$, we obtain

$$2\|x - x^k\|_1 \geq \|h_{T_0^c}\|_1 - \|h_{T_0}\|_1 \geq (1 - \rho_k)\|h_{T_0^c}\|_1. \quad (1.20)$$

If Φ satisfies the null space property then $\rho_k < 1$ and we obtain

$$\|h_{T_0^c}\|_1 \leq \frac{2}{1 - \rho_k} \|x - x^k\|_1.$$

Since $\hat{x} = x + h$, it follows that

$$\|\hat{x} - x\|_1 = \|h\|_1 = \|h_{T_0}\|_1 + \|h_{T_0^c}\|_1 \leq (\rho_k + 1)\|h_{T_0^c}\|_1 \leq 2\frac{1 + \rho_k}{1 - \rho_k} \|x - x^k\|_1.$$

Theorem 3 *If Φ has the null space property for k -sparse signals then*

$$\|\hat{x} - x\|_1 \leq 2 \frac{1 + \rho_k}{1 - \rho_k} \|x - x^k\|_1. \quad (1.21)$$

The proof of the theorem shows that the k -sparse solution \hat{x} of the minimization problem is unique, and allows us to get a bound on the error of reconstruction of the nearly k -sparse signal x by \hat{x} . In particular, if $x = x^k$, i.e., x is k -sparse, then $x = \hat{x}$. It is still possible that there is more than one k -sparse signal with sample y even though any other solution doesn't satisfy the ℓ_1 minimum property. In the next section we shall see how that possibility can be eliminated.

1.3.3 RIP

Theorem 3 is an important theoretical result but it will not be of much practical use unless we find methods for determining ρ and for designing encoder/decoder pairs with $\rho_k < 1$ for k small with respect to n . The null space property doesn't have direct intuitive meaning for us. A more intuitive concept is the Uniform Uncertainty Principle or Restricted Isometry Property (RIP). Let Φ be our usual $m \times n$ sampling matrix, For each k we define the restricted isometry constant δ_k as the smallest nonnegative number such that

$$(1 - \delta_k) \|x\|_2^2 \leq \|\Phi(x)\|_2^2 \leq (1 + \delta_k) \|x\|_2^2 \quad (1.22)$$

for all k -sparse signals x , i.e., $x \in \Sigma_k$. (Note that here we are using the ℓ_2 norm.) Expression (1.22) captures in an efficient and easily understandable manner how well the k -sparse signal x is captured by the sample $y = \Phi x \in R_m$. Indeed if $\delta_k \geq 1$ so that the left-hand term is ≤ 0 then we could have a nonzero signal that produces a zero sample, so that we would have no capability of recapturing it. Thus a desirable property for an encoder is $\delta_k < 1$.

Definition 2 *Φ has the Restricted Isometry Property (RIP) for k if it has a restricted isometry constant such that $0 < \delta_k < 1$.*

Another way to understand RIP is in terms of eigenvalues. If x is k -sparse with index set T then

$$\|\Phi x\|_2^2 = \langle \Phi_T x, \Phi_T x \rangle = \langle \Phi_T^{\text{tr}} \Phi_T x, x \rangle$$

As pointed out in the proof of Theorem 1, $\Phi^{\text{tr}}\Phi$ is self-adjoint and positive definite. Thus all eigenvalues $\lambda_i(T)$ of this square matrix are strictly positive and $\lambda_{\min}(T)\|x\|_2^2 \leq \langle \Phi_T^{\text{tr}}\Phi_T x, x \rangle \leq \lambda_{\max}(T)\|x\|_2^2$. It follows that

$$1 - \delta_k = \lambda_{\min} \leq \lambda_{\max} = 1 + \delta_k, \quad (1.23)$$

where the maximum and minimum are taken over all index sets T with $\leq k$ indices.

Exercise 11 Let Φ be a sample matrix with identical columns $\Phi = (c, c, \dots, c)$ where $\|c\|_2 = 1$. Show that $\delta_1 = 0$.

Exercise 12 Use the eigenvalue property (1.23) to show that for any sample matrix Φ it is not possible for $\delta_k = 0$ if $k \geq 2$.

Exercise 13 Use the $m \times n$ sample matrix Φ with constant elements $\Phi_{ij} = 1/\sqrt{m}$ and k -sparse signals x such that $x_j = 1$ for $j \in T$ to show that this sample matrix is not RIP for $k \geq 2$.

If we have two signals $x, x' \in \Sigma_k$ then $x - x'$ may not be k -sparse, but it is $2k$ -sparse. Thus $(x - x') \in \Sigma_{2k}$ and

$$(1 - \delta_{2k})\|x - x'\|_2^2 \leq \|\Phi(x - x')\|_2^2 \leq (1 + \delta_{2k})\|x - x'\|_2^2. \quad (1.24)$$

If $x \neq x'$ then, to distinguish the signals, we must have $\delta_{2k} < 1$, i.e., Φ must satisfy RIP for $2k$. Moreover, for x very close to x' in norm we want the samples $y = \Phi x$ and $y' = \Phi x'$ to be very close in norm, and this would be implied by $\|\Phi(x - x')\|_2^2 \leq (1 + \delta_{2k})\|x - x'\|_2^2$.

It is worth pointing out here that the basic idea behind RIP is pervasive in signal processing theory. For signals in Hilbert spaces the concept is called a **frame**. Frames will be discussed in Section ???. The concepts are not identical because frames refer to vector spaces and Σ_k is not a vector space.

Now we will show that the null space property can be implied by the more intuitive RIP. Since RIP is expressed in terms of ℓ_2 whereas the null space property is expressed in terms of ℓ_1 , we need the result from Section ??? that the norms of a k -sparse signal x satisfy the relation $\|x\|_2 \leq \|x\|_1 \leq \|x\|_2/\sqrt{k}$. Now suppose Φ satisfies RIP for some fixed $2k$. This will imply that for each sample $y \in R_m$ there will be at most one k -sparse solution x to $y = \Phi x$. In particular $N(\Phi)$ will contain no k -sparse vectors.

Now let us examine what RIP implies about the null space property. We want to guarantee that for all k -index sets T the inequality $\|h_T\|_1 < \rho_k \|h_{T^c}\|_1$ holds for a constant $\rho_k < 1$ and all $h \in N(\Phi)$. We start by taking any $h \in N(\Phi)$. Our approach will be to take the worst case possible for h : the $T = T_0$ is the index set of the k largest components of h , in absolute value. We want to guarantee that $\|h_{T_0}\|_1 < \rho_k \|h_{T_0^c}\|_1$. Note that RIP applies only to k -sparse signals and, in general, the h is not k -sparse. In order to apply RIP we will decompose $h_{T_0^c}$ as a sum of signals h_{T_ℓ} , each of which is at most k -sparse. Noting that h is an n -tuple we use the Euclidean algorithm to write $n = ak + r$ where a is a positive integer and the integer r is the remainder, $0 \leq r < k$. We divide the indices of h into $a + 1$ disjoint sets. The first set T_0 contains the k indices i for which $|h_i|$ is maximal, the set T_1 contains the indices of the next k maximal components, and so forth. The last index set T_a contains the indices i of the remaining r components, for which $|h_i|$ is the smallest. Now let h_{T_ℓ} be the n -tuple such that the component $(h_{T_\ell})_i = h_i$ for $i \in T_\ell$ and $(h_{T_\ell})_i = 0$ for $i \in T_\ell^c$.

Exercise 14 Show that 1) $h_{T_\ell} \in \Sigma_k$, 2) $h_{T_\ell} \cdot h_{T_{\ell'}} = 0$ for $\ell \neq \ell'$, and 3)

$$h = \sum_{\ell=0}^a h_{T_\ell}. \quad (1.25)$$

The following material is quite technical, but it leads to the easily understood result: Theorem 4 with an upper bound for the constant ρ_k in the null space property as a function of the index δ_{2k} in RIP. Let $h_{T_0 \cup T_1}$ be the n -tuple such that $(h_{T_0 \cup T_1})_i = h_i$ for $i \in T_0 \cup T_1$ with all other components zero. Thus $h_{T_0 \cup T_1} \in \Sigma_{2k}$. We can apply RIP to each k -sparse signal h_{T_ℓ} and from (1.25) we have

$$\Phi(h) = \Phi(h_{T_0 \cup T_1}) + \sum_{\ell=2}^a \Phi(h_{T_\ell}) = \theta. \quad (1.26)$$

Thus,

$$\|\Phi h_{T_0 \cup T_1}\|_2^2 = \langle \Phi h_{T_0 \cup T_1}, -\sum_{j=2}^a \Phi h_{T_j} \rangle = -\sum_{j=2}^a \langle \Phi h_{T_0}, \Phi h_{T_j} \rangle - \sum_{j=2}^a \langle \Phi h_{T_1}, \Phi h_{T_j} \rangle$$

so

$$\|\Phi h_{T_0 \cup T_1}\|_2^2 \leq \sum_{j=2}^a (|\langle \Phi h_{T_0}, \Phi h_{T_j} \rangle| + |\langle \Phi h_{T_1}, \Phi h_{T_j} \rangle|). \quad (1.27)$$

To obtain an upper bound for the right-hand side we need the following result.

