Course Notes (Paris 2009)

Ronald DeVore

June 14, 2009

Abstract

1 Lecture 1-2: Foundations of Compressed Sensing

1.1 Sparsity

Let X be a Banach space. By a dictionary $\mathcal{D} \subset X$ we mean any set of norm one elements. Let us consider $\Sigma_k := \Sigma_k(\mathcal{D})$ the set of all $S \in X$ such that

$$S = \sum_{g \in \Lambda} c_g g, \quad \#(\Lambda) \le k \tag{1.1}$$

We say the elements in Σ_k are *k-sparse*. Notice that Σ_k is generally not a linear space: $\Sigma_k + \Sigma_k \neq \Sigma_k$.

In applications, we cannot expect our target function (image/signal/ solution to PDE) to be sparse so we consider how well we can approximate it by k-sparse elements. This is measured by

$$\sigma_k(f) := \sigma_k(f)_X := \inf_{S \in \Sigma_k} \|f - S\|_X, \tag{1.2}$$

which is called *the error of k term approximation* in $\mathbb{R}$. To measure how fast $\sigma_k(f)$ tends to zero we introduce the *primary approximation spaces* $\mathcal{A}^r = \mathcal{A}^r(\mathcal{D}, X)$, $r > 0$, which consists of all f such that

$$\|f\|_{\mathcal{A}^r} := \sup_{k \ge 1} k^r \sigma_k(f) < \infty. \tag{1.3}$$

In some cases, one can characterize the space $\mathcal{A}^r$. To get a feeling for results of this type let us consider the following setting where $\mathcal{D} = (\phi_j)_{j \ge 1}$ is an orthonormal basis for a real Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle$ and corresponding norm $\|\cdot\| := \|\cdot\|_{\mathcal{H}}$. Each $f \in \mathcal{H}$ has an expansion

$$f = \sum_{j=1}^{\infty} a_j \phi_j, \quad a_j := a_j(f) := \langle f, \phi_j \rangle. \tag{1.4}$$

We say a sequence (b_j) is in weak ℓ_q if

$$\|(b_j)\|_{w\ell_q}^q := \sup_{\epsilon > 0} \epsilon^q \#\{j : \ |b_j| > \epsilon\} < \infty. \tag{1.5}$$

An equivalent definition is that the sequence b_j^* of rearrangements of the absolute values of the b_j into non-increasing order satisfies

$$b_n^* \le Mn^{-1/q}, \quad n \ge 1, \tag{1.6}$$

with the smallest M being the norm in $w\ell_q$.

Theorem 1.1 *A function $f \in \mathcal{H}$ is in $\mathcal{A}^r$, $r > 0$, if and only if $(a_j(f)) \in w\ell_q$ with $1/q = r+1/2$ with equivalent norms: there exists constants c_1, c_2 such that*

$$c_1\|f\|_{\mathcal{A}^r} \le \|(a_j(f))\|w\ell_q \le c_2\|f\|_{\mathcal{A}^r}. \tag{1.7}$$

Proof: Suppose $(a_j(f)) \in w\ell_q$, then taking $\epsilon = M^{-1}2^{-j/q}$ with $M := \|(a_j)\|_{w\ell_q}$, we find from the definition of $w\ell_q$ that $\Lambda_j := \Lambda_j(f) := \{i : |a_i| > M2^{-j/q}\}$, has cardinality $\#(\Lambda_j) \le 2^j$ for each $j \in \mathbb{Z}$. Hence,

$$\sigma_{2^k}^2 \le \sum_{j>k} \sum_{i\in\Lambda_j\setminus\Lambda_{j-1}} |a_i|^2 \le \sum_{j>k} 2^j M^2 2^{-2(j-1)/q} \le 2M^2 \sum_{j\ge k} 2^{-2jr} \le \frac{2}{1-2^{-2r}} M^2 2^{-2kr},$$

from which the left inequality easily follows.

If $f \in \mathcal{A}^r$, then with (a_n^*) the decreasing rearrangement of $(|a_j|)$ we have

$$n[a_{2n}^*]^2 \le \sum_{n+1}^{2n} [a_k^*]^2 \le \sigma_n^2(f) \le \|f\|_{\mathcal{A}^r}^2 n^{-2r}.$$

A similar inequality holds for a_{2n+1}^* and we derive the right inequality in (1.7). □

Frequently, we want to measure approximation error in non-Hilbertian norms. In this case, one needs further properties of the basis relative to that norm. In classical settings such as for wavelets or Fourier decompositions, this is provided by Littlewood-Paley theory and square functions. For our purposes, it will be enough to consider sequence norms $\ell_p(\Lambda)$ where Λ is a finite or countably infinite set.

If we fix the $\ell_p = \ell_p(\Lambda)$ norm in which approximation error is to be measured, then for any $x \in \mathbb{R}^N$, we have for $q := (r + 1/p)^{-1}$,

$$c_0\|x\|_{w\ell_q} \le \|x\|_{\mathcal{A}^r} \le c_1 r^{-1/p}\|x\|_{w\ell_q}, \quad x \in \mathbb{R}^N, \tag{1.8}$$

for two absolute constants $c_0, c_1 > 0$. This is proved in a similar manner to Theorem 1.1 where the constants in these inequalities do not depend on N. Therefore, $x \in \mathcal{A}^r$ is equivalent to $x \in w\ell_q$ with equivalent norms.

Since the ℓ_q norm is larger than the weak ℓ_q norm, we can replace the weak ℓ_q norm by the ℓ_q norm in the right inequality of (1.8). However, the constant can be improved via a direct argument. Namely, if $1/q = r + 1/p$, then for any $x \in \ell_q$, $q < p$,

$$\sigma_k(x)_{\ell_p} \le \|x\|_{\ell_q} k^{-r}, \quad k = 1, 2, \ldots. \tag{1.9}$$

To prove this, take Λ' as the set of indices corresponding to the k largest entries in x. If ϵ is the size of the smallest entry in Λ', then $\epsilon \leq \|x\|_{w\ell_q} k^{-1/q} \leq \|x\|_{\ell_q} k^{-1/q}$ and therefore

$$\sigma_k(x)_{\ell_p}^p = \sum_{i \notin \Lambda'} |x_i|^p \leq \epsilon^{p-q} \sum_{i \notin \Lambda'} |x_i|^q \leq k^{-\frac{p-q}{q}} \|x\|_{\ell_q}^{p-q} \|x\|_{\ell_q}^q, \tag{1.10}$$

so that (1.9) follows.

From this, we see that if we consider the class $K = U(\ell_q^N)$, we have

$$\sigma_k(K)_{\ell_p} \leq k^{-r}, \tag{1.11}$$

with $r = 1/q - 1/p$. On the other hand, taking $x \in K$ such that $x_i = (2k)^{-1/q}$ for $2k$ indices and 0 otherwise, we find that

$$\sigma_k(x)_{\ell^p} = [k(2k)^{-p/q}]^{1/p} = 2^{-1/q} k^{-r}, \tag{1.12}$$

so that $\sigma_k(K)_{\ell_p}$ can be framed by

$$2^{-1/q} k^{-r} \leq \sigma_k(K)_{\ell_p} \leq k^{-r}. \tag{1.13}$$

2 Compressed sensing

The typical paradigm for obtaining a compressed version of a discrete signal represented by a vector $x \in \mathbb{R}^N$ is to choose an appropriate basis, compute the coefficients of x in this basis, and then retain only the k largest of these with $k < N$. If we are interested in a bit stream representation, we also need in addition to quantize these k coefficients.

Assuming, without loss of generality, that x already represents the coefficients of the signal in the appropriate basis, this means that we pick an approximation to x from Σ_k. The best performance that we can achieve by such an approximation process in some given norm $\|\cdot\|_X$ of interest is described by $\sigma_k(x)_X$.

The above compression scheme requires us to know all the entries in x. Compressed sensing asks whether we can obtain the same performance with less information about x. To formulate the problem, we are given a budget of n questions we can ask about x. These questions are required to take the form of asking for the values $\lambda_1(x), \ldots, \lambda_n(x)$ where the λ_j are fixed linear functionals. The information we gather about x can therefore by described by

$$y = \Phi x, \tag{2.1}$$

where Φ is an $n \times N$ matrix called the *encoder* and $y \in \mathbb{R}^n$ is the *information vector*. The rows of Φ are representations of the linear functionals λ_j, $j = 1, \ldots, n$.

To extract the information that y holds about x, we use a *decoder* Δ which is a mapping from $\mathbb{R}^n \to \mathbb{R}^N$. We emphasize that Δ is not required to be linear. Thus, $\Delta(y) = \Delta(\Phi x)$ is our approximation to x from the information we have retained. We shall denote by $\mathcal{A}_{n,N}$ the set of all encoder-decoder pairs (Φ, Δ) with Φ an $n \times N$ matrix.

The most common way of evaluating the performance of an encoding-decoding pair $(\Phi, \Delta) \in \mathcal{A}_{n,N}$ is to ask for the largest value of k such that the encoding-decoding is exact for all k-sparse vectors, i.e.

$$x \in \Sigma_k \Rightarrow \Delta(\Phi x) = x. \tag{2.2}$$

This has an easy solution [10]

Lemma 2.1 *If Φ is any $n \times N$ matrix and k is a positive integer, then the following are equivalent:*
(i) *There is a decoder Δ such that $\Delta(\Phi x) = x$, for all $x \in \Sigma_k$,*
(ii) $\Sigma_{2k} \cap \mathcal{N} = \{0\}$,
(iii) *For any set T with $\#T = 2k$, the matrix Φ_T has rank $2k$.*
(iv) *For any set T with $\#(T) = 2k$, the columns indexed by T are linearly independent.* (v) *The symmetric non-negative matrix $\Phi_T^t \Phi_T$ is invertible, i.e. positive definite.*

Proof: The equivalence of (ii-v) is linear algebra.

(i)⇒(ii): Suppose (i) holds and $x \in \Sigma_{2k} \cap \mathcal{N}$. We can write $x = x_0 - x_1$ where both $x_0, x_1 \in \Sigma_k$. Since $\Phi x_0 = \Phi x_1$, we have, by (i), that $x_0 = x_1$ and hence $x = x_0 - x_1 = 0$.

(ii)⇒(i): Given any $y \in \mathbb{R}^n$, we define $\Delta(y)$ to be any element in $\mathcal{F}(y)$ with smallest support. Now, if $x_1, x_2 \in \Sigma_k$ with $\Phi u = \Phi d$, then $x_1 - x_2 \in \mathcal{N} \cap \Sigma_{2k}$. From (ii), this means that $x_1 = x_2$. Hence, if $x \in \Sigma_k$ then $\Delta(\Phi x) = x$ as desired. □

It is easy to construct examples of matrices of size $n \times N$ with $n = 2k$ which satisfy the requirements of the Lemma. For example, if $0 < x_1 < \cdots < x_N = 1$, then the matrix $\Phi = (x_i^j)_{0 \le i \le n; 1 \le j \le N}$ works. Thus $2k$ measurements suffice to recover every k sparse vectors.