Lemma 1 Suppose $x, x' \in \Sigma_k$ with k -index sets S, S' respectively and such that S and S' don't intersect. Let $\langle \cdot, \cdot \rangle$ be the ℓ_2 inner product on R_n . Then $|\langle \Phi x, \Phi x' \rangle| \leq 2\delta_{2k} \|x\|_2 \|x'\|_2$.

PROOF: Let

$$\kappa = \max_{\|z\|_2 = \|z'\|_2 = 1} |\langle \Phi z, \Phi z' \rangle|$$

where the maximum is taken for all k -sparse unit vectors z with index set S and k -sparse unit vectors z' with index set S' . Then, by renormalizing, it follows that $|\langle \Phi x, \Phi x' \rangle| \leq \kappa \|x\|_2 \|x'\|_2$. Since $z + z'$ and $z - z'$ are $2k$ -sparse and $\|z \pm z'\|_2^2 = \|z\|_2^2 + \|z'\|_2^2 = 2$, we have the RIP inequalities

$$2(1 - \delta_{2k}) \leq \|\Phi(z + z')\|_2^2 \leq 2(1 + \delta_{2k}),$$

$$2(1 - \delta_{2k}) \leq \|\Phi(z - z')\|_2^2 \leq 2(1 + \delta_{2k}).$$

From the parallelogram law for ℓ_2 , see (??), we have

$$\langle \Phi z, \Phi z' \rangle = \frac{1}{4} (\|\Phi(z + z')\|_2^2 - \|\Phi(z - z')\|_2^2),$$

so the RIP inequalities imply $|\langle \Phi z, \Phi z' \rangle| \leq \delta_{2k}$. Thus $\kappa \leq \delta_{2k}$. Q.E.D.

Exercise 15 Verify the details of the proof of Lemma 1.

Applying Lemma 1 to the right-hand side of (1.27) we find

$$\|\Phi h_{T_0 \cup T_1}\|_2^2 \leq \delta_{2k} (\|h_{T_0}\|_2 + \|h_{T_1}\|_2) \left(\sum_{j=2}^a \|h_{T_j}\|_2 \right). \quad (1.28)$$

On the right we use the standard inequality $\|h_{T_0}\|_2 + \|h_{T_1}\|_2 \leq \sqrt{2} \|h_{T_0 \cup T_1}\|_2$, see (??), and on the left-hand side we use RIP for $2k$ to obtain

$$(1 - \delta_{2k}) \|h_{T_0 \cup T_1}\|_2^2 \leq \|\Phi h_{T_0 \cup T_1}\|_2^2 \leq \sqrt{2} \delta_{2k} \|h_{T_0 \cup T_1}\|_2 \left(\sum_{j=2}^a \|h_{T_j}\|_2 \right),$$

so

$$\|h_{T_0 \cup T_1}\|_2 \leq \frac{\sqrt{2} \delta_{2k}}{1 - \delta_{2k}} \sum_{\ell=2}^a \|h_{T_\ell}\|_2. \quad (1.29)$$

Now for some tricky parts and the reason that we have chosen the T_ℓ by putting the indices of h in decreasing order of magnitude.

Lemma 2

$$\|h_{T_\ell}\|_2 \leq \frac{1}{\sqrt{k}} \|h_{T_{\ell-1}}\|_1, \quad \ell = 2, \dots, a.$$

PROOF: For any $j \in T_\ell, i \in T_{\ell-1}$, we have $|h_j| \leq |h_i|$, so $|h_j|$ is also bounded by the average of the $|h_i|$ as i ranges over $T_{\ell-1}$:

$$|h_j| \leq \frac{1}{k} \sum_{i \in T_{\ell-1}} |h_i|.$$

Thus

$$\|h_{T_\ell}\|_2^2 = \sum_{j \in T_\ell} |h_j|^2 \leq k \left(\frac{\sum_{i \in T_{\ell-1}} |h_i|}{k} \right)^2.$$

Taking the square root, we have $\|h_{T_\ell}\|_2 \leq \|h_{T_{\ell-1}}\|_1 / \sqrt{k}$. Q.E.D.

From the lemma we have

$$\sum_{\ell=2}^a \|h_{T_\ell}\|_2 \leq (k)^{-1/2} \sum_{\ell=1}^{a-1} \|h_{T_\ell}\|_1 \leq (k)^{-1/2} \sum_{\ell=1}^a \|h_{T_\ell}\|_1 = (k)^{-1/2} \|h_{T_0^c}\|_1,$$

which gives an upper bound for the right-hand side of (1.29). From $\|h_{T_\ell}\|_1 / \sqrt{k} \leq \|h_{T_\ell}\|_2$ we get a lower bound for the left-hand side of (1.29): $\|h_{T_0}\|_1 / \sqrt{k} \leq \|h_{T_0}\|_2 \leq \|h_{T_0 \cup T_1}\|_2$. Putting this all together we find

$$\|h_{T_0}\|_1 \leq \frac{\sqrt{2}\delta_{2k}}{1 - \delta_{2k}} \|h_{T_0^c}\|_1 = \rho_k \|h_{T_0^c}\|_1. \quad (1.30)$$

To guarantee the null space property we must have $\rho_k < 1$.

Theorem 4 *A sufficient condition for the null space property to hold and for each class $F(y)$ to contain at most one k -sparse signal is that RIP holds for $2k$ -sparse signals with*

$$\rho_k = \frac{\sqrt{2}\delta_{2k}}{1 - \delta_{2k}} < 1.$$

It follows that if $\delta_{2k} < 1/(1 + \sqrt{2}) \approx 0.4142$ then the null space property is satisfied.

Exercise 16 *If $\delta_{2k} = 1/4$ for some k and the sample matrix Φ , verify from Theorem 1.21 that the estimate $\|\hat{x} - x\|_1 \leq 5.5673 \|x - x^k\|_1$ holds for the approximation of a signal x by a k -sparse signal \hat{x} .*

Many different bounds for ρ_k can be derived from RIP using modifications of the preceding argument, but the result of Theorem 4, due to Candés and Tao, is sufficient for our purposes. The point is that for a given $m \times n$ sample matrix Φ we would like to be able to guarantee the null space property for k as large as possible, and for a given k we would like to design appropriate sample matrices with m as small as possible.

1.4 Probabilistic Theory

For practical computation of sampling matrices Φ for compressive sensing we employ a probabilistic approach. Rather than designing a sampling matrix especially adapted for a given type of signal, we chose a general purpose sampling matrix at random. To motivate this approach we recall from Section [] that all eigenvalues $\lambda_i(T)$ of the self-adjoint matrix $\Phi_T^{\text{tr}}\Phi_T$ are strictly positive and

$$1 - \delta_k = \lambda_{\min} \leq \lambda_{\max} = 1 + \delta_k$$

where the maximum and minimum are taken over all index sets T with $\#(T) = k$. Thus RIP is satisfied provided $\lambda_i(T) \approx 1$. This means that $\Phi_T^{\text{tr}}\Phi_T$ must be close to the $T \times T$ identity matrix for every k -index set T . Thus the column vectors c_j of the sampling matrix $\Phi = (c_1, \dots, c_m)$ should satisfy $c_{j_1} \cdot c_{j_2} \approx 0$ for $j_1 \neq j_2$ and $c_j \cdot c_j \approx 1$ to obtain RIP. We want each column vector to be approximately an ℓ_2 unit vector but distinct columns should be nearly orthogonal, i.e., uncorrelated. One way to try to achieve this is to choose the elements of Φ as independent samples from a probability space with mean 0 and standard deviation 1.

We recall some basic definitions from probability theory. A continuous probability distribution for the random variable t on the real line R is a continuous function $\rho(t)$ on R such that $0 \leq \rho(t) \leq 1$ and $\int_{-\infty}^{\infty} \rho(t) dt = 1$. We also require that $\int_{-\infty}^{\infty} p(t)\rho(t) dt$ converges for any polynomial $p(t)$. Here $\int_{t_1}^{t_2} \rho(t) dt$ is interpreted as the probability that a sample t taken from R falls in the interval $t_1 \leq t \leq t_2$. The expectation (or mean) \bar{t} of the distribution is $\bar{t} = E_{\rho}(t) \equiv \int_{-\infty}^{\infty} t\rho(t) dt$ and the standard deviation $\sigma \geq 0$ is defined by

$$\sigma^2 = \int_{-\infty}^{\infty} (t - \bar{t})^2 \rho(t) dt = E_{\rho}((t - \bar{t})^2).$$

Here σ is a measure of the concentration of the distribution about its mean. The most famous continuous distribution is the normal (or Gaussian) distribution func-

tion

$$\rho_0(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/2\sigma^2} \quad (1.31)$$

where μ is a real parameter and $\sigma > 0$. This is just the bell curve, centered about $t = \mu$. In this case $E_{\rho_0}(t) = \mu$ and $\sigma^2 = E_{\rho_0}(t^2)$. In Section §?? the normal distribution has already arisen through its connection with the Heisenberg uncertainty principle. The standard notation for the normal distribution with mean μ and standard deviation σ is $N(\mu, \sigma)$.