2.1 Instance optimality

We would like to measure the performance of a compressed sensing scheme (Δ, Φ) in a more robust way so that it includes all vectors x. Accordingly, we given the following definition:
We say that (Φ, Δ) is instance optimal in $\|\cdot\|_X$ of order k with constant C_0 if

$$\|x - \Delta(\Phi x)\|_X \le C_0 \sigma_k(x)_X, \tag{2.3}$$

holds for all $x \in \mathbb{R}^N$.

Notice that if we have instance optimality of order k for some norm then any k sparse vector x is captured exactly since $\sigma_k(x)_X = 0$. We shall see that the range of k for which instance optimality holds strongly depends on the norm X under consideration.

We have already seen in Lemma 2.1 that the performance of a matrix Φ in compressed sensing is determined by the null space

$$\mathcal{N} = \mathcal{N}(\Phi) := \{x \in \mathbb{R}^N : \Phi x = 0\}. \tag{2.4}$$

The importance of $\mathcal{N}$ is that if we observe $y = \Phi x$ without any a-priori information on x, the set of z such that $\Phi z = y$ is given by the affine space

$$\mathcal{F}(y) := x + \mathcal{N}. \tag{2.5}$$

The following result from [10] shows how the null space determines whether or not we have instance optimality.

Theorem 2.2 *Given an $n \times N$ matrix Φ, a norm $\|\cdot\|_X$ and a value of k, then a sufficient condition that there exists a decoder Δ such that (2.3) holds with constant C_0 is that*

$$\|\eta\|_X \leq \frac{C_0}{2}\sigma_{2k}(\eta)_X, \quad \eta \in \mathcal{N}. \tag{2.6}$$

A necessary condition is that

$$\|\eta\|_X \leq C_0\sigma_{2k}(\eta)_X, \quad \eta \in \mathcal{N}. \tag{2.7}$$

Proof: To prove the sufficiency of (2.6), we will define a decoder Δ for Φ as follows. Given any $y \in \mathbb{R}^N$, we consider the set $\mathcal{F}(y)$ and choose

$$\Delta(y) := \operatorname*{argmin}_{z \in \mathcal{F}(y)} \sigma_k(z)_X. \tag{2.8}$$

We shall prove that for all $x \in \mathbb{R}^N$

$$\|x - \Delta(\Phi x)\|_X \leq C_0\sigma_k(x)_X. \tag{2.9}$$

Indeed, $\eta := x - \Delta(\Phi x)$ is in $\mathcal{N}$ and hence by (2.6), we have

$$\begin{aligned} \|x - \Delta(\Phi x)\|_X &\leq (C_0/2)\sigma_{2k}(x - \Delta(\Phi x))_X \\ &\leq (C_0/2)(\sigma_k(x)_X + \sigma_k(\Delta(\Phi x)_X) \\ &\leq C_0\sigma_k(x)_X, \end{aligned}$$

where the second inequality uses the fact that $\sigma_{2k}(x+z)_X \leq \sigma_k(x)_X + \sigma_k(z)_X$ and the last inequality uses the fact that $\Delta(\Phi x)$ minimizes $\sigma_k(z)$ over $\mathcal{F}(y)$.

To prove the necessity of (2.7), let Δ be any decoder for which (2.3) holds. Let η be any element in $\mathcal{N} = \mathcal{N}(\Phi)$ and let η_0 be the best $2k$-term approximation of η in X. Let $\eta_0 = \eta_1 + \eta_2$ be any splitting of η_0 into two vectors of support size k, we can write

$$\eta = \eta_1 + \eta_2 + \eta_3, \tag{2.10}$$

with $\eta_3 = \eta - \eta_0$. Since $-\eta_1 \in \Sigma_k$ we have by (2.3) that $-\eta_1 = \Delta(\Phi(-\eta_1))$, but since $\eta \in \mathcal{N}$, we also have $-\Phi\eta = \Phi(\eta_2 + \eta_3)$ so that $-\eta_1 = \Delta(\Phi(\eta_2 + \eta_3))$. Using again (2.3) we derive

$$\begin{aligned} \|\eta\|_X &= \|\eta_2 + \eta_3 - \Delta(\Phi(\eta_2 + \eta_3))\|_X \leq C_0\sigma_k(\eta_2 + \eta_3) \\ &\leq C_0\|\eta_3\|_X = C_0\sigma_{2k}(\eta), \end{aligned}$$

which is (2.7). □

When X is an ℓ_p space, the best k term approximation is obtained by leaving the k largest components of x unchanged and setting all the others to 0. Therefore the property

$$\|\eta\|_X \leq C\sigma_k(\eta)_X, \tag{2.11}$$

can be reformulated by saying that

$$\|\eta\|_X \leq C\|\eta_{T^c}\|_X, \tag{2.12}$$

holds for all $T \subset \{1, \cdots, N\}$ such that $\#T \leq k$, where T^c is the complement set of T in $\{1, \cdots, N\}$. In going further, we shall say that Φ has the *null space property* in X of order k with constant C if (2.12) holds for all $\eta \in \mathcal{N}$ and $\#T \leq k$. Thus, we have

Corollary 2.3 *Suppose that X is an ℓ_p^N space, $k > 0$ an integer and Φ an encoding matrix. If Φ has the null space property (2.12) in X of order $2k$ with constant $C_0/2$, then there exists a decoder Δ so that (Φ, Δ) satisfies (2.3) with constant C_0. Conversely, the validity of (2.3) for some decoder Δ implies that Φ has the null space property (2.12) in X of order $2k$ with constant C_0.*

3 Gelfand widths: bounds for the range of k

Given a norm $\|\cdot\|_X$ in which we wish to measure error, we would like to know the largest range of k for which we can obtain instance optimality and then understand which schemes (Φ, Δ) achieve this range. We shall bound k by considering the performance of compressed sensing systems on compact sets K and showing this is related to certain well-known n widths.

Given K and X, we define

$$E_n(K)_X := \inf_{(\Phi,\Delta)\in\mathcal{A}_{n,N}} \sup_{x\in K} \|x - \Delta(\Phi x)\|_X, \tag{3.1}$$

which is a measure of the performance of the best compressed sensing systems on the set K.

We shall show that $E_n(K)_X$ is equivalent to the following Gelfand width:

$$d^n(K)_X := \inf_Y \sup\{\|x\|_X \ ; \ x \in K \cap Y\}, \quad n = 1, 2, \ldots, \tag{3.2}$$

where the infimum is taken over all subspaces Y of X of codimension less or equal to n.

Lemma 3.1 *Let $K \subset \mathbb{R}^N$ be any set for which $K = -K$ and for which there is a $C_0 > 0$ such that $K + K \subset C_0 K$. If $X \subset \mathbb{R}^N$ is any normed space, then*

$$d^n(K)_X \leq E_n(K)_X \leq C_0 d^n(K)_X, \quad 1 \leq n \leq N. \tag{3.3}$$

Proof: The proof will again bring out the role of the null space of Φ in the performance of Φ. Indeed, this null space $Y = \mathcal{N}$ of Φ is of codimension less or equal to n. Conversely, given any space $Y \subset \mathbb{R}^N$ of codimension n, we can associate its orthogonal complement $Y^\perp$ which is of dimension n and the $n \times N$ matrix Φ whose rows are formed by any basis for $Y^\perp$. Through this identification, we see that

$$d^n(K)_X = \inf_\Phi \sup\{\|\eta\|_X : \eta \in \mathcal{N} \cap K\}, \tag{3.4}$$

where the infimum is taken over all $n \times N$ matrices Φ.

Now, if (Φ, Δ) is any encoder-decoder pair and $z = \Delta(0)$, then for any $\eta \in \mathcal{N}$, we also have $-\eta \in \mathcal{N}$. It follows that either $\|\eta - z\|_X \geq \|\eta\|_X$ or $\| - \eta - z\|_X \geq \|\eta\|_X$. Since $K = -K$ we conclude that

$$d^n(K)_X \leq \sup_{\eta \in \mathcal{N} \cap K} \|\eta - \Delta(\Phi\eta))\|_X. \tag{3.5}$$

Taking an infimum over all encoder-decoder pairs in $\mathcal{A}_{n,N}$, we obtain the left inequality in (3.3).

To prove the right inequality, we choose an optimal Y for $d^n(K)_X$ and use the matrix Φ associated to Y (i.e., the rows of Φ are a basis for $Y^\perp$). We define a decoder Δ for Φ as follows. Given y in the range of Φ, we recall that $\mathcal{F}(y)$ is the set of x such that $\Phi x = y$. If $\mathcal{F}(y) \cap K \neq \emptyset$, we take any $\bar{x}(y) \in \mathcal{F}(y) \cap K$ and define $\Delta(y) := \bar{x}(y)$. When $\mathcal{F}(y) \cap K = \emptyset$, we define $\Delta(y)$ as any element from $\mathcal{F}(y)$. This gives

$$E_n(K)_X \leq \sup_{x,x' \in \mathcal{F}(y) \cap K} \|x - x'\|_X \leq \sup_{\eta \in C_0[K \cap \mathcal{N}]} \|\eta\|_X \leq C_0 d^n(K)_X, \tag{3.6}$$

where we have used the fact that $x - x' \in \mathcal{N}$ and $x - x' \in C_0 K$ by our assumptions on K. This proves the right inequality in (3.3). □

The orders of the Gelfand widths of ℓ_q balls in ℓ_p are known. Historically, the most famous of these results is the following

$$c_0 \min\left\{1, \sqrt{\frac{\log(N/n)}{n}}\right\} \leq d^n(U(\ell_1^N))_{\ell_2^N} = E_n(U(\ell_1^N)_{\ell_2^N} \leq c_1 \min\left\{1, \sqrt{\frac{\log(N/n)}{n}}\right\}. \tag{3.7}$$

The upper bound in (3.7) was first proved by Kashin [17] with a slightly worse power of the logarithm. The above form was given by Gluskin and Garneev [14]. These results could be thought of as the start of compressed sensing. We will have more to say on this in a moment. For now let us mention another result (which can be proved using the techniques in Chapter 13 of [19]). For any $0 < q < 1$,

$$c_0\Big[\min\left\{1, \frac{\log(N/n)}{n}\right\}\Big]^{1/q-1} \leq E_n(U(\ell_q^N)_{\ell_1^N} \leq c_1\Big[\min\left\{1, \frac{\log(N/n)}{n}\right\}\Big]^{1/q-1}. \tag{3.8}$$

Let us see how we can use this last result to give a bound on the optimal range of k for which instance optimality can hold. Suppose that we have an $n \times N$ matrix which gives ℓ_1^N instance optimality for some C_0 and k. For any vector in $U(\ell_q^N)$ we know from (1.13) that $\sigma_k(x) \leq \|x\|_{\ell_q^N} k^{-1/q-1}$. It follows that if we have instance optimality of order k for some sensing system of size $n \times N$, then $E_n(U(\ell_q^N)_{\ell_1^N} \leq C_0 k^{-1/q+1}$. Applying (3.8) gives

$$c_0\Big[\frac{\log(N/n)}{n}\Big]^{1/q-1} \leq E_n(U(\ell_q^N)_{\ell_1^N} \leq C_0 k^{-1/q+1}. \tag{3.9}$$

This means that $k \leq \frac{Cn}{\log(N/n)}$ with $C = (\frac{C_0}{c_0})^{1/q-1}$. Thus, this is the largest range of k for which we can have ℓ_1 instance optimality. Similar bounds can be established for instance optimality in other space $X = \ell_p^N$ and will be discussed shortly. For now we set out to see if we can find matrices that give instance optimality for this range of k.