Now suppose we choose each of the mn matrix elements of $\Phi = (\Phi_{ij}) = (t_{ij})$, $1 \leq i \leq m$, $1 \leq j \leq n$, independently from the population with normal distribution $N(0, 1/\sqrt{m})$. Then $\mathbf{t} = (t_{ij})$ is a matrix random variable on R_{mn} with probability density function

$$\Delta(\mathbf{t}) = \left(\frac{m}{2\pi}\right)^{mn/2} \exp\left(-\frac{m}{2} \sum_{i=1}^m \sum_{j=1}^n t_{ij}^2\right). \quad (1.32)$$

Given any function $f(\mathbf{t}) \in L_1(R_{mn}, \Delta)$ we define its expectation by

$$\bar{f}(\mathbf{t}) = E(f(\mathbf{t})) = \int_{R_{mn}} f(\mathbf{t}) \Delta(\mathbf{t}) \Pi_{ij} dt_{ij}.$$

Note that E is linear, i.e.,

$$E(\alpha f_1(\mathbf{t}) + \beta f_2(\mathbf{t})) = \alpha E(f_1(\mathbf{t})) + \beta E(f_2(\mathbf{t}))$$

for real parameters α, β . Further we have the properties

$$E(1) = 1, \quad E(t_{i_1 j_1} t_{i_2 j_2}) = \delta_{i_1 i_2} \delta_{j_1 j_2} / m. \quad (1.33)$$

The second identity is a consequence of the property that two distinct matrix elements of Φ are uncorrelated.

Now let x be an n -tuple and choose the matrix elements of Φ independently from the normal distribution $N(0, 1/\sqrt{m})$ as just described. Then

$$E_{\Delta}(\|\Phi x\|_{\ell_2^m}^2) = E_{\Delta}(\langle \Phi_T^{\text{tr}} \Phi_T x, x \rangle) = \sum_{i=1}^m \sum_{j_1, j_2=1}^n x_{j_1} x_{j_2} E_{\Delta}(t_{i j_1} t_{i j_2}) = \|x\|_{\ell_2^n}^2$$

Thus, if Φ lies sufficiently close to its mean value then it will satisfy RIP. In this chapter we will show that for n sufficiently large with respect to m the probability that the random matrix Φ lies very close to the mean and satisfies RIP is near certainty. Such random matrices are relatively easy to construct. Even standard spreadsheets will do the job.

Another simple means of constructing random matrices that might satisfy RIP is through use of the distribution $\rho_1(t)$ on R :

$$\rho_1(t) = \begin{cases} \sqrt{\frac{m}{12}}, & -\sqrt{\frac{3}{m}} \leq t \leq \sqrt{\frac{3}{m}} \\ 0, & \text{otherwise.} \end{cases} \quad (1.34)$$

Here,

$$\bar{t} = E_{\rho_1}(t) = 0, \quad \sigma^2 = E_{\rho_1}(t^2) = \frac{1}{m}.$$

We choose $\Phi_{ij} = t_{ij}$ where $\mathbf{t} = (t_{ij})$ is a matrix random variable on R_{nm} with density function

$$\Delta_1(\mathbf{t}) = \begin{cases} \left(\frac{m}{12}\right)^{mn/2} & \text{if } -\sqrt{\frac{3}{m}} \leq t_{ij} \leq \sqrt{\frac{3}{m}} \text{ for all } i,j \\ 0, & \text{otherwise.} \end{cases} \quad (1.35)$$

Given any function $f(\mathbf{t})$ of the mn components of \mathbf{t} we define its expectation by

$$\bar{f}(\mathbf{t}) = E_{\rho_1}(f(\mathbf{t})) = \int_{R_{mn}} f(\mathbf{t}) \Delta_1(\mathbf{t}) \Pi_{ij} dt.$$

Again we find $E_{\rho_1}(\|\Phi x\|_2^2) = 1$ so these random matrices appear to be good candidates to satisfy RIP for large n . A related construction is where each column vector Φ_j of the sample matrix $\Phi = (\Phi_{ij})$ is chosen randomly as a vector on the unit m -sphere S_m : $\sum_{i=1}^m \Phi_{ij}^2 = \|\Phi_j\|_{\ell_2^m}^2 = 1$. Each unit column vector is chosen independently of the others. Here the probability density is just the area measure on the n -sphere. In this case

$$\begin{aligned} \|\Phi x\|_{\ell_2^m}^2 &= \sum_{j=1}^n \|x_j \Phi_j\|_{\ell_2^m}^2 = \sum_{j=1}^n x_j^2 \|\Phi_j\|_{\ell_2^m}^2 + 2 \sum_{1 \leq j < \ell \leq n} x_j x_\ell \langle \Phi_j, \Phi_\ell \rangle_{\ell_2^m} \\ &= \|x\|_{\ell_2^n}^2 + 2 \sum_{1 \leq j < \ell \leq n} x_j x_\ell \langle \Phi_j, \Phi_\ell \rangle_{\ell_2^m}. \end{aligned} \quad (1.36)$$

Since the unit vectors Φ_j and Φ_ℓ are chosen randomly on the unit sphere and independently if $j \neq \ell$, the inner product $\langle \Phi_j, \Phi_\ell \rangle_{\ell_2^m}$ will be a random variable with mean 0. Thus $E(\|\Phi x\|_{\ell_2^m}^2) = \|x\|_{\ell_2^n}^2$, and this is another good candidate to satisfy RIP.

A third method of construction is in terms of the discrete probability measure

$$\rho_2(t) = \begin{cases} \frac{1}{2}, & t = \pm 1 \\ 0 & \text{otherwise.} \end{cases} \quad (1.37)$$

We define the expectation of any polynomial $p(t)$ in t by

$$\bar{p}(t) = E_{\rho_2}(p(t)) \equiv \frac{1}{2} \sum_{t=\pm 1} p(t).$$

Here, $\bar{t} = 0$ and $\sigma^2 = E_{\rho_2}(t^2) = 1$. In this case we choose $\Phi_{ij} = t_{ij}/\sqrt{m}$ where each t_{ij} is obtained by a ‘‘coin flip’’. Given any polynomial function $p(\mathbf{t})$ of the mn components of $\mathbf{t} = \mathbf{t}_{ij}$ we define its expectation by

$$\bar{p}(\mathbf{t}) = E(p(\mathbf{t})) = \frac{1}{2^{mn}} \sum_{\mathbf{t}_{ij}=\pm 1} p(\mathbf{t}). \quad (1.38)$$

Again E is linear and we have the properties

$$E(1) = 1, \quad E(t_{i_1 j_1} t_{i_2 j_2}) = \delta_{i_1 i_2} \delta_{j_1 j_2}. \quad (1.39)$$

Further, $E_{\rho_2}(\|\Phi x\|_2^2) = 1$ so these random matrices are again good candidates to satisfy RIP for large n .

1.4.1 Covering numbers

In the preceding section we have exhibited three families of random matrices such that for any real n -tuple x the random variable $\|\Phi x\|_{\ell_2^m}^2$ has expected value $E(\|\Phi x\|_{\ell_2^m}^2) = \|x\|_{\ell_2^n}^2$. In order to prove RIP for matrices Φ so chosen we need to show that this random variable is concentrated about its expected value. For this we will make use of the concentration of measure inequalities

$$\Pr \left(\left| \|\Phi x\|_{\ell_2^m}^2 - \|x\|_{\ell_2^n}^2 \right| \geq \epsilon \|x\|_{\ell_2^n}^2 \right) \leq 2e^{-mc_0(\epsilon)}, \quad 0 < \epsilon < 1. \quad (1.40)$$

The left hand side of the inequality is the probability that $\left| \|\Phi x\|_{\ell_2^m}^2 - \|x\|_{\ell_2^n}^2 \right| \geq \epsilon \|x\|_{\ell_2^n}^2$, where the probability is taken over all $m \times n$ matrices Φ in one of the three families. Here $c_0(\epsilon) > 0$ depends only on ϵ . Relations (1.40) quantify the degree to which each of the probability distributions is concentrated about its expected value. In §?? we will derive (1.40) for each of the three families and give explicit expressions for $c_0(\epsilon)$. Here we assume these results and proceed with the verification of RIP.

Our proof, taken from [], makes use of an estimate for the number of closed balls of radius η that are needed to completely cover the unit ball B_1 centered at the origin in N dimensional Euclidean space. (The ball of radius r centered at the

origin is $B_r = \{x \in \ell_2^N : \|x\| \leq r\}$.) This geometrical result is particularly interesting because it has applications to many practical problems concerning high dimensional spaces, such as point clouds and learning theory, as well as compressive sampling.

Let S be a closed bounded subset of N dimensional Euclidean space. We define the covering number $\mathcal{N}(S, \eta)$ as the smallest number n of closed balls $D_1(\eta), \dots, D_n(\eta)$ of radius η that are needed to cover S : $S \subseteq \cup_{i=1}^n D_i(\eta)$. There is no explicit formula for this number, even for the unit ball $S = B_1$. However, it is relatively easy to derive an upper bound for $\mathcal{N}(B_1, \eta)$ and that is our aim.

We say that m points $x^{(1)}, \dots, x^{(m)}$ in S are η -distinguishable if $\|x^{(i)} - x^{(j)}\| > \eta$ for $i \neq j$. We define $\mathcal{M}(S, \eta)$ as the maximal number of η -distinguishable points in S . Since S is closed and bounded this maximum exists, although the points $x^{(i)}$ are not unique.