4 Constructing good matrices

Now that we know the largest range of k possible in various settings of compressed sensing, we set out to see if we can construct matrices with this range of performance. All constructions of good CS matrices Φ are probabilistic.

We shall limit ourselves to random matrices of the following form (other possibilities can also be treated). We suppose that $\Phi = \Phi(\omega)$, $\omega \in \Omega$, is a family of random $n \times N$ matrices whose entries are given by independent realizations of a fixed symmetric random variable η defined on a probability space (Ω, ρ) with expectation $\mathbb{E}\eta = 0$ and variance $\mathbb{E}\eta^2 = 1/n$. The columns Φ_j, $j = 1, \ldots, N$, of Φ will be vectors in $\mathbb{R}^n$ with $\mathbb{E}\|\Phi_j\|_{\ell_2^n} = 1$.

We shall show that under rather mild conditions on η, the matrices $\Phi(\omega)$ will be good matrices with very high probability. Indeed, it will be enough to assume that $r := \sqrt{n}\eta$ is sub-Gaussian, i.e.

$$\Pr\{|\eta| > \delta\} \leq C_o e^{-c_0\delta^2}, \quad \delta > 0. \tag{4.1}$$

Two simple instances of random matrices which are often considered in compressed sensing are

(i) **Gaussian matrices:** $\Phi_{i,j} = \mathcal{N}(0, \frac{1}{n})$ are i.i.d. Gaussian variables of variance $1/n$.

(ii) **Bernoulli matrices:** $\Phi_{i,j} = \frac{\pm 1}{\sqrt{n}}$ are i.i.d. Bernoulli variables of variance $1/n$.

To understand the performance of the random matrices $\Phi(\omega)$ generated by such a choice η, we first examine the mapping properties of Φ. From the sub-Gaussian property one deduces:

Concentration of Measure Property (CMP) : *For any $x \in \mathbb{R}^N$ and any $0 < \delta < 1$, there is a set $\Omega_0(x, \delta)$ with*

$$\rho(\Omega_0(x, \delta)^c) \leq C_0 e^{-nc_0(\delta)}, \tag{4.2}$$

such that for each $\omega \in \Omega_0(x, \delta)$ we have

$$(1 - \delta)\|x\|_{\ell_2^N}^2 \leq \|\Phi(\omega)x\|_{\ell_2^n}^2 \leq (1 + \delta)\|x\|_{\ell_2^N}^2. \tag{4.3}$$

Lemma 4.1 *Let r be a zero mean random variable that satisfies* (4.1)*. Then, the $n \times N$ random family $\Phi(\omega)$, whose entries $\phi_{i,j}$ are independent realizations of $\eta = \frac{1}{\sqrt{n}}r$ satisfies the CMP for all n and N.*

Proof: For a not too difficult proof of this fact see [12]. □

For specific random variables such as Gaussian or Bernoulli random variables, there are several proofs in the literature of CMP. For example, it is proved in [1] that CMP holds with $c_0(\delta) = \delta^2/4 - \delta^3/6$ and $C_0 = 2$ for Bernoulli random variables.

There are several important consequences that can be drawn from the CMP. For us, the most important example is the Restricted Isometry Property (RIP) as introduced by Candés, Romberg, and Tao [5] which examines the mapping properties of Φ on Σ_k.

Restricted Isometry Property (RIP): *An $n \times N$ matrix A is said to have RIP of order k with constant δ if*

$$(1 - \delta)\|z\|_{\ell_2^N} \leq \|Az\|_{\ell_2^n} \leq (1 + \delta)\|z\|_{\ell_2^N}, \quad \forall z \in \Sigma_k. \tag{4.4}$$

We shall now show that random matrices with CMP will satisfy RIP for the large range of k.

Theorem 4.2 *Any random family of $n \times N$ matrices which satisfies CMP will automatically satisfy the RIP of order k and constant δ for any $k \leq c(\delta)n/\log(N/n)$ with probability $\geq 1-e^{-c_2 n}$ where c and c_2 depend only on δ.*

For the proof of this theorem we follow [3]. For any index set $T \subset \{1, \ldots, N\}$, let X_T be the linear space of all vectors in $\mathbb{R}^N$ which are supported on T.

Lemma 4.3 *Let $\Phi(\omega)$, $\omega \in \Omega$, satisfies* **CMP**. *Then, for any set T with $\#(T) = k < n$ and any $0 < \delta < 1$, we have*

$$(1-\delta)\|x\|_{\ell_2^N} \leq \|\Phi(\omega)x\|_{\ell_2^n} \leq (1+\delta)\|x\|_{\ell_2^N}, \quad \textit{for all} \quad x \in X_T, \tag{4.5}$$

with probability

$$\geq 1 - 2(12/\delta)^k \, e^{-c_0(\delta/2)n}. \tag{4.6}$$

Proof: First note that it is enough to prove (4.5) in the case $\|x\|_{\ell_2^N} = 1$, since Φ is linear. Next, we choose a finite set of points Q_T such that $Q_T \subseteq X_T$, $\|q\|_{\ell_2^N} \leq 1$ for all $q \in Q_T$, and for all $x \in X_T$ with $\|x\|_{\ell_2^N} \leq 1$ we have

$$\min_{q \in Q_T} \|x - q\|_{\ell_2^N} \leq \delta/4. \tag{4.7}$$

It is well known from covering numbers and easy to prove (see e.g. Chapter 13 of [19]) that we can choose such a set Q_T with $\#(Q_T) \leq (12/\delta)^k$. We next use **CMP** with $\delta/2$, with the result that, with probability exceeding the right side of (4.6), we have

$$(1-\delta/2)\|q\|_{\ell_2^N}^2 \leq \|\Phi q\|_{\ell_2^n}^2 \leq (1+\delta/2)\|q\|_{\ell_2^N}^2, \quad \text{for all } q \in Q_T, \tag{4.8}$$

which trivially gives us

$$(1-\delta/2)\|q\|_{\ell_2^N} \leq \|\Phi q\|_{\ell_2^n} \leq (1+\delta/2)\|q\|_{\ell_2^N}, \quad \text{for all } q \in Q_T. \tag{4.9}$$

We now define A as the smallest number such that

$$\|\Phi x\|_{\ell_2^n} \leq (1+A)\|x\|_{\ell_2^N}, \quad \text{for all } x \in X_T, \ \|x\|_{\ell_2^N} \leq 1. \tag{4.10}$$

Our goal is to show that $A \leq \delta$. For this, we recall that for any $x \in X_T$ with $\|x\|_{\ell_2^N} \leq 1$, we can pick a $q \in Q_T$ such that $\|x - q\|_{\ell_2^N} \leq \delta/4$. In this case we have

$$\|\Phi x\|_{\ell_2^n} \leq \|\Phi q\|_{\ell_2^n} + \|\Phi(x-q)\|_{\ell_2^n} \leq 1 + \delta/2 + (1+A)\delta/4. \tag{4.11}$$

Since by definition A is the smallest number for which (4.10) holds, we obtain $A \leq \delta/2 + (1 + A)\delta/4$. Therefore $A \leq \frac{3\delta/4}{1-\delta/4} \leq \delta$, as desired. We have proved the upper inequality in (4.5). The lower inequality follows from this since

$$\|\Phi x\|_{\ell_2^n} \geq \|\Phi q\|_{\ell_2^n} - \|\Phi(x-q)\|_{\ell_2^n} \geq 1 - \delta/2 - (1+\delta)\delta/4 \geq 1 - \delta, \tag{4.12}$$

which completes the proof. □

Proof of Theorem : We know that for each of the k dimensional spaces X_T, the matrix $\Phi(\omega)$ will fail to satisfy (4.5) with probability

$$\leq 2(12/\delta)^k e^{-c_0(\delta/2)n}. \tag{4.13}$$

There are $\binom{N}{k} \leq (eN/k)^k$ such subspaces. Hence, the RIP will fail to hold with probability

$$\leq 2(eN/k)^k(12/\delta)^k e^{-c_0(\delta/2)n} = e^{-c_0(\delta/2)n+k[\log(eN/k)+\log(12/\delta)]+\log(2)}. \tag{4.14}$$

Thus, for a fixed $c_1 > 0$, whenever $k \leq c_1 n/\log(N/k)$, we will have that the exponent in the exponential on the right side of (4.14) is $\leq -c_2 n$ provided that $c_2 > c_0(\delta/2) - c_1[1 + (1 + \log(12/\delta))/\log(N/k)]$. Hence, we can always choose $c_1 > 0$ sufficiently small to ensure that $c_2 > 0$. This proves the theorem. From the validity of the theorem for the range of $k \leq c_1 n/\log(N/k)$, one can easily deduce its validity for $k \leq c_1' n/[\log(N/n) + 1]$ for $c_1' > 0$ depending only on c_1. □

Remarks: The above theorem holds for any random family satisfying **CMP** not necessarily generated by draws of a single random variable η. For the matrices generate by a single random variable, we have shown that if $r := \sqrt{n}\eta$ is subgaussian then it has the **CMP**. Therefore, **SG→CMP → RIP**. Much more is known about RIP. Two papers to look at are Rudelson and Vershynin [21] which treats RIP for Fourier matrices where there are still fundamental open questions and Adamczak, Litvak, Pajor, Tomczack-Jaegermann [2] which shows that weaker assumptions than **SG** suffice for **RIP**

5 Verifying instance optimality

We have claimed that matrices which satisfy **CMP** are good matrices for compressed sensing. To illustrate this fact, we shall now show that they satisfy instance optimality in ℓ_1^N for the largest range of k. The following lemma is proved using the method of Candés and Tao[6].

Lemma 5.1 *Let $a = \ell/k$, $b = \ell'/k$ with $\ell, \ell' \geq k$ integers. If Φ is any matrix which satisfies the RIP of order $(a+b)k$ with $\delta = \delta_{(a+b)k} < 1$. Then Φ satisfies the null space property in ℓ_1 of order ak with constant $C_0 = 1 + \frac{\sqrt{a}(1+\delta)}{\sqrt{b}(1-\delta)}$.*

Proof: It is enough to prove (2.12) in the case when T is the set of indices of the largest ak entries of η. Let $T_0 = T$, T_1 denote the set of indices of the next bk largest entries of η, T_2 the next bk largest, and so on. The last set T_s defined this way may have less than bk elements.