Lemma 3

$$\mathcal{M}(S, 2\eta) \leq \mathcal{N}(S, \eta) \leq \mathcal{M}(S, \eta).$$

PROOF: For the right hand inequality, note that if $m = \mathcal{M}(S, \eta)$ then there exist m points $x^{(1)}, \dots, x^{(m)}$ that are η -distinguishable in S . By the maximality of m it follows that any $x \in S$ satisfies $\|x - x^{(i)}\| \leq \eta$ for some i . Hence $S \subseteq \cup_{j=1}^m D_j(\eta)$ and $\mathcal{N}(S, \eta) \leq m$.

To prove the left hand inequality we use the pigeon hole principal. Let $M = \mathcal{M}(S, 2\eta)$. Then there exist M points $x^{(1)}, \dots, x^{(M)}$ in S that are 2η -distinguishable. If $\mathcal{N}(S, \eta) < M$ then S is covered by fewer than M balls of radius η . If so, at least two of the M distinct points $x^{(i)}, x^{(j)}$ must lie in the same ball $D_h(\eta)$ with center h . By the triangle inequality

$$\|x^{(i)} - x^{(j)}\| \leq \|x^{(i)} - h\| + \|h - x^{(j)}\| \leq \eta + \eta = 2\eta.$$

However, this is impossible since $x^{(i)}$ and $x^{(j)}$ are 2η -distinguishable. Hence $M \leq \mathcal{N}(S, \eta)$. Q.E.D.

Theorem 5 For the unit ball B_1 in N dimensional Euclidean space and any $0 < \eta \leq 1$, we have

$$(1/2\eta)^N \leq \mathcal{N}(B_1, \eta) \leq (3/\eta)^N.$$

PROOF: If $m = \mathcal{M}(B_1, \eta)$ there must exist m η -distinguishable points $x^{(1)}, \dots, x^{(m)}$ in B_1 . By the maximality of m it follows that any point $x \in B_1$ satisfies $\|x -$

$x^{(i)}\| \leq \eta$ for some i . Hence, if $D_j(\eta)$ is the ball of radius η centered at $x^{(j)}$, we have $S \subseteq \cup_{j=1}^m D_j(\eta)$ and $\sum_{j=1}^{k=1} V(D_j) \geq V(B_1)$ where $V(D)$ is the volume of the ball D . Since $V(D_j) = \eta^N V(B_1)$ we find $m\eta^N \geq 1$. This implies $(1/\eta)^N \leq \mathcal{M}(B_1, \eta)$.

On the other hand, we can construct balls $D_j(\eta/2)$ about each $x^{(j)}$. If $x \in D_j(\eta/2)$ then, since $x = (x - x^{(j)}) + x^{(j)}$, we have

$$\|x\| \leq \|x - x^{(j)}\| + \|x^{(j)}\| \leq \frac{\eta}{2} + 1 \leq \frac{3}{2}.$$

Thus each ball $D_j(\eta/2)$ is contained in the ball $B_{3/2}$. Further, since the m points $x^{(i)}$ are η -distinguishable, no two of these balls overlap. Thus, $\sum_{j=1}^{k=1} V(D_j) \leq V(B_{3/2})$ where $V(D)$ is the volume of the ball D . Since $V(D_j) = (\eta/2)^N V(B_1)$ and $V(B_{3/2}) = (3/2)^N V(B_1)$ we have $m(\eta/2)^N \leq (3/2)^N$. This implies $\mathcal{M}(B_1, \eta) \leq (3/\eta)^N$. The theorem is now an immediate consequence of Lemma 3. Q.E.D.

Lemma 4 *Let Φ be an $m \times n$ matrix whose matrix elements Φ_{ij} are drawn randomly and independently from a probability distribution that satisfies the concentration of measure inequality (1.40). Let T be an index set with $\#(T) = k < m$, and X_T be the set of all n -tuples x with index set T . Then for any $0 < \delta < 1$ and all $x \in X_T$ we have*

$$(1 - \delta)\|x\|_{\ell_2^m} \leq \|\Phi x\|_{\ell_2^m} \leq (1 + \delta)\|x\|_{\ell_2^m} \quad (1.41)$$

with probability at least $1 - 2(12/\delta)^k e^{-c_0(\delta/2)^m}$.

NOTE 1: Although this result appears to be RIP, it is not. First of all, the lemma applies only to k -sparse signals x with a specific index set T , not to all k -sparse signals. Secondly the inequalities (1.41) do not hold in general, but only with a guaranteed probability. The lemma says that the probability of failure of (1.41) to hold is bounded above by $2(12/\delta)^k e^{-c_0(\delta/2)^m}$. This will be a practical method for constructing sample matrices only if we can show that this bound is so close to 0 that failure virtually never occurs.

NOTE 2: The proof of the lemma and results to follow depend on a simple but basic result from probability theory, the union bound. We sketch the derivation. Suppose $p(A^{(1)})$ is the probability that inequalities (1.41) fail to hold for $x = x^{(1)}$, and $p(A^{(2)})$ is the probability of failure for $x = x^{(2)}$. Then

$$p(A^{(1)}) = p(A^{(1)} \cap A^{(2)c}) + p(A^{(1)} \cap A^{(2)}),$$

i.e., the probability of failure for $x^{(1)}$ is the probability of simultaneous failure for $x^{(1)}$ and success for $x^{(2)}$ plus the probability of simultaneous failure for both $x^{(1)}$ and $x^{(2)}$. Here, $0 \leq p(C) \leq 1$ for all these probabilities. Similarly we have the decomposition

$$p(A^{(2)}) = p(A^{(1)c} \cap A^{(2)}) + p(A^{(1)} \cap A^{(2)}).$$

The probability that there is failure for at least one of $x^{(1)}, x^{(2)}$ is denoted $p(A^{(1)} \cup A^{(2)})$ and it has the decomposition

$$p(A^{(1)} \cup A^{(2)}) = p(A^{(1)} \cap A^{(2)c}) + p(A^{(1)c} \cap A^{(2)}) + p(A^{(1)} \cap A^{(2)}),$$

i.e., the probability of simultaneous failure for $x^{(1)}$ and success for $x^{(2)}$, plus the probability of simultaneous success for $x^{(1)}$ and failure for $x^{(2)}$, plus the probability of simultaneous failure for both $x^{(1)}$ and $x^{(2)}$. Comparing these identities and using the fact that $p(A^{(1)} \cap A^{(2)}) \geq 0$ we obtain the inequality

$$p(A^{(1)} \cup A^{(2)}) \leq p(A^{(1)}) + p(A^{(2)}).$$

By a simple induction argument we can establish

Lemma 5 (*union bound*) $p(A^{(1)} \cup A^{(2)} \dots \cup A^{(h)}) \leq \sum_{i=1}^h p(A^{(i)})$.

Thus the probability that at least one of h events occurs is bounded above by the sum of the probabilities of occurrence of each of the events separately.

PROOF OF LEMMA 4: We can assume $\|x\|_{\ell_2^n} = 1$ since this can be achieved for any nonzero x by multiplying (1.41) by $1/\|x\|_{\ell_2^n}$. Now consider a finite set Q_T of n -tuples $q^{(1)}, \dots, q^{(K)}$, such that Q_T is contained in the unit ball in X_T , i.e., $q^{(i)} \in X_T$ and $\|q^{(i)}\|_{\ell_2^n} \leq 1$. We choose these vectors such that the unit ball in X_T is covered by the closed balls D_1, \dots, D_K where D_i is centered at $q^{(i)}$ and has radius $\delta/4$. Thus, if $x \in X_T$ with $\|x\|_{\ell_2^n} \leq 1$ then there is some point $q^{(j)} \in Q_T$ such that

$$\|x - q^{(j)}\|_{\ell_2^n} \leq \delta/4.$$

From Theorem 5, $\mathcal{N}(B_1, \delta/4) \leq (12/\delta)^k$, so we can require $\#(Q_T) \leq (12/\delta)^k$.

The concentration of measure inequality (1.40) can be written in the form

$$(1 - \epsilon)\|x\|_{\ell_2^n}^2 \leq \|\Phi x\|_{\ell_2^m}^2 \leq (1 + \epsilon)\|x\|_{\ell_2^n}^2, \quad 0 < \epsilon < 1, \quad (1.42)$$

with probability of failure bounded above by $2e^{-mc_0(\epsilon)}$. Now we set $\epsilon = \delta/2$ and use the union bound for $\#(Q_T)$ events to obtain

$$(1 - \delta/2)\|x\|_{\ell_2^m}^2 \leq \|\Phi x\|_{\ell_2^m}^2 \leq (1 + \delta/2)\|x\|_{\ell_2^m}^2, \text{ for all } q \in Q_T$$

with probability of failure bounded above by $2(12/\delta)^k e^{-mc_0(\delta/2)}$. We can take square roots on each side of the inequality and use the facts that $1 - \delta/2 \leq \sqrt{1 - \delta/2}$ and $\sqrt{1 + \delta/2} \leq (1 + \delta/2)$ for $0 < \delta/2 < 1$ to obtain

$$(1 - \delta/2)\|x\|_{\ell_2} \leq \|\Phi x\|_{\ell_2^m} \leq (1 + \delta/2)\|x\|_{\ell_2}, \text{ for all } q \in Q_T$$

with the same probability of failure as before.