We define $\eta_0 := \eta_{T_0} + \eta_{T_1}$. Since $\eta \in \mathcal{N}$, we have $\Phi\eta_0 = -\Phi(\eta_{T_2} + \ldots + \eta_{T_s})$, so that

$$\begin{aligned} \|\eta_T\|_{\ell_2} &\leq \|\eta_0\|_{\ell_2} \leq (1-\delta)^{-1}\|\Phi\eta_0\|_{\ell_2} = (1-\delta)^{-1}\|\Phi(\eta_{T_2} + \ldots + \eta_{T_s})\|_{\ell_2} \\ &\leq (1-\delta)^{-1}\sum_{j=2}^{s}\|\Phi\eta_{T_j}\|_{\ell_2} \leq (1+\delta)(1-\delta)^{-1}\sum_{j=2}^{s}\|\eta_{T_j}\|_{\ell_2}, \end{aligned}$$

where we have used the **RIP** repeatedly. Now for any $i \in T_{j+1}$ and $i' \in T_j$, we have $|\eta_i| \leq |\eta_{i'}|$ so that $|\eta_i| \leq (bk)^{-1}\|\eta_{T_j}\|_{\ell_1}$. It follows that

$$\|\eta_{T_{j+1}}\|_{\ell_2} \leq (bk)^{-1/2}\|\eta_{T_j}\|_{\ell_1}, \quad j = 1, 2, \ldots, s-1, \tag{5.1}$$

so that

$$\|\eta_T\|_{\ell_2} \leq (1+\delta)(1-\delta)^{-1}(bk)^{-1/2}\sum_{j=1}^{s-1}\|\eta_{T_j}\|_{\ell_1} \leq (1+\delta)(1-\delta)^{-1}(bk)^{-1/2}\|\eta_{T^c}\|_{\ell_1}. \tag{5.2}$$

By the Cauchy-Schwartz inequality $\|\eta_T\|_{\ell_1} \leq (ak)^{1/2}\|\eta_T\|_{\ell_2}$, and we therefore obtain

$$\|\eta\|_{\ell_1} = \|\eta_T\|_{\ell_1} + \|\eta_{T^c}\|_{\ell_1} \leq (1 + \frac{\sqrt{a}(1+\delta)}{\sqrt{b}(1-\delta)})\|\eta_{T^c}\|_{\ell_1} \tag{5.3}$$

which verifies the null space property with the constant C_0. □

Since we know the null space property is sufficient for instance optimality, we have proved the following.

Theorem 5.2 *Let Φ be any matrix which satisfies the RIP of order $3k$. Define the decoder Δ for Φ as in* (8.18) *for $X = \ell_1$. Then* (2.3) *holds in $X = \ell_1$ with constant $C_0 = 2(1 + \sqrt{2}\frac{1+\delta}{1-\delta})$.*

Remarks: Candés [4] has shown that $3k$ can be replaced by $2k$ in the above theorem. There are also some papers trying to understand the weakest assumption on δ.

Let us also note that the same arguments as given above give the following *mixed norm instance optimality*

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq Ck^{-1/2}\sigma_k(f)_{\ell_1^N}, \tag{5.4}$$

which holds for any matrix satisfying RIP of order $3k$ and an appropriate decoder Δ.

6 Instance optimality in ℓ_2

The reader may be curious as to why we concentrated on instance optimality in ℓ_1^N and not in the more natural space ℓ_2^N. The reason is that instance optimality fails miserably in ℓ_2^N. The reason for this is that any properly normalized $n \times N$ compressed sensing matrix Φ with $n << N$ will necessarily have large norm on ℓ_2^N. Here is one particular way to fetter this out [10].

Theorem 6.1 *Any $n \times N$ matrix Φ of which satisfies instance optimality with $k = 1$ necessarily has $N \leq C_0^2 n$.*

Proof: We know that a necessary and sufficient condition for instance optimality is the null space property. So for any vector η in the null space of Φ, we have

$$\|\eta\|_{\ell_2}^2 \leq C_0^2\|\eta_{T^c}\|_{\ell_2}^2, \quad \#T \leq 1, \tag{6.1}$$

or equivalently for all $j \in \{1, \cdots, N\}$,

$$\sum_{i=1}^{N}|\eta_i|^2 \leq C_0^2\sum_{i\neq j}|\eta_i|^2. \tag{6.2}$$

From this, we derive that for all $j \in \{1, \cdots, N\}$,

$$|\eta_j|^2 \leq (C_0^2 - 1) \sum_{i \neq j} |\eta_i|^2 = (C_0^2 - 1)(\|\eta\|_{\ell_2}^2 - |\eta_j|^2), \tag{6.3}$$

and therefore

$$|\eta_j|^2 \leq A \|\eta\|_{\ell_2}^2, \tag{6.4}$$

with $A = 1 - \frac{1}{C_0^2}$.

Let $(e_j)_{j=1,\cdots,N}$ be the canonical basis of $\mathbb{R}^N$ so that $\eta_j = \langle \eta, e_j \rangle$ and let $v_1, \ldots, v_{N-n}$ be an orthonormal basis for $\mathcal{N}$. Denoting by $P = P_{\mathcal{N}}$ the orthognal projection onto $\mathcal{N}$, we apply (6.4) to $\eta := P(e_j) \in \mathcal{N}$ and find that for any $j \in \{1, \ldots, N\}$

$$|\langle P(e_j), e_j \rangle|^2 \leq A. \tag{6.5}$$

This means

$$\sum_{i=1}^{N-n} |\langle e_j, v_i \rangle|^2 \leq A, \quad j = 1, \ldots, N. \tag{6.6}$$

We sum (6.6) over $j \in \{1, \ldots, N\}$ and find

$$N - n = \sum_{i=1}^{N-n} \|v_i\|_{\ell_2}^2 \leq AN. \tag{6.7}$$

It follows that $(1 - A)N \leq n$. That is, $N \leq nC_0^2$ as desired. □

7 Instance optimality in probability

While it is disturbing that instance optimality does not hold in ℓ_2^N, the situation is not so bleak if we rethink what we are doing. To obtain instance optimality for the large range of k for ℓ_1, we need to use probabilistic constructions since there are no known deterministic constructions. On the other hand, even if we had one of the favorable random matrices we would not be able to verify it since the RIP property cannot be checked in any reasonable computational time. Hence ultimately we are in a situation where we draw a matrix at random and know only that it will work with high probability. Then why not evaluate performance in this probabilistic setting as well?

So let us embed ourselves into the following setting. We let Ω be a probability space with probability measure ρ and let $\Phi = \Phi(\omega)$, $\omega \in \Omega$ be an $n \times N$ random matrix. To keep matters simple, let us assume that the entries of Φ are generated by independent draws of a random variable as we have previously considered. We seek results of the following type:

Instance Optimality in Probability: *for any $x \in \mathbb{R}^N$, if we draw Φ at random with respect to P, then*

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \sigma_k(x)_{\ell_2} \tag{7.1}$$

holds for this particular x with high probability for some decoder Δ (dependent on the draw Φ).

It should be understood that Φ is drawn independently for each x in contrast to building a Φ such that (7.1) holds simultaneously for all $x \in \mathbb{R}^N$ which was our original definition of instance optimality.

We now describe our process for decoding $y = \Phi x$, when $\Phi = \Phi(\omega)$ is our given realization of the random matrix. (This method is numerically impractical but will be sufficient for theoretical results. Later we shall turn to more practical decoders.) Let $T \subset \{1, \ldots, N\}$ be any subset of column indices with $\#(T) = k$ and let X_T be the linear subspace of $\mathbb{R}^N$ which consists of all vectors supported on T. For this T, we define

$$x_T^* := \operatorname*{argmin}_{z \in X_T} \|\Phi z - y\|_{\ell_2}. \tag{7.2}$$

In other words, x_T^* is chosen as the least squares minimizer of the residual in approximation by elements of X_T. Notice that x_T^* is supported on T. If Φ satisfies RIP of order k then the matrix $\Phi_T^t \Phi_T$ is nonsingular and the nonzero entries of x_T^* are given by

$$(\Phi_T^t \Phi_T)^{-1} \Phi_T^t y. \tag{7.3}$$

To decode y, we search over all subsets T of cardinality k and choose

$$T^* := \operatorname*{argmin}_{\#(T)=k} \|y - \Phi x_T^*\|_{\ell_2^n}. \tag{7.4}$$

Our decoding of y is now given by

$$x^* = \Delta(y) := x_{T^*}^*. \tag{7.5}$$

Theorem 7.1 [10] *Assume that Φ is a random matrix which satisfies RIP of order $2k$ and also satisfies CMP each with probability $1 - \epsilon$. Then, for each $x \in \mathbb{R}^N$, the estimate* (7.1) *holds with $C_0 = 1 + \frac{2C}{1-\delta}$ and probability $1 - 2\epsilon$.*

Proof: Let $x \in \mathbb{R}^N$ be arbitrary and let $\Phi = \Phi(\omega)$ be the draw of the matrix Φ from the random ensemble. We denote by T the set of indices corresponding to the k largest entries of x. Thus

$$\|x - x_T\|_{\ell_2} = \sigma_k(x)_{\ell_2}. \tag{7.6}$$

Then,

$$\|x - x^*\|_{\ell_2} \leq \|x - x_T\|_{\ell_2} + \|x_T - x^*\|_{\ell_2} \leq \sigma_k(x)_{\ell_2} + \|x_T - x^*\|_{\ell_2}. \tag{7.7}$$

We bound the second term by

$$\begin{aligned}
\|x_T - x^*\|_{\ell_T^N} &\leq (1-\delta)^{-1} \|\Phi(x_T - x^*)\|_{\ell_2} \\
&\leq (1-\delta)^{-1} (\|\Phi(x - x_T)\|_{\ell_2} + \|\Phi(x - x^*)\|_{\ell_2}) \\
&= (1-\delta)^{-1} (\|y - \Phi x_T\|_{\ell_2} + \|y - \Phi x^*\|_{\ell_2}) \\
&\leq 2(1-\delta)^{-1} \|y - \Phi x_T T\|_{\ell_2} = 2(1-\delta)^{-1} \|\Phi(x - x_T)\|_{\ell_2} \\
&\leq 2C(1-\delta)^{-1} \|x - x_T\|_{\ell_2} = 2C(1-\delta)^{-1} \sigma_k(x)_{\ell_2}.
\end{aligned}$$

where the first inequality uses the RIP and the fact that $x_T - x^*$ is a vector with support of size less than $2k$, the third inequality uses the minimality of T^* and the fourth inequality uses the boundedness property in probability for $x - x_T$. □

8 Decoding

Up to this point we have completely ignored the practicality of the decoders used in our compressed sensing results. We shall now remedy this situation. The two most common decoders are constructed by ℓ_1 minimization and greedy algorithms. Both of these are reasonable to implement numerically. We shall only have time to discuss ℓ_1 minimization but there are now nice results for greedy decoders (see [20], [9]). We concentrate on how this decoder performs in terms of instance optimality in ℓ_1^N and instance optimality with high probability in ℓ_2^N?