We have verified (1.41) for $x \in Q_T$. We extend the result to any x in the unit sphere of X_T by using the fact that $\|x - q\|_{\ell_2} \leq \delta/4$ for some $q \in Q_T$. First we prove the right hand inequality. let A be the smallest number such that

$$\|\Phi x\|_{\ell_2^m} \leq (1 + A)\|x\|_{\ell_2^m} \tag{1.43}$$

for all x in the unit sphere of X_T . (Since this sphere is closed and bounded, A must exist and be finite.) Choosing any such x in the unit sphere, and approximating it by q as above, we find

$$\begin{aligned} \|\Phi x\|_{\ell_2^m} &\leq \|\Phi q\|_{\ell_2^m} + \|\Phi(x - q)\|_{\ell_2^m} \leq (1 + \delta/2)\|q\|_{\ell_2} + (1 + A)\|x - q\|_{\ell_2} \\ &\leq 1 + \delta/2 + (1 + A)\delta/4. \end{aligned}$$

Since A is the smallest number such that (1.43) holds we have $1 + A \leq 1 + \delta/2 + (1 + A)\delta/4$ or $A \leq (3\delta/4)/(1 - \delta/4) \leq \delta$. (This last inequality follows from a simple calculus argument and verifies the right hand side of (1.41).) The left hand side of (1.41) follows from

$$\|\Phi x\|_{\ell_2^m} \geq \|\Phi q\|_{\ell_2^m} - \|\Phi(x - q)\|_{\ell_2^m} \geq (1 - \delta/2) - (1 + \delta)\delta/4 \geq 1 - \delta.$$

Q.E.D.

Theorem 6 *Let Φ be an $m \times n$ matrix whose matrix elements Φ_{ij} are drawn randomly and independently from a probability distribution that satisfies the concentration of measure inequality (1.40). Let $0 < \delta < 1$. Then there exist constants $c_1, c_2 > 0$, depending only on δ , such that for all k -sparse signals x with $1 \leq k < m$ and $k \leq c_1 m / \ln(n/k)$ we have*

$$(1 - \delta)\|x\|_{\ell_2} \leq \|\Phi x\|_{\ell_2^m} \leq (1 + \delta)\|x\|_{\ell_2} \tag{1.44}$$

with probability of failure bounded above by e^{-mc_2} .

PROOF: From Lemma 4 we have established (1.44) for all k -sparse signals $x \in X_T$, with failure probability bounded by $2(12/\delta)^k e^{-mc_0(\delta/2)}$. We employ the union bound to establish the result for all k -sparse signals. There are

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \leq \left(\frac{en}{k}\right)^k$$

possible index sets T , where the last inequality follows from Stirling's formula []. Therefore, (1.44) will fail to hold for all k -sparse signals with probability bounded by

$$2\left(\frac{en}{k}\right)^k (12/\delta)^k e^{-mc_0(\delta/2)} = \exp(-mc_0(\delta/2) + k[\ln(en/k) + \ln(12/\delta)] + \ln 2).$$

To finish the proof we need to find a constant c_2 such that

$$0 < c_2 \leq c_0\left(\frac{\delta}{2}\right) - \frac{k}{m}[\ln\left(\frac{n}{k}\right) + 1 + \ln(12/\delta)] - \frac{\ln 2}{m}. \quad (1.45)$$

If we limit k to values such that $k \leq c_1 m / \ln(n/k)$ for some positive constant c_1 then (1.45) will hold if $0 < c_2 \leq c_0\left(\frac{\delta}{2}\right) - c_1\left(1 + \frac{(2+\ln(12/\delta))}{\ln(n/k)}\right)$. (Here we have used the bound $\ln 2 \leq k$.) By choosing the positive constant c_1 sufficiently small we can guarantee that $c_2 > 0$. Q.E.D.

1.4.2 Concentration of Measure Inequalities

The normal distribution

We proceed with the derivation of the concentration of measure inequality (1.40) for sample matrices chosen by means of the normal distribution $N(0, 1/\sqrt{m})$. Then $\Phi = \mathbf{t} = (t_{ij})$ is a matrix random variable on R_{nm} with distribution function

$$\Delta(\mathbf{t}) = \left(\frac{m}{2\pi}\right)^{mn/2} \exp\left(-\frac{m}{2} \sum_{i=1}^m \sum_{j=1}^n t_{ij}^2\right). \quad (1.46)$$

We chose an n -tuple x with $\|x\|_{\ell_2^n} = 1$ as signal and determine the probability distribution function for the random variable $\|\Phi x\|_{\ell_2^m}^2 = \sum_{i=1}^m (\sum_{j=1}^n t_{ij} x_j)^2$. First we determine the cumulative probability function for $\|\Phi x\|_{\ell_2^m}$, i.e.,

$$P(\tau) = \Pr(\|\Phi x\|_{\ell_2^m} \leq \tau) = \left(\frac{m}{2\pi}\right)^{mn/2} \int_{\|\Phi x\|_{\ell_2^m} \leq \tau} \exp\left(-\frac{m}{2} \sum_{i=1}^m \sum_{j=1}^n t_{ij}^2\right) d\mathbf{t} \quad (1.47)$$

where $\tau \geq 0$. This multiple integral appears difficult to evaluate because of the dependence on the vector x . However, there is rotational symmetry here so that $P(\tau)$ is the same for all unit vectors x . To see this, note that we can use the Gram-Schmidt process to construct an orthonormal basis $\{O^{(1)}, O^{(2)}, \dots, O^{(n)}\}$ for ℓ_2^n such that $O^{(1)} = x$. Then the $n \times n$ matrix O whose j -th row is $O^{(j)}$ will be orthogonal, i.e.,

$$O = \begin{pmatrix} O^{(1)} \\ O^{(2)} \\ \vdots \\ O^{(n)} \end{pmatrix} = (O_{ij}), \quad OO^{\text{tr}} = I$$

where I is the $n \times n$ identity matrix. Using well known properties of the determinant $\det(A)$ of the matrix A we have

$$1 = \det(I) = \det(OO^{\text{tr}}) = \det(O)\det(O^{\text{tr}}) = (\det(O))^2,$$

so $\det(O) = \pm 1$. Now we make the orthogonal change of variables,

$$T_{ij} = \sum_{k=1}^n t_{ik} O_{jk}, \quad i = 1, \dots, m, \quad j = 1, \dots, n,$$

so that, in particular, $\|\Phi x\|_{\ell_2^m}^2 = \sum_{i=1}^m T_{i1}^2$. Since O is orthogonal this coordinate transformation preserves the sum of squares

$$\sum_{i=1}^m \sum_{j=1}^n T_{ij}^2 = \sum_{i=1}^m \sum_{j=1}^n t_{ij}^2. \quad (1.48)$$

Exercise 17 *Verify the identity (1.48).*

Furthermore, the Jacobian of the coordinate transformation has determinant $[\det(O)]^m$, so its absolute value is 1. Thus via the standard rules for change of variables in multiple integrals we have

$$P(\tau) = \left(\frac{m}{2\pi}\right)^{mn/2} \int_{\sum_{i=1}^m T_{i1}^2 \leq \tau} \exp\left(-\frac{m}{2} \sum_{i=1}^m \sum_{j=1}^n T_{ij}^2\right) d\mathbf{T}.$$

Here the region of integration is $-\infty < T_{ij} < \infty$ for $j \neq 1$, so we can carry out $m(n-1)$ integrations immediately. For the remaining integrations we introduce spherical coordinates $T_{i1} = r\omega_i$ where $\sum_{i=1}^m \omega_i^2 = 1$, i.e. r is a radial coordinate and $(\omega_1, \dots, \omega_m)$ ranges over the unit sphere in m -space. Then

$d\mathbf{T} = r^{m-1} dr d\Omega$, where $d\Omega$ is the area measure on the sphere, and $\sum_{i=1}^m T_{i1}^2 = r^2$.

Thus

$$P(\tau) = c \int_0^{\sqrt{\tau}} e^{-mr^2/2} r^{m-1} dr = c' \int_0^{m\tau/2} e^{-R} R^{(m/2-1)} dR$$

where c' is a positive constant. The easiest way to compute c' is to note that $P(+\infty) = 1$, so $c' = 1/\Gamma(m/2)$ by the definition of the Gamma function. Now $P(\tau) = \int_0^\tau p(\alpha) d\alpha$ where $p(\tau)$ is the probability distribution function for $\|\Phi x\|_{\ell_2^m}^2$, so

$$p(\tau) = \frac{(m/2)^{m/2}}{\Gamma(m/2)} \tau^{m/2-1} e^{-m\tau/2}. \quad (1.49)$$

This is a famous distribution in probability theory, the Chi-square (or Gamma) distribution.

Exercise 18 Verify that $E(\|\Phi x\|_{\ell_2^m}^2) = \int_0^\infty \tau p(\tau) d\tau = 1$ by evaluating the integral explicitly.

We need to make use of a form of the Markov inequality, a simple but powerful result from probability theory.

Theorem 7 Let X be a nonnegative random variable with continuous probability distribution function $p(x)$ and let d be a positive constant. Then

$$\Pr(X \geq d) \leq \frac{1}{d} E(X). \quad (1.50)$$

PROOF:

$$\begin{aligned} E(X) &= \int_0^\infty xp(x) dx = \int_0^d xp(x) dx + \int_d^\infty xp(x) dx \\ &\geq \int_0^d xp(x) dx + d \int_d^\infty p(x) dx \geq d \int_d^\infty p(x) dx = d \Pr(X \geq d). \end{aligned}$$

Q.E.D.

Now we are ready to verify the concentration of measure inequalities for the probability distribution $p(\tau)$ of the random variable $X = \|\Phi x\|_{\ell_2^m}^2$ with $E(X) = 1$. For any $\epsilon > 0$ us first consider

$$\Pr(X - E(X) > \epsilon) = \int_{E(X)+\epsilon}^\infty p(\tau) d\tau.$$

We will use the idea that leads to the famous Chernoff inequality. Let $u > 0$ and note that

$$\begin{aligned}\Pr(X - E(X) > \epsilon) &= \Pr(X > E(X) + \epsilon) = \Pr(e^{uX} > e^{u(E(X)+\epsilon)}) \\ &= \Pr(e^{u(X-E(X)-\epsilon)} > 1).\end{aligned}$$

This is because the exponential function is one-one and order preserving. Thus, applying the Markov inequality to the random variable $Y = e^{u(X-E(X)-\epsilon)}$ with $d = 1$ and $E(X) = 1$ we have

$$\Pr(X - E(X) > \epsilon) \leq E(e^{u(X-E(X)-\epsilon)}) = \int_0^\infty e^{u(\tau-1-\epsilon)} p(\tau) d\tau = e^{-u(1+\epsilon)} E(e^{uX}),$$

for $u > 0$. Using the explicit formula (1.49) for $p(\tau)$ and assuming $0 < u < m/2$ we find

$$E(e^{uX}) = \frac{(m/2)^{m/2}}{(m/2 - u)^{m/2}}. \quad (1.51)$$

Thus

$$\Pr(X - 1 > \epsilon) \leq \frac{(m/2)^{m/2} e^{-u(1+\epsilon)}}{(m/2 - u)^{m/2}}$$

for all $0 < u < m/2$. This gives us a range of inequalities. The strongest is obtained by minimizing the right hand side in u . A first year calculus computation shows that the minimum occurs for $u = m\epsilon/2(1 + \epsilon)$. Thus

$$\Pr(X - 1 > \epsilon) \leq [e^{-\epsilon}(1 + \epsilon)]^{m/2} \leq e^{-m(\epsilon^2/4 - \epsilon^3/6)}. \quad (1.52)$$

Exercise 19 Verify the right hand inequality in (1.52) by showing that the maximum value of the function $f(\epsilon) = (1 + \epsilon) \exp(-\epsilon + \epsilon^2/2 - \epsilon^3/3)$ on the interval $0 \leq \epsilon$ is 1.

For the other inequality we reason in a similar manner.

$$\Pr(X - E(X) < -\epsilon) = \Pr(e^{u(E(X)-X-\epsilon)} > 1) < e^{u(1-\epsilon)} E(e^{-uX}),$$

for $u > 0$, so from (1.51),

$$\Pr(X - E(X) < -\epsilon) < e^{u(1-\epsilon)} \frac{(m/2)^{m/2}}{(m/2 + u)^{m/2}}.$$

For $0 < \epsilon < 1$ the minimum of the right hand side occurs at $u = m\epsilon/2(1 - \epsilon)$ and equals $[(1 - \epsilon)e^\epsilon]^{m/2}$, smaller than the right hand side of (1.52). Indeed

$$\Pr(X - 1 < -\epsilon) < [(1 - \epsilon)e^\epsilon]^{m/2} \leq e^{-m(\epsilon^2/4 - \epsilon^3/6)}. \quad (1.53)$$

We conclude from (1.52), (1.53) and the union bound that the concentration of measure inequality (1.40) holds with $c_0 = \epsilon^2/4 - \epsilon^3/6$.

Exercise 20 Verify the right hand inequality in (1.53) by showing that $g(\epsilon) = (1 - \epsilon)e^{\epsilon + \epsilon^2/2 - \epsilon^3/3} \leq 1$ for $0 < \epsilon < 1$.

The coin flip distribution

For the coin flip (Bernoulli) distribution we choose the sample matrix via $\Phi_{ij} = t_{ij}/\sqrt{m}$ where the mn random variables t_{ij} take the values ± 1 independently and with equal probability $1/2$. The expectation is now given by the sum (1.38) and has properties (1.39). If x is an n -tuple with $\|x\|_{\ell_2^n} = 1$ then

$$\|\Phi x\|_{\ell_2^m}^2 = \sum_{i=1}^m Q_i(x)^2, \quad Q_i(x) = \sum_j x_j \Phi_{ij}. \quad (1.54)$$

Note that each of the random variables $Q_i(x)$ has mean 0, i.e., $E(Q_i(x)) = 0$ and variance $E(Q_i^2(x)) = 1/m$. One consequence of this and (1.54) is $E(\|\Phi x\|^2) = E(1) = 1$. Since for any $\epsilon > 0$ we have $\Pr(\|\Phi x\|_{\ell_2^m}^2 - E(\|\Phi x\|^2) \geq \epsilon) = \Pr(\sum_{i=1}^m Q_i^2(x) - 1 \geq \epsilon)$ the concentration of measure inequalities for the Bernoulli distribution reduce to properties of sums of independent, identically distributed, random variables.

We follow the treatment in []. Just as with the normal distribution, we make use of the Markov inequality and introduce a parameter $u > 0$ that we can adjust to get an optimal inequality. Now

$$\begin{aligned} \Pr(\sum_{i=1}^m Q_i^2(x) - 1 \geq \epsilon) &= \Pr(e^{u\sum Q_i^2(x)} \geq e^{u(1+\epsilon)}) = \Pr(\prod_{i=1}^m e^{uQ_i^2(x)} \geq e^{u(1+\epsilon)}) \\ &= \Pr(e^{-u(1+\epsilon)} \prod_{i=1}^m e^{uQ_i^2(x)} \geq 1) \leq e^{-u(1+\epsilon)} \prod_{i=1}^m E(e^{uQ_i^2(x)}), \end{aligned}$$

where the last step is the Markov inequality. Finally we have

$$\Pr(\sum_{i=1}^m Q_i^2(x) - 1 \geq \epsilon) \leq e^{-u(1+\epsilon)} [E(e^{uQ_1^2(x)})]^m, \quad (1.55)$$

since the $Q_i(x)$ are identically distributed. similarly we have

$$\Pr(\sum_{i=1}^m Q_i^2(x) - 1 \leq -\epsilon) \leq e^{u(1-\epsilon)} [E(e^{-uQ_1^2(x)})]^m, \quad (1.56)$$

Returning to inequality (1.55) note that

$$E(e^{uQ_1^2(x)}) = E\left(\sum_{k=0}^{\infty} \frac{u^k Q_1(x)^{2k}}{k!}\right) = \sum_{k=0}^{\infty} \frac{u^k}{k!} E(Q_1(x)^{2k}), \quad (1.57)$$

where the interchange in order of summation and integration can be justified by the monotone convergence theorem of Lebesgue integration theory. Thus, to bound (1.55) it is sufficient to bound $E(Q_1(x)^{2k})$ for all k . However, in distinction to the case of the normal distribution, the bounds for $E(e^{uQ_1^2(x)})$ and $E(Q_1(x)^{2k})$ depend on x . Thus to obtain a concentration of measure inequality valid uniformly for all signals we need to find the “worst case”, i.e., to determine the vector $x = w$ on the unit n -sphere such that $E(e^{uQ_1^2(w)})$ is maximal. We will show that this worst case is achieved for $w = (1, 1, \dots, 1)/\sqrt{n}$.

We focus our attention on two components of the unit vector x . By relabeling we can assume they are the first two components $x_1 = a, x_2 = b$, so that $Q_1(x) = (at_{11} + bt_{12} + U)/\sqrt{m}$ where $U = \sum_{j=3}^n x_j t_{1j}$. Let $\hat{x} = (c, c, x_3, \dots, x_n)$ where $c = \sqrt{(a^2 + b^2)/2}$. Note that $\|\hat{x}\|_{\ell_2^n} = 1$, i.e., \hat{x} is also a unit vector.

Lemma 6 *For $k = 1, 2, \dots$ we have $E(Q_1(x)^{2k}) \leq E(Q_1(\hat{x})^{2k})$. Thus $E(e^{uQ_1^2(x)}) \leq E(e^{uQ_1^2(\hat{x})})$.*

PROOF: We fix U and average $Q_1(\hat{x})^{2k} - Q_1(x)^{2k}$ over the four possibilities $t_{11} = \pm 1, t_{12} = \pm 1$. This average is $S_k/4m^k$ where

$$S_k = (U+2c)^{2k} + 2U^{2k} + (U-2c)^{2k} - (U+a+b)^{2k} - (U+a-b)^{2k} - (U-a+b)^{2k} - (U-a-b)^{2k}.$$

We will show that $S_k \geq 0$. To see this we use the binomial theorem to expand each term of S_k except $2U^{2k}$ and regroup to obtain

$$S_k = 2U^{2k} + \sum_{i=0}^{2k} \binom{2k}{i} U^{2k-i} D_i,$$

$$D_i = (2c)^i + (-2c)^i - (a+b)^i - (a-b)^i - (-a+b)^i - (-a-b)^i.$$

If i is odd it is clear that $D_i = 0$. If $i = 2j$ is even then since $2c^2 = a^2 + b^2$ and $2(a^2 + b^2) = (a+b)^2 + (a-b)^2$ we have

$$D_{2j} = 2(2a^2 + 2b^2)^j - 2(a+b)^{2j} - 2(a-b)^{2j} = 2[(X+Y)^j - X^j - Y^j],$$

where $X = (a+b)^2, Y = (a-b)^2$. It is an immediate consequence of the binomial theorem that $(X + Y)^j \geq X^j + Y^j$ for $X, Y \geq 0$. Thus $D_{2j} \geq 0$ and

$$S_k = 2U^{2k} + \sum_{j=0}^k \binom{2k}{2j} U^{2(k-j)} D_{2j} \geq 0.$$

To compute the expectations $E(Q_1(x)^{2k}), E(Q_1(\hat{x})^{2k})$ we can first average over t_{11}, t_{12} as above and then average over U . This process will preserve the inequality $S_k \geq 0$ so we have $E(Q_1(x)^{2k}) \leq E(Q_1(\hat{x})^{2k})$. Q.E.D.