The decoder for ℓ_1 minimization is

$$\Delta(y) := \operatorname*{argmin}_{\Phi z = y} \|z\|_{\ell_1}, \quad y \in \mathbb{R}^n. \tag{8.1}$$

It can be implemented numerically with linear programming using the simplex algorithm or interior point methods. The fact that ℓ_1-minimization is a good decoder was one of the main contributions of Donoho [13] and Candés, Romberg, and Tao [5, 7] and their results were the beginning of the subject of compressed sensing as it is now called. The following theorem is contained in [10] but can also be derived from the techniques in [5].

Theorem 8.1 *Let Φ be any matrix which satisfies the RIP of order $3k$ with $\delta_{3k} \leq \delta < (\sqrt{2}-1)^2/3$. Define the decoder Δ for Φ as in (8.2). Then, (Φ, Δ) satisfies (2.3) in $X = \ell_1$ with $C_0 = \frac{2\sqrt{2}+2-(2\sqrt{2}-2)\delta}{\sqrt{2}-1-(\sqrt{2}+1)\delta}$.*

Remark: *Again, Candés [4] shows that $3k$ can be replaced by $2k$ with a somewhat more involved argument.*

Proof: We apply Lemma 5.1 with $a = 1$, $b = 2$ to see that Φ satisfies the null space property in ℓ_1 of order k with constant $C = 1 + \frac{1+\delta}{\sqrt{2}(1-\delta)} < 2$. This means that for any $\eta \in \mathcal{N}$ and T such that $\#T \leq k$, we have

$$\|\eta\|_{\ell_1} \leq C\|\eta_{T^c}\|_{\ell_1}, \tag{8.2}$$

and therefore

$$\|\eta_T\|_{\ell_1} \leq (C-1)\|\eta_{T^c}\|_{\ell_1}. \tag{8.3}$$

Let $x^* = \Delta(\Phi x)$ be the solution of (8.1) so that $\eta = x^* - x \in \mathcal{N}$ and

$$\|x^*\|_{\ell_1} \leq \|x\|_{\ell_1}. \tag{8.4}$$

Denoting by T the set of indices of the largest k coefficients of x, we can write

$$\|x_T^*\|_{\ell_1} + \|x_{T^c}^*\|_{\ell_1} \leq \|x_T\|_{\ell_1} + \|x_{T^c}\|_{\ell_1}. \tag{8.5}$$

It follows that

$$\|x_T\|_{\ell_1} - \|\eta_T\|_{\ell_1} + \|\eta_{T^c}\|_{\ell_1} - \|x_{T^c}\|_{\ell_1} \leq \|x_T\|_{\ell_1} + \|x_{T^c}\|_{\ell_1}, \tag{8.6}$$

and therefore

$$\|\eta_{T^c}\|_{\ell_1} \leq \|\eta_T\|_{\ell_1} + 2\|x_{T^c}\|_{\ell_1} = \|\eta_T\|_{\ell_1} + 2\sigma_k(x)_{\ell_1}. \tag{8.7}$$

Using (8.3) and the fact that $C < 2$ we thus obtain

$$\|\eta_{T^c}\|_{\ell_1} \leq \frac{2}{2-C}\sigma_k(x)_{\ell_1}. \tag{8.8}$$

We finally use again (8.2) to conclude that

$$\|x - x^*\|_{\ell_1} \leq \frac{2C}{2-C}\sigma_k(x)_{\ell_1}, \tag{8.9}$$

which is the announced result. □

Our next goal is to show that the ℓ_1 minimization decoder can be used together with general random matrices to give instance optimality in probability for the large range of k. To establish this fact we need another mapping property of random matrices.

Lemma 8.2 *Let $\Phi(\omega)$ be an $n \times N$ random matrix which satisfies CMP. For each $x \in \mathbb{R}^N$ there is a set $\Omega_1(x)$ with*

$$\rho(\Omega_1(x)^c) \leq Ce^{\frac{-n}{2L}} \tag{8.10}$$

such that for all $\omega \in \Omega_1(x)$,

$$\|\Phi x\|_{\ell_\infty^n} \leq \frac{1}{\sqrt{L}}\|x\|_{\ell_2^N}, \quad \text{where } L := \log N/n. \tag{8.11}$$

Proof: We shall prove this lemma in the case that $\eta = \frac{1}{\sqrt{n}}r$ where r is the Bernoulli random variable taking values ± 1. In the general **SG** case, one has to analyze moments (see [12]. Without loss of generality we can assume that $\|x\|_{\ell_2^N} = 1$. Fix such an x. We note that each entry y_i of y is of the form

$$y_i = \frac{1}{\sqrt{n}}\sum_{j=1}^{N} x_j r_{i,j}, \tag{8.12}$$

where the $r_{i,j}$ are independent random variables and $x = (x_1, \ldots, x_N)$. We shall use Hoeffding's inequality (see page 596 of [15]) which says that for independent mean zero random variables ϵ_j taking values in $[a_j, b_j]$, $j = 1, \ldots, N$, we have

$$\Pr\left(|\sum_{j=1}^{N}\epsilon_j| \geq \delta\right) \leq 2e^{\frac{-2\delta^2}{\sum_{j=1}^{N}(b_j-a_j)^2}}. \tag{8.13}$$

We apply this to the random variables $\epsilon_j := \frac{1}{\sqrt{n}}x_j r_{i,j}$, $j = 1, \ldots, N$, which take values in $\frac{1}{\sqrt{n}}[-x_j, x_j]$. Since $\sum_{j=1}^{N}(2x_j)^2 = 4$, we deduce that

$$\Pr\left(|y_i| \geq \delta\right) \leq 2e^{\frac{-n\delta^2}{2}}. \tag{8.14}$$

Applying a union bound, we get

$$\Pr\left(\|y\|_{\ell_\infty^n} \geq \delta\right) \leq 2ne^{\frac{-n\delta^2}{2}}. \tag{8.15}$$

If we now take $\delta = 1/\sqrt{L}$ we arrive at the lemma. □

There is one additional mapping property of random matrices which is instrumental in showing that ℓ_1 minimization can be used as a decoder and attain instance optimality in probability.

Clipped Ball Mapping Property (CBMP): *Let $\Phi(\omega)$ be a random family of $n \times N$ matrices whose entries are given by random draws of the random variable $\eta = \frac{1}{\sqrt{n}} r$ with r a* **SG** *random variable. Let $L := \log(N/n)$ as before. Then, with probability $\geq 1 - Ce^{-c\sqrt{nN}}$ on the draw of Φ the following holds: for each vector $y \in \mathbb{R}^n$ with $\|y\|_{\ell_2^n}, L^{-1/2}\|y\|_{\ell_\infty^n} \leq 1$, there is a $z \in \mathbb{R}^N$ such that $\Phi(z) = y$ and $\|z\|_{\ell_1^N} \leq C'\sqrt{\frac{n}{L}}$. In other words, with high probability the unit ball in ℓ_1^N is mapped onto a clipped ball in $\mathbb{R}^n$.*

Remark: Using arguments similar to the proof of ℓ_1 instance optimality we can also require that the vector z in **CBMP** satisfies $\|z\|_{\ell_2^N} \leq C\|y\|_{\ell_2^n}$

This mapping property was proved by A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann [18] and reproved in [12]. We now use this mapping property to prove ℓ_2 instance optimality in probability.

Theorem 8.3 *Let $\Phi(\omega)$ be a random family of $n \times N$ matrices whose entries are given by random draws of the random variable $\eta = \frac{1}{\sqrt{n}} r$ with r a* **SG** *random variable and let Δ be the ℓ_1-minimization decoder. Let $L := \log(N/n)$ as before. For each $x \in \mathbb{R}^N$ and each $k \leq \tilde{a}n/\log(N/n)$, $N \geq [\ln 6]^2 n$, there is a set $\Omega(x,k)$ with*

$$\rho(\Omega(x,k)^c) \leq C[e^{-\tilde{c}_1 n} + e^{-\sqrt{Nn}} + e^{-n/24} + ne^{\frac{-n}{2\log(N/n)}}], \tag{8.16}$$

such that for each $\omega \in \Omega(x,k)$, we have

$$\|x - \Delta(\Phi x)\|_{\ell_2^N} \leq C'\sigma_k(x)_{\ell_2^N}, \tag{8.17}$$

where C and C' are absolute constants.

Proof: We will prove the theorem for the largest k satisfying $k \leq \tilde{a}n/L$. The theorem follows for all other k from the monotonicity of σ_k. Let x_k be a best approximation to x from Σ_k, so $\|x - x_k\|_{\ell_2^N} = \sigma_k(x)_{\ell_2^N} =: \sigma_k(x)$, and let $y' = \Phi(x - x_k)$. From **CMP** and Lemma 8.2, we have with high probability

$$\|y'\|_{\ell_2^n} \leq \sqrt{\frac{3}{2}}\|x - x_k\|_{\ell_2^N} = \sqrt{\frac{3}{2}}\sigma_k(x),$$

and

$$\|y'\|_{\ell_\infty^n} \leq \frac{1}{\sqrt{L}}\|x - x_k\|_{\ell_2^N} = \frac{1}{\sqrt{L}}\sigma_k(x).$$

The **CBMP** and the Remark following its definition says that there is a vector $z' \in \mathbb{R}^N$, such that $\Phi(x - x_k) = y' = \Phi z'$ and

$$\|z'\|_{\ell_2^N} \leq C\sigma_k(x), \quad \text{and } \|z'\|_{\ell_1^N} \leq C\sqrt{\frac{n}{L}}\sigma_k(x). \tag{8.18}$$

Note that $\sigma_k(x_k + z')_{\ell_1^N} \leq \|z'\|_{\ell_1^N}$, and therefore using (8.18) it follows that

$$\sigma_k(x_k + z')_{\ell_1^N} \leq C\sqrt{\frac{n}{L}}\sigma_k(x). \tag{8.19}$$

Since $\Phi x = \Phi(x_k + z')$, we have that $\bar{x} := \Delta(\Phi(x_k + z')) = \Delta(\Phi x)$. We know that with high probability Φ satisfies RIP of order $2k$ and hence the mixed-norm instance optimality (5.4). This means that

$$\|x_k + z' - \bar{x}\|_{\ell_2^N} \leq \frac{C}{\sqrt{k}} \sigma_k(x_k + z')_{\ell_1^N} \leq C' \sigma_k(x).$$

where the last inequality uses the definition of k. Therefore, it follows from (8.18) that

$$\begin{aligned} \|x - \bar{x}\|_{\ell_2^N} &\leq \|x - x_k - z'\|_{\ell_2^N} + \|x_k + z' - \bar{x}\|_{\ell_2^N} \\ &\leq \|x - x_k\|_{\ell_2^N} + \|z'\|_{\ell_2^N} + \|x_k + z' - \bar{x}\|_{\ell_2^N} \\ &\leq C\sigma_k(x), \end{aligned} \tag{8.20}$$

which proves the theorem. □

References

[1] D. Achilioptas, Database-friendly random projections, Proc. ACM Symposium on the Principles of Database Systems, pp. 274-281, 2001

[2] R. Adamczak, A.E. Litvak, A. Pajor, N. Tomczak-Jaegermann, *Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling*, preprint

[3] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, A simple proof of the restricted isometry property for random matrices, Constructive Approximation, to appear.

[4] E. Candés, *The restricted isometry property and its implications for compressed sensing*, Compte Rendus de l'Academie des Sciences, Paris, Series I, **3**46(2008), 589–592.

[5] E. J. Candès, J. Romberg and T. Tao, Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information, *IEEE Trans. Inf. Theory*, **5**2(2006), 489–509.

[6] E. Candès, J. Romberg, and T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Comm. Pure and Appl. Math.*, **5**9(2006), 1207–1223.

[7] E. Candès and T. Tao, Decoding by linear programming, *IEEE Trans. Inf. Theory* **5**1(2005), 4203–4215.

[8] E. Candès and T. Tao, Near optimal signal recovery from random projections:universal encoding strategies, *IEEE Trans. Inf. Theory* **5**2(2006), 5406–5425.

[9] A. Cohen, W. Dahmen, and R. DeVore, Near optimal approximation of arbitrary vectors from highly incomplete measurements, preprint.

[10] A. Cohen, W. Dahmen, and R. DeVore, *Compressed sensing and best k-term approximation*, JAMS, **22**(2009), 211–231.

[11] R. DeVore, Nonlinear approximation, *Acta Numer.* **7** (1998), 51–150.

[12] R. DeVore, G. Petrova, and P. Wojtaczszyk, *Instance-optimality in Probability with an ℓ_1-Minimization Decoder*, ACHA (to appear).

[13] D. Donoho, Compressed Sensing, EEE Trans. Information Theory, **52** (2006), 1289-1306.

[14] A.Garnaev, E.D. Gluskin, The widths of a Euclidean ball, Doklady AN SSSR, 277 (1984), 1048-1052.

[15] Györfy, L., M. Kohler, A. Krzyzak, A. and H. Walk (2002) *A distribution-free theory of nonparametric regression*, Springer Verlag, Berlin.

[16] E.D. Gluskin, Norms of random matrices and widths of finite-dimensional sets, Math. USSR Sbornik, 48(1984), 173–182.

[17] B. Kashin, The widths of certain finite dimensional sets and classes of smooth functions, *Izvestia* **41**(1977), 334–351.

[18] A. Litvak, A. Pajor, M. Rudelson, N. Tomczak-Jaegermann, *Smallest singular value of random matrices and geometry of random polytopes*, Advances in Math. **1**95 (2005), 491–523.

[19] G.G. Lorentz, M. von Golitschek and Yu. Makovoz, *Constructive Approximation:Advanced Problems*, Springer Grundlehren, vol. 304, Springer Berlin Heidelberg, 1996.

[20] D. Needel and J. Tropp, *CoSaMP: Iterative signal recovery from incomplete and inaccurate samples*, preprint 2008.

[21] M. Rudelson and R. Vershynin, *On sparse reconstruction from Gaussian and Fourier measurements*, CPAM **61**(2008), 1025–1045.

[22] A. Pajor and N. Tomczak-Jaegermann, *Subspaces of small codimension of finite dimensional Banach spaces*, Proc. Amer. Math. Soc., vol. 97, 1986, pp. 637–642.

Course Notes (Paris 2009)

Ronald DeVore

June 22, 2009

Abstract

1 Lecture 3: Capturing Functions in High Dimensions

1.1 Classifying High Dimensional Functions:

Our last two lectures will study the problem of approximating (or capturing through querries) a function f defined on $\Omega \subset \mathbb{R}^N$ with N very large. The usual way of classifying functions is by smoothness. The more derivatives a function has the nicer it is and the more efficiently it can be numerically approximated. However, as we move into high space dimension, this type of classification will suffer from the so-called *curse of dimensionality* which we shall now quantify.

2 Widths and Entropy of Classes

We can quantify the curse of dimensionality through concepts like entropy and widths.

Kolmogorov Entropy: Let us first consider entropy. Suppose that K is a compact set in the Banach space X with norm $\|\cdot\|_X$. If $\epsilon > 0$, we consider all possible coverings of $K \subset \cup_{i=1}^m B(x_i, \epsilon)$ using balls $B(x_i, \epsilon)$ of radius ϵ with centers $x_i \in X$. The smallest number $m = N_\epsilon(K)_X$ is called the covering number of K. The Kolmogorov entropy of K is then defined as

$$H_\epsilon(K)_X := \log_2(N_\epsilon(K))_X. \tag{2.1}$$

The Kolmogorov entropy measures the size or massivity of K. It has another important property of determining optimal encoding of the elements of K. Namely, if $x \in K$ then we can assign to x the binary bits of an index i for which $x \in B(x_i, \epsilon)$. Each x is then encoded to accuracy ϵ with $\leq \lceil H_\epsilon(K)_X \rceil$ bits and no other encoder can do better.

It is frequently more convenient to consider the entropy numbers

$$\epsilon_n(K)_X := \inf\{\epsilon : H_\epsilon(K)_X \leq n\}. \tag{2.2}$$

Typically, $\epsilon_n(K)_X$ decay like n^{-r} for standard compact sets. Not only does $\epsilon_n(K)_X$ tell us the minimal distortion we can achieve with n bit encoding, it also indicates that any numerical

algorithm which computes an approximation to each of the elements of K to accuracy $\epsilon_n(K)_X$ will require at least n operations.

Widths: There are several types of widths. The most prominent of these is the Kolmogorov width which measures how well K can be approximated through linear spaces of fixed dimension n. However, for us, the following definition of manifold width [1] will be more useful since it also governs nonlinear processes. We consider two continuous functions. The first function a maps each element $x \in K$ into $\mathbb{R}^n$ and the second function M maps $\mathbb{R}^n$ into the set $\mathcal{M}$ (which we view as an n-dimensional manifold although we make no assumptions about the smoothness of the image $\mathcal{M}$). The manifold width of the compact set K is then defined by

$$\delta_n(K)_X := \inf_{M,a} \sup_{x \in K} \|x - M(a(x))\|_X. \tag{2.3}$$

For typical compact sets K of functions, the manifold widths behave like the entropy numbers. For example, if K is the unit ball of any Besov or Sobolev space of smoothness s which compactly embeds into $L_p(\Omega)$ with $\Omega \subset \mathbb{R}^N$, then (see [2])

$$C_0 n^{-s/N} \le \delta_n(K)_{L_p(\Omega)} \le C_1 n^{-s/N}. \tag{2.4}$$

We see in (2.4) the curse of dimensionality. In order to obtain just moderate rates of convergence with $n \to \infty$ we need s to be comparable with N.

3 Model classes for functions in high dimension

We are interested in what are reasonable models for function classes $\mathcal{F}$ in $\mathbb{R}^N$ when N is large. The ultimate test is whether a proposed model class is commensurate with applications. However, we also need this class to be amenable to computation which in light of the previous section means that its manifold width or Kolmogorov entropy should tend to zero like n^{-r} for not too small of values of r. Of course this can be achieved by assuming the elements in $\mathcal{F}$ are very smooth but this may not be reasonable from the viewpoint of the specified application. Another approach is to suppose that althought the elements in $\mathcal{F}$ depend on N variables there are relatively few variables which dominate. Our goal here is to introduce and consider model classes that quantify this idea.

We shall begin by considering the following model classes. We assume that Φ is a compact set of continuous functions $\phi : \Omega \to \Omega_0$ with $\Omega \subset \mathbb{R}^N$ compact and $\Omega_0 \subset \mathbb{R}^k$ also compact. Here k is fixed and should be viewed as much smaller than N. Let K be a compact set of functions in $C(\Omega_0)$ (which is typically the unit ball of some smoothness norm $\|\cdot\|_{W^s}$). We consider the class $\mathcal{F} := \mathcal{F}(K, \Phi) := \{f = g(\phi) : \ \phi \in \Phi, g \in K\}$ which will be compact in $C(\Omega)$. Notice that $\mathcal{F}$ combines the notions of smoothness and variable reduction. In the case that K corresponds to a smoothness class for nonlinear approximation (like certain Besov classes) then $\mathcal{F}$ is also incorporating sparsity.

4 Recovering functions of few variables in high dimension

We begin by considering the special case of the above model classes where $K := U(C^s)$ and Φ is the set of coordinate projections onto a k-dimensional coordinate space. This is part of an ongoing study with Przemek Wojtaczszyk and Guergana Petrova. Each $f \in \mathcal{F}(K, \Phi)$ is of the form

$$f(x_1, \ldots, x_N) = g(x_{i_1}, \ldots, x_{i_k}), \quad |g|_{C^s} \leq 1 \tag{4.1}$$

where $C^s = C^s([0,1]^k)$ is the set of functions which have s continuous deriviatives, equipped with its usual semi-norm

$$|f|_{C^s(\Omega)} := \max_{|\nu| \leq s} \|D^\nu f\|_{C(\Omega)}.$$

We can allow s to be non-integer by working with Lipshcitz spaces. Notice that $\mathcal{F}$ is not a linear space.

The first problem we wish to consider is the following. Suppose we are given a budget of n points for which we can ask the values of f. Where should we choose these points in order to best recover f in the norm of $C(\Omega)$ and what is the best error we can receive. This is a special case of what is called *optimal recovery*. It is also sometimes called *directed learning*. There are two settings to consider. The first *adaptively* asks the questions. The j-th querry can depend on the answer to the first $j-1$ querries. The second one sets once and for all the n questions (non-adaptive). We shall consider both settings in what follows.

4.1 One variable of dependence

It will be instructive to consider first the case when f depends only on one coordinate variable where the arguments and proofs are most transparent. That is, we suppose $f(x_1, \ldots, x_N) = g(x_j)$ with both g and j unknown to us. We take the domains $\Omega = [0,1]^N$ and $\Omega_0 := [0,1]$ for simplicity.

We shall describe first a nonadaptive set of points where we will ask for the values of f. This set will be the union of two point sets. The first of these sets is $\mathcal{P} := \{P_i := m^{-1}(i, i, \ldots, i)\}_{i=0}^m\}$ which we call the set of *base points*. The important property of this set is that when its points are projected onto any of the coordinate axes, we get a set of $m+1$ equally spaced points with spacing $h := 1/m$.