If x is any unit vector with two components that are not equal, say $x_1 = a, x_2 = b$ with $a \neq b$ then we can use Lemma 6 to obtain a new unit vector $\hat{x} = (c, c, x_3, \dots, x_n)$ such that $E(e^{uQ_1^2(x)}) \leq E(e^{uQ_1^2(\hat{x})})$. Proceeding in this way it will take at most $n-1$ steps to construct the unit vector $w = (C, \dots, C)$ where $C = 1/\sqrt{n}$ and $E(e^{uQ_1^2(x)}) \leq E(e^{uQ_1^2(w)})$. If all components of x are equal then $x = \pm w$ and $Q_1^2(\pm w) = Q_1^2(w)$. Thus the ‘‘worst case’’ is achieved for $x = w$.

For the worst case we have $Q_1(w) \equiv Q = (\tau_1 + \dots + \tau_n)/\sqrt{n}$ where the τ_j are independent random variables each taking the values $\pm 1/\sqrt{m}$ with probability $1/2$. Substituting this expression for Q_1 in (1.57) and using the multinomial expansion we can obtain an expression of the form

$$E(e^{uQ^2}) = \sum_{k_1, \dots, k_n} K_{k_1, \dots, k_n} u^{k_1 + \dots + k_n} E(\tau_1^{k_1} \dots \tau_n^{k_n})$$

where the k_j range over the nonnegative integers and $K_{k_1, \dots, k_n} \geq 0$. Moreover $E(\tau_1^{k_1} \dots \tau_n^{k_n}) = 0$ if any of the k_j are odd. Thus the only nonzero terms are of the form

$$E(\tau_1^{2\ell_1} \dots \tau_n^{2\ell_n}) = \prod_{j=1, \dots, n} E(\tau_j^{2\ell_j}) = \prod_{j=1, \dots, n} E(\tau^{2\ell_j})$$

where the ℓ_j range over the nonnegative integers and τ is a single random variable taking the values $\pm 1/\sqrt{m}$, each with probability $1/2$. This is because the τ_j are independently distributed. (Note that τ has mean 0 and standard deviation $1/\sqrt{m}$.) We conclude that the value of $E(e^{uQ^2})$ is uniquely determined by a power series in the values $E(\tau^{2k})$ for $k = 1, 2, \dots$, such that all terms are nonnegative:

$$E(e^{uQ^2}) = \sum_{\ell_1, \dots, \ell_n} K_{2\ell_1, \dots, 2\ell_n} u^{2(\ell_1 + \dots + \ell_n)} E(\tau^{2\ell_1}) \dots E(\tau^{2\ell_n}). \quad (1.58)$$

We will not attempt to evaluate this sum! What we need is an upper bound and we can get it by comparing this computation with corresponding computations for the normal distribution. There we had $X = \|\Phi x\|_{\ell_2^m}^2 = \sum_{i=1}^m Q_i^2(x)$ with $Q_i(x) =$

$\sum_{j=1}^n x_j t_{ij}$, where the t_{ij} are independently distributed normal random variables with distribution $N(0, 1/\sqrt{m})$. In that case we found, using spherical symmetry, that $E(e^{uQ_i(x)})$ was independent of the unit vector x . Thus we could choose $x = w$ without changing the result, and we could write $Q_i = Q = (t_1 + \dots + t_n)/\sqrt{n}$ where the t_j are independently distributed normal random variables each with distribution $N(0, 1/\sqrt{m})$. Then the expansion for $E(e^{uQ^2})$ would take exactly the form (1.58), with the same coefficients $K_{2\ell_1, \dots, 2\ell_n}$ and with $E(\tau^{2\ell}/m^\ell)$ replaced by $E(t^{2\ell})$. A straightforward computation for the normal distribution yields

$$E(t^{2k}) = \frac{(2k)!}{k!(2m)^k} \geq \frac{1}{m^k}, \quad (1.59)$$

whereas, since $\tau^2 = 1$ we have $E(\tau^{2k}/m^k) = 1/m^k$. Thus $E(\tau^{2k}/m^k) \leq E(t^{2k})$, so $[E(e^{uQ_1^2(x)})]^m$ from (1.55) is bounded by the corresponding expression (1.51) for the normal distribution

$$[E(e^{uQ_1^2(x)})]^m \leq \frac{(m/2)^{m/2}}{(m/2 - u)^{m/2}}. \quad (1.60)$$

Exercise 21 *Derive (1.59).*

A similar argument gives $E(Q_i^{2k}(x)) \leq (2k)!/k!(2m)^k$.

Using the techniques for the Chernoff inequality we achieve the same estimate as (1.52) for the normal distribution,

$$\Pr(\|\Phi x\|_{\ell_2^m}^2 - 1 > \epsilon) \leq [e^{-\epsilon}(1 + \epsilon)]^{m/2} \leq e^{-m(\epsilon^2/4 - \epsilon^3/6)}. \quad (1.61)$$

Rather than follow the exact same method for $\Pr(\|\Phi x\|_{\ell_2^m}^2 - 1 < -\epsilon)$, (1.56), we recast this inequality in the form

$$\Pr(\|\Phi x\|_{\ell_2^m}^2 - 1 < -\epsilon) \leq e^{u(1-\epsilon)} [E(e^{-uQ_1^2(x)})]^m \leq \quad (1.62)$$

$$e^{u(1-\epsilon)} [E(1 - uQ_1^2(x) + \frac{u^2}{2!}Q_1^4(x)/2!)]^m = e^{u(1-\epsilon)} [1 - uE(Q_1^2(x)) + \frac{u^2}{2}E(Q_1^4(x))]^m,$$

taking advantage of the fact that we are dealing with a convergent alternating series. Now always $E(Q_1^2(x)) = 1/m$ and $E(Q_i^{2k}(x)) \leq (2k)!/k!(2m)^k$, a bound for the right hand side of (1.62) is

$$\Pr(\|\Phi x\|_{\ell_2^m}^2 - 1 < -\epsilon) \leq e^{u(1-\epsilon)} [1 - \frac{u}{m} + \frac{3u^2}{2m^2}]^m, \quad 0 \leq u < m/2. \quad (1.63)$$

Just as in the derivation of (1.52) we make the substitution $u = m\epsilon/2(1 + \epsilon)$, not optimal in this case but good enough. The result, after some manipulation, is

$$\Pr(\|\Phi x\|_{\ell_2^m}^2 - 1 < -\epsilon) \leq e^{-m(\epsilon^2/4 - \epsilon^3/6)}. \quad (1.64)$$

It follows directly from (1.61), (1.64) and the union bound that the concentration of measure inequality (1.40) holds with $c_0 = \epsilon^2/4 - \epsilon^3/6$, the same as for the normal distribution case.

Exercise 22 Verify the right hand inequality in (1.64).

The constant distribution

A modification of the method applied to the Bernoulli distribution will also work for the constant distribution $\rho_1(t)$ on R :

$$\rho_1(t) = \begin{cases} \sqrt{\frac{m}{12}}, & -\sqrt{\frac{3}{m}} \leq t \leq \sqrt{\frac{3}{m}} \\ 0, & \text{otherwise.} \end{cases} \quad (1.65)$$

Here, the sample matrices are defined by $\Phi_{ij} = t_{ij}$ where $\mathbf{t} = (t_{ij})$ is a matrix random variable on R_{nm} with density function

$$\Delta_1(\mathbf{t}) = \begin{cases} \left(\frac{m}{12}\right)^{mn/2} & \text{if } -\sqrt{\frac{3}{m}} \leq t_{ij} \leq \sqrt{\frac{3}{m}} \text{ for all } i,j \\ 0, & \text{otherwise.} \end{cases} \quad (1.66)$$

A function $f(\mathbf{t})$ has expectation

$$\bar{f}(\mathbf{t}) = E_{\rho_1}(f(\mathbf{t})) = \int_{R_{nm}} f(\mathbf{t}) \Delta_1(\mathbf{t}) \Pi_{ij} d\mathbf{t}.$$

Here $Q_i(x) = \sum_{j=1}^n x_j t_{ij}$ and we take x to be a unit vector. Much of the derivation is word for word the same as in the Bernoulli case. Just as in that case the bounds for $E(e^{uQ_1^2(x)})$ and $E(Q_1(x)^{2k})$ depend on x . To obtain a concentration of measure inequality valid uniformly for all signals we need to find a “worst case” vector $x = w$ on the unit n -sphere such that $E(e^{uQ_1^2(w)})$ is maximal.