The second set of points we shall need are *padding points*. These points will be used to find the coordinate j. Padding points are associated to a pair of points $P, P' \in \mathcal{P}$ and are constructed as follows. Every integer $j \in \{1, ..., N\}$ has a unique binary representation $j = 1 + \sum_{k=0}^{L} b_k(j) 2^k$ where $L := \lceil \log_2(N) \rceil - 1$ and each bit $b_k(j) \in \{0, 1\}$. Given a pair of points $P, P' \in \mathcal{P}$ and a value of $k \in \{0, 1, ..., L\}$, we define the point $[P, P']_k$ whose j-th coordinate is $P(j)$ (i.e. the same as the j-th coordinate of P) if $b_k(j) = 0$ and otherwise this coordinate of $[P, P']_k$ is defined to be the same as the j-th coordinate of P'. Thus any of these padding points corresponds to incrementing about half of the coordinate values from P to P'.

We ask for the values of f at the base points in $\mathcal{P}$ given above. We also ask for the values of f at the following set $\mathcal{Q}$ of padding points. To each pair P_{i-1}, P_i, $i = 1, \ldots, m$, of consecutive

points, we associate the padding points $[P_{i-1}, P_i]_k$, $k = 0, \ldots, L$. Thus the collection of padding points is $\mathcal{Q} := \{[P_{i-1}, P_i]_k,\ i = 1, \ldots m,\ k = 0, \ldots, L\}$. Clearly there are $m(L+1)$ points in $\mathcal{Q}$.

After receiving the values of f at the base points $\mathcal{P}$ (which are the values $g(i/m)$, $i = 0, \ldots, m$, of the univariate function g) , we can construct a function $\hat{g} = A_m(g)$ which approximates g to accuracy h^s, $h := 1/m$ (for example using piecewise polynomials). Namely,

$$\|g - \hat{g}\|_{C[0,1]} \leq C_s |g|_{C^s} h^s. \tag{4.2}$$

We can also assume that when $g = c$ is constant on $\mathcal{P}$ then $A_m(g) = c$. The function $\hat{g}(x_j)$ would provide a good approximation to f if we knew j.

Notice that if f is constant on $\mathcal{P}$ then we do not need to know j. If f is not constant on $\mathcal{P}$ then we shall use the padding points to find j. There is an i such that $f(P_{i-1}) \neq f(P_i)$ with $1 \leq i \leq m$. The value $f([P_{i-1}, P_i]_k)$ is either $f(P_{i-1})$ or $f(P_i)$. If it is $f(P_{i-1})$, then we know that $b_k(j) = 0$; if it is $f(P_i)$ then we know $b_k(j) = 1$. Thus from the answer to these questions, we know all the bits of j and hence we know j. We define our approximation to f to be $\hat{f}(x_1, \ldots, x_N) := \hat{g}(x_j)$.

Theorem 4.1 *If $f(x_1, \ldots, x_N) = g(x_j)$ with $g \in C^s$, then the function $\hat{f}$ defined above satisfies*

$$\|f - \hat{f}\|_{C(\Omega)} \leq C_s |g|_{C^s} h^s. \tag{4.3}$$

Let us note that Algorithm 1 asks for the values of f at $m + 1 + m(L+1)$ points. The logarithm in L is the price we pay for not knowing the change coordinate j.

Gain of Adaptivity: *If we are willing to use an adaptive algorithm we can save on the number of querries as follows. From the base points we determine an i for which $f(P_i) \neq f(P_{i+1})$. Then we only need the padding points for the pair P_i, P_{i+1}. Hence, we need a total of $m+1+L+1$ points in all. This matches the entropy and manifold width for $\mathcal{F}$ and hence the above result is optimal.*

We can also construct an algorithm which will handle the case where f is not necessarily a function of just one variable but is approximated by such a function. Given the univariate function g, we define the multivariate functions $G_\nu(x_1, \ldots, x_N) := g(x_\nu)$ for any $\nu = 1, \ldots, N$. Our new model for f is that there is a $j \in \{1, \ldots, N\}$ and an $\epsilon > 0$ such that

$$\|f - G_j\|_{C(\Omega)} \leq \epsilon. \tag{4.4}$$

We do not know g, j or ϵ.

Let us describe the adaptive version of an algorithm for this model class. We use the values of f at the points $P_i \in \mathcal{P}$, $i = 0, \ldots, m$ to determine $\hat{g}$ as before. Now to find a change coordinate j from this information, we choose a pair $(P_i, P_{i'})$, $i < i'$, for which $|f(P_i) - f(P_{i'})|$ is the largest among all such pairs. To identify the changing coordinate, we proceed as follows. We consider the value $f(Q_k)$ at each of the points $Q_k := [P_i, P_{i'}]_k$, $k = 0, \ldots, L$. If this value is closest to $f(P_i)$ we assign the bit $b_k = 0$. If this value is closest to $f(P_{i'})$ or if there is a tie, we assign the bit $b_k = 1$. These bits determine an integer $j \in \{1, \ldots, N\}$. We define $\hat{f}(x_1, \ldots, x_N) = \hat{g}(x_j)$.

Theorem 4.2 *Suppose that f is a function of N variables for which there is a function $g \in C^s$ and a $\nu \in \{1, \ldots, N\}$ such that*

$$\|f - G_\nu\|_{C(\Omega)} \leq \epsilon. \tag{4.5}$$

Then the function $\hat{f}$ defined above satisfies

$$\|f - \hat{f}\|_{C(\Omega)} \leq C_s[\epsilon + |g|_{C^s} h^s]. \tag{4.6}$$

Proof: We consider two cases.

Case 1: We assume in this case that the maximum deviation in the values $f(P_i)$, $i = 0, \ldots, m$, is $> 4\epsilon$. By the definition of i, i' we have that $|f(P_i) - f(P_{i'})| > 4\epsilon$. At each of the padding points $Q_k := [P_i, P_{i'}]_k$, $k = 0, \ldots, L$, we have $|f(Q_k) - G_\nu(Q_k)| \leq \epsilon$. Now if $b_k(\nu) = 0$ then $G_\nu(Q_k) = G_\nu(P_i)$ (since G_ν is a function only of the ν-th variable) and therefore $f(Q_k)$ is within 2ϵ of $f(P_i)$ but further than 2ϵ from $f(P_{i'})$. This means that the bit b_k assigned by the algorithm will be zero and therefore $b_k = b_k(\nu)$. The same conclusion holds if $b_k(\nu) = 1$. Hence the value j determined by the algorithm is equal to ν. We therefore obtain

$$\|f - \hat{f}\|_{C(\Omega)} \leq \|f - G_\nu\|_{C(\Omega)} + \|g - \hat{g}\|_{C([0,1])} \leq \epsilon + C_s |g|_{C^s} h^s. \tag{4.7}$$

which is the desired inequality.

Case 2: In this case, the maximum deviation of f over the points P_i, $i = 0, \ldots, m$, is $\leq 4\epsilon$. Hence the maximum deviation of g over the points $h\mathcal{L}_1 = \{0, 1/m, \ldots, 1\}$ is at most 6ϵ. We consider the function $\tilde{g} = g - c$ where c is the median value of g on $h\mathcal{L}_1$. Then $|\tilde{g}| \leq 3\epsilon$ on $h\mathcal{L}_1$. Using the fact that $\tilde{g}$ is in C^s, it follows that $\|\tilde{g}\|_{C([0,1])} \leq C_s(\epsilon + |g|_{C^s} h^s)$ and further that $\|G_\nu - G_j\|_{C(\Omega)} \leq C_s(\epsilon + |g|_{C^s} h^s)$. Hence,

$$\|f - G_j\|_{C(\Omega)} \leq \|f - G_\nu\|_{C(\Omega)} + \|G_\nu - G_j\|_{C(\Omega)} \leq \epsilon + C_s(\epsilon + |g|_{C^s} h^s), \tag{4.8}$$

where j is the coordinate assigned by our algorithm. Finally, using that $\|\hat{f} - G_j\|_{C(\Omega)} \leq C_s |g|_{C^s} h^s$ we obtain the theorem. Notice that in this case we may have selected a wrong change coordinate j. However, since the maximum deviation of f over P_i, $i = 1, \ldots, m$, is small, estimate (4.6) still holds. $\square$.

5 The general case of k variables

The results we have just obtained for the case of one change variable extend to the general case of k change variables but the constructions and proofs are more substantial. We shall point out some of the essential new ingredients.

Our starting point is to assume that we have a set $\mathcal{P}$ of base points in $\mathbb{R}^N$ with certain properties. Let $\mathcal{A}$ be a collection of partitions $\mathbf{A}$ of $\{1, 2, \ldots, N\}$. Each $\mathbf{A}$ consists of k disjoint sets $A_1, \ldots, A_k$. We require:

Partition Assumption *The collection $\mathcal{A}$ is rich enough that it has the following two properties:*

(i) *given any k distinct integers $i_1, \ldots, i_k \in \{1, \ldots, N\}$, there is a partition $\mathbf{A}$ in $\mathcal{A}$ such that each set in $\mathbf{A}$ contains precisely one of the integers $i_1, \ldots, i_k$.*

(ii) *For any $k+1$ distinct integers $i, i_1, \ldots, i_k$, there is a partition $\mathbf{A}$ such that one of the sets A_ν contains i but none of the other $i_1, \ldots, i_k$.*

Property (i) is called *perfect hashing* in combinatorics/ computer science. It is an interesting question to understand the fewest number of sets $\mathcal{A}$ that satisfy the **Partition Assumption**. We shall show later that $C(k)\log N$ partitions suffice. But for now, we assume that we have such a collection $\mathcal{A}$ in hand and proceed to construct an algorithm for approximating f.

Corresponding to any $\mathbf{A} \in \mathcal{A}$ we construct the set of *base points*

$$P = \sum_{i=1}^{k} \alpha_i \chi_{A_i}, \quad \alpha_i \in \{0, 1/m, \ldots, 1\}. \tag{5.1}$$

In other words, P has coordinate value α_i at each of the coordinate indices in A_i. We denote by $\mathcal{P}$ the set of all such base points. Note that there are $(m+1)^k \#(\mathcal{A})$ such base points.

The important property of the base points is:

Projection Property: *Given any $\mathbf{j} = (j_1, \ldots, j_k)$ and any integers $0 \le i_{j_1}, \ldots, i_{j_k} \le m$, there is a point $P \in \mathcal{P}$ such that the coordinate j_ν of P is i_{j_ν}/m. Said in another way, given any k dimensional coordinate plane, the projection of $\mathcal{P}$ onto this plane gives a uniform grid with spacing $h := 1/m$.*

Indeed, it is enough to take a partition $\mathbf{A}$ from $\mathcal{A}$ such that each j_ν is in a different set A_i of $\mathbf{A}$. Then the point (5.1) with the appropriate value of $\alpha_i = i_{j_\nu}/m$ when $j_\nu \in A_i$ will have the value i_{j_ν}/m at coordinate j_ν.