We focus our attention on two components of the unit vector x . By relabeling we can assume they are the first two components $x_1 = a, x_2 = b$, so that $Q_1(x) = (at_{11} + bt_{12} + U)$ where $U = \sum_{j=3}^n x_j t_{1j}$. Let $\hat{x} = (c, c, x_3, \dots, x_n)$ where $c = \sqrt{(a^2 + b^2)/2}$. Note that $\|\hat{x}\|_{\ell_2^n} = 1$, i.e., \hat{x} is also a unit vector.

Lemma 7 Suppose $ab < 0$. Then for $k = 1, 2, \dots$ we have $E(Q_1(x)^{2k}) \leq E(Q_1(\hat{x})^{2k})$, and $E(e^{uQ_1^2(x)}) \leq E(e^{uQ_1^2(\hat{x})})$ if $u > 0$.

PROOF: We fix U and average $Q_1(\hat{x})^{2k} - Q_1(x)^{2k}$ over the range $-\sqrt{3/m} \leq t_{11}, t_{12} \leq \sqrt{3/m}$, i.e. we integrate over this square with respect to the measure $12dt_{11}dt_{12}/m$. This average is $4S_k K^{2k}/(2k+1)(2k+2)$ where $K = \sqrt{3/m}$, $V = U/K$ and

$$S_k = \frac{1}{c^2} [(V + 2c)^{2k+2} - 2V^{2k+2} + (V - 2c)^{2k+2}] - \frac{1}{ab} [(V + a + b)^{2k+2} - (V + a - b)^{2k+2} - (V - a + b)^{2k+2} + (V - a - b)^{2k+2}].$$

We will show that $S_k \geq 0$. To see this we use the binomial theorem to expand each term of S_k except $2V^{2k}$ and regroup to obtain

$$S_k = \sum_{i=1}^{2k+2} \binom{2k+2}{i} V^{2k+2-i} D_i,$$

$$D_i = \frac{1}{c^2} [(2c)^i + (-2c)^i] - \frac{1}{ab} [(a+b)^i - (a-b)^i - (-a+b)^i + (-a-b)^i].$$

If i is odd it is clear that $D_i = 0$. If $i = 2j$ is even then since $2c^2 = a^2 + b^2$ and $2(a^2 + b^2) = (a+b)^2 + (a-b)^2$ we have

$$D_{2j} = 4(2a^2 + 2b^2)^{j-1} - \frac{1}{ab} [2(a+b)^{2j} - 2(a-b)^{2j}] = 2[(X+Y)^j - X^j - Y^j],$$

where $X = (a+b)^2, Y = (a-b)^2$. If $ab < 0$ then $-\frac{1}{ab} > 0$ and $(a+b)^2 > (a-b)^2$, so $D_{2j} > 0$ and

$$S_k = \sum_{j=1}^k \binom{2k+2}{2j} V^{2(k+1-j)} D_{2j} > 0.$$

To compute the expectations $E(Q_1(x)^{2k}), E(Q_1(\hat{x})^{2k})$ we can first average over t_{11}, t_{12} as above and then average over U . This process will preserve the inequality $S_k \geq 0$ so we have $E(Q_1(x)^{2k}) \leq E(Q_1(\hat{x})^{2k})$. Q.E.D.

Since $\|\Phi x\|^2 = \|\Phi(-x)\|^2$ we can always assume that the unit vector x has at least one component $x_\ell > 0$. The lemma shows that the ‘‘worst case’’ must occur for all $x_j \geq 0$. Now, assume the first two components are $x_1 = a > 0, x_2 = b > 0$, so $Q_1(x) = (at_{11} + bt_{12} + U)$ where $U = \sum_{j=3}^n x_j t_{1j}$. Then we can write $a = r \cos \theta, b = r \sin \theta$ where $r > 0$ and $0 < \theta < \pi/2$. we investigate the dependence of $E(Q_1^{2k}(x_\theta))$ on θ for fixed r . Here x_θ is a unit vector for all θ .

Lemma 8 Let $f_k(\theta) = E(Q_1^{2k}(x_\theta))$. Then

$$f'_k(\theta) \begin{cases} > 0 & \text{for } 0 < \theta < \pi/4 \\ = 0 & \text{for } \theta = \pi/4 \\ < 0 & \text{for } \pi/4 < \theta < \pi/2. \end{cases}$$

It follows from this result and the mean value theorem of calculus that the “worst case” is again $w = (1, 1, \dots, 1)/\sqrt{n}$.

PROOF: We fix U and average $Q_1(x_\theta)^{2k}$ over the range $-\sqrt{3/m} \leq t_{11}, t_{12} \leq \sqrt{3/m}$, with respect to the measure $12dt_{11}dt_{12}/m$. This average is $4S_k(rK)^{2k}/(2k+1)(2k+2)^2$ where $K = \sqrt{3/m}$, $V = U/rK$ and

$$S_k = \frac{1}{\sin \theta \cos \theta} [(V + \cos \theta + \sin \theta)^{2k+2} - (V + \cos \theta - \sin \theta)^{2k+2} - (V - \cos \theta + \sin \theta)^{2k+2} + (V - \cos \theta - \sin \theta)^{2k+2}].$$

We use the binomial theorem to expand each term of S_k and regroup to obtain

$$S_k = \sum_{i=1}^{2k+2} \binom{2k+2}{i} V^{2k+2-i} D_i,$$

$$D_i = \frac{1}{\sin \theta \cos \theta} [(\cos \theta + \sin \theta)^i - (\cos \theta - \sin \theta)^i - (-\cos \theta + \sin \theta)^i + (-\cos \theta - \sin \theta)^i].$$

If i is odd it is clear that $D_i = 0$. If $i = 2j$ is even we have $D_0 = 0$ and

$$D_{2j}(\theta) = \frac{2}{\cos \theta \sin \theta} [(\cos \theta + \sin \theta)^{2j} - (\cos \theta - \sin \theta)^{2j}], \quad j \geq 1,$$

where

$$S_k = \sum_{j=1}^k \binom{2k+2}{2j} V^{2(k+1-j)} D_{2j}(\theta). \quad (1.67)$$

We investigate the dependence of D_{2j} on θ by differentiating:

$$\begin{aligned} \frac{d}{d\theta} D_{2j}(\theta) &= (\cos^2 \theta - \sin^2 \theta) \left[(\cos \theta + \sin \theta)^{2j} \left(\frac{2j \cos \theta \sin \theta}{(\cos \theta + \sin \theta)^2} - 1 \right) \right. \\ &\quad \left. + (\cos \theta - \sin \theta)^{2j} \left(\frac{2j \cos \theta \sin \theta}{(\cos \theta - \sin \theta)^2} + 1 \right) \right] \\ &= (\cos^2 \theta - \sin^2 \theta) \sum_{\ell=0}^{j-1} \left[4j \binom{2j-2}{2\ell} - 2 \binom{2j}{2\ell+1} \right] \cos^{2\ell+1} \theta \sin^{2j-2\ell-1} \theta. \end{aligned}$$

Now

$$4j \binom{2j-2}{2\ell} - 2 \binom{2j}{2\ell+1} > 0, \quad \ell = 0, 1, \dots, j = 1, 2, \dots, \quad (1.68)$$

so

$$\frac{d}{d\theta} D_{2j}(\theta) \begin{cases} > 0 & \text{for } 0 < \theta < \pi/4 \\ = 0 & \text{for } \theta = \pi/4 \\ < 0 & \text{for } \pi/4 < \theta < \pi/2. \end{cases}$$

Now we average $4S_k(rK)^{2k}/(2k+1)(2k+2)^2$ over U to get $EQ_1(x_\theta)^{2k}$. From (1.67) we obtain the same result for $f'_k(\theta)$ as we got uniformly for each term $D'_{2j}(\theta)$. Q.E.D.

Exercise 23 Verify the inequality (1.68).

Exercise 24 Using Lemmas 7, 8 and the mean value theorem, show that for the constant probability distribution ρ_1 the “worst case” is $w = (1, \dots, 1)/\sqrt{n}$.

Now that we know that $w = (1, \dots, 1)/\sqrt{n}$ is the “worst case” we can parallel the treatment for the Bernoulli distribution virtually word for word to obtain the concentration of measure inequality for the constant probability distribution ρ_1 . To show that the “worst case” is dominated by the normal distribution we have to verify

$$E_{N(0,1/\sqrt{m})}(t^{2k}) = \frac{(2k)!}{k!(2m)^k} \geq E_{\rho_1}(t^{2k}) = \sqrt{\frac{m}{12}} \int_{-\sqrt{3/m}}^{\sqrt{3/m}} t^{2k} dt = \frac{3^k}{(2k+1)m^k}, \quad (1.69)$$

for all $k = 1, 2, \dots$, and this is straight forward.

Exercise 25 Verify the inequality (1.69).

Thus the concentration of measure inequality (1.40) again holds with $c_0 = \epsilon^2/4 - \epsilon^3/6$, the same as for the normal distribution and Bernoulli cases.

1.5 Discussion and practical implementation