Next we show how the partitions $\mathcal{A}$ give a bit representation for any integer $j \in \{1, \ldots, N\}$. We denote by $b_{\mathbf{A}}(j)$ the integer $i \in \{1, \ldots, k\}$ for which $j \in A_i$. We view the vector $(b_{\mathbf{A}}(j))_{\mathbf{A}\in\mathcal{A}}$ as a bitstream associated to j. [1] Let us observe

Uniqueness Property: *If two integers j, j' have identical bitstreams, then $j = j'$.*

Indeed, if $j \neq j'$ then we can find $\mathbf{A}$ such that j and j' lie in different partition sets of $\mathbf{A}$ because of the **Partition Assumption**. Therefore, $b_{\mathbf{A}}(j) \neq b_{\mathbf{A}}(j')$.

We shall also need padding points. In our analog of the first univariate algorithm, we construct the padding points for certain ordered pairs of base points which we call *admissible pairs. An ordered pair (P, P') with $P \neq P'$* of base points is admissible if: (a) they are subordinate to the same partition $\mathbf{A}$, (b) there is an integer i such that P and P' agree on all of the sets $A_\nu \neq A_i$ and on A_i, the coordinate values of P and P' are each constant and differ by $1/m$. Given P, we see that there are $2k$ choices of P' for which P, P' is an admissible pair. Namely, we have k choices for A_i and once this choice is made there are two choices for the values of P' on A_i. Hence, there are $2k\#(\mathcal{P}) = 2k(m+1)^k\#(\mathcal{A})$ admissible pairs.

For each admissible pair P, P' we define a collection of padding points $Q = [P, P']_{\mathbf{B},\nu}$ for each $\mathbf{B} \in \mathcal{A}$, $\mathbf{B} \neq \mathbf{A}$, and $\nu \in \{1, \ldots, k\}$ as follows. The j-th coordinate of Q for each $j \in A_\mu$, $\mu \neq i$, is the common j-th coordinate of P and P'. In other words, we do not alter the padding points except on A_i. For each $j \in A_i$, the j-th coordinate of Q is the same as that of P' if $j \in A_i \cap B_\nu$. Otherwise it is the same as that of P. In other words, the padding points change some of the coordinates of P to that of P'. The coordinates that are changed are precisely those in $B_\nu \cap A_i$.

[1] In the language of theoretical computer science we are using here a perfect hash function.

We denote the set of all such padding points by $\mathcal{Q}$. Since there are $\#(\mathcal{A})-1$ choices of $\mathbf{B}$ and k choices of ν there are at most $k(\#(\mathcal{A})-1)$ padding points associated to each admissible pair. Thus,

$$\#(\mathcal{Q}) \leq 2k^2\#(\mathcal{P})(\#(\mathcal{A})-1) = 2k^2(m+1)^k(\#(\mathcal{A}))(\#(\mathcal{A})-1). \tag{5.2}$$

We now proceed to describe an algorithm which is the non- adaptive version of the algorithm for one change coordinate. Again, adaptive questioning would help reduce some on the number of questions we ask.

We assume that f is a function that depends on at most k variables (unknown to us) whose indexing we denote by $\mathbf{j} = (j_1, \ldots, j_k)$ with $1 \leq j_1 < j_2 < \cdots < j_k \leq N$. We call the entries in $\mathbf{j}$ the *change coordinates* of f. We introduce the following general notation. Given a vector $\mathbf{l} = (l_1, \ldots, l_k)$, with $1 \leq l_1 < \ldots < l_k \leq N$, we define the function $G_{\mathbf{l}}$ by $G_{\mathbf{l}}(x_1, \ldots, x_N) = g(x_{l_1}, \ldots, x_{l_k})$. Thus, we know that $f = G_{\mathbf{j}}$.

The Algorithm starts by asking for the values of f at all points in $\mathcal{P} \cup \mathcal{Q}$. We then proceed to find the change coordinates $\mathbf{j}$ as follows. Given any admissible pair P, P', let $\mathbf{A}$ be the subordinating partition of P and P' and let A_i be the set in $\mathbf{A}$ where P and P' take differing values. We examine the values of f at all the padding points Q associated to this pair. We say the pair P, P' is *useful* if for each $\mathbf{B} \in \mathcal{A}$, we have exactly one value $\nu = \nu(\mathbf{B})$ where $f([P, P']_{\mathbf{B},\nu}) = f(P')$ and for all $\mu \neq \nu$, we have $f([P, P']_{\mathbf{B},\mu}) = f(P)$. For each such admissible and useful pair, we then define

$$J_{P,P'} := \bigcap_{\mathbf{B}\in\mathcal{A}} B_{\nu(\mathbf{B})} \cap A_i \tag{5.3}$$

The following lemma will show that we can identify the change coordinates of f. We say that the change coordinate j_ν is visible at scale m if there exists two points $m^{-1}(i_1, \ldots, i_N)$ and $m^{-1}(i'_1, \ldots, i'_N)$, $0 \leq i_1, i'_1, \ldots, i_N, i'_N \leq m$, which are identical in all coordinates except for the j_ν-th coordinate and $f(m^{-1}(i_1, \ldots, i_N)) \neq f(m^{-1}(i'_1, \ldots, i'_N))$. We can always take $i'_{j_\nu} = i_{j_\nu} + 1$

Lemma 5.1 *The following properties hold:*

(i) *For each admissible and useful pair P, P' the set $J_{P,P'}$ is either empty or it contains precisely one integer j,*

(ii) *this integer j is one of the change coordinates $j_1, \ldots, j_k$ of f.*

(iii) *For each change coordinate j which is visible at scale m, there is an admissible, useful pair P, P' for which $J_{P,P'} = \{j\}$*

Proof: (i) Given any two distinct integers $j, j' \in A_i$, we want to show that not both of these integers can be in $J_{P,P'}$. To see this, we take a partition $\mathbf{B}$ such that $j \in B_\nu$ and $j' \in B_{\nu'}$ and $\nu \neq \nu'$. The existence of such a partition follows from the **Partition Assumption**. From the definition of useful, we cannot have both $f([P, P']_{\mathbf{B},\nu}) = f(P')$ and $f([P, P']_{\mathbf{B},\nu'}) = f(P')$. So only one of these integers j, j' can be in $J_{P,P'}$.

(ii) Suppose $J_{P,P'} = \{j\}$. If j is not a change coordinate then by virtue of the **Partition Assumption** there is a partition $\mathbf{B}$ in which j appears in one set B_μ and all the change coordinates are outside B_μ. But since there are no change coordinates in B_μ, we have

$f([P,P']_{\mathbf{B},\mu}) = f(P) \neq f(P')$. Hence $\mu \neq \nu(B)$ which is a contradiction to j being in $J_{P,P'}$. Thus, j must be a change coordinate.

(iii) Given a change coordinate j which is visible at scale m, we know there is a point $R := m^{-1}(i_1, \ldots, i_N)$ such that incrementing i_j by one gives a point $R' := m^{-1}(i'_1, \ldots, i'_N)$ at which the value of f changes, i.e. $f(R') \neq f(R)$. By the **Partition Assumption**, we can choose a partition $\mathbf{A}$ such that each cell A_μ, $\mu = 1, \ldots, k$, contains exactly one change coordinate. Let j be in cell A_i. Consider the pairs P, P' subordinate to $\mathbf{A}$, for which P and P' differ only on A_i. We can take such a pair so that P is identical to R at each change coordinate and P' is identical to R' at each change coordinate Then, for any $\mathbf{B}$, we have $f([P,P']_{\mathbf{B},\nu}) = f(P')$ if and only if $j \in B_\nu$. Thus, P, P' is useful and this also shows that $j \in J_{P,P'}$, as desired. □

The lemma proves that the change coordinates of f that are visible at scale m have been identified. There may be $\ell \leq k$ of these coordinates, so we add arbitrarily $k - \ell$ coordinates to obtain $\mathbf{j}'$ such that $j'_\nu = j_\nu$ for any ν such that j_ν is visible at scale m. On the lattice $h(i_1, \ldots, i_N)$, $0 \leq i_1, \ldots, i_N \leq m$, with $h := m^{-1}$, we have $f = G_{\mathbf{j}} = G_{\mathbf{j}'}$.

We can now identify the values of g at each of the lattice points $h\mathcal{L}_k := \{h(i_1, \ldots, i_k)\}$, where $0 \leq i_1, \ldots, i_k \leq m$ as follows. We take a partition $\mathbf{A} \in \mathcal{A}$ for which each coordinate j'_ν lies in a different set of the partition. If we take the coordinates of P to be $i_{j'_\nu}/m$ on the set of $\mathbf{A}$ which contains j'_ν, then we obtain a base point P such that $f(P) = G_{\mathbf{j}}(P) = G_{\mathbf{j}'}(P) = g(h(i_1, \ldots, i_k))$.

Now that we have the values of g on the lattice $h\mathcal{L}_k$, we can find the approximation $A_m(g)$ which satisfies $\|g - A_m(g)\|_{C(\Omega_0)} \leq C_s |g|_{C^s} h^s$. We define $\hat{f}(x_1, \ldots, x_N) := A_m(g)(x_{j'_1}, \ldots, x_{j'_k})$.

Theorem 5.2 *The number of point values used in Algorithm 1 is* $\leq 2k^2(m+1)^k(\#(\mathcal{A}))^2$. *If* $f = G_{\mathbf{j}}$ *with* $g \in C^s$, *then the function* $\hat{f}$ *defined by Algorithm 1 satisfies*

$$\|f - \hat{f}\|_{C(\Omega)} \leq C_s |g|_{C^s} h^s. \tag{5.4}$$

Proof: The number of point values used is

$$\#(\mathcal{P}) + \#(\mathcal{Q}) \leq (m+1)^k(\#(\mathcal{A})) + 2k^2(m+1)^k(\#(\mathcal{A}))(\#(\mathcal{A}) - 1) \leq 2k^2(m+1)^k(\#(\mathcal{A}))^2.$$

To prove the bound on the approximation error, we note that

$$f - \hat{f} = G_{\mathbf{j}} - \hat{f} = G_{\mathbf{j}} - G_{\mathbf{j}'} + G_{\mathbf{j}'} - \hat{f} = G_{\mathbf{j}} - G_{\mathbf{j}'} + [g - A_m(g)](x_{j'_1}, \ldots, x_{j'_k}). \tag{5.5}$$

The first term on the right satisfies

$$[G_{\mathbf{j}} - G_{\mathbf{j}'}](x_1, \ldots, x_N) = g(x_{j_1}, \ldots, x_{j_k}) - g(x_{j'_1}, \ldots, x_{j'_k}) = \bar{g}(x_{j_1}, \ldots, x_{j_k}, x_{j'_1}, \ldots, x_{j'_k}).$$

Since $\bar{g}$ is a C^s function of $\ell < 2k$ variables which vanishes on the lattice $h\mathcal{L}_\ell$, one finds $\|G_{\mathbf{j}} - G_{\mathbf{j}'}\| \leq C_s |g|_{C^s} h^s$. The second term on the right of (5.5) can also be bounded by $C_s |g|_{C^s} h^s$ and we therefore obtain (5.4). □.

Remarks :

1. We show next that there are sets $\mathcal{A}$ which satisfy the **Partition Assumption** which contain $\approx 2ke^k[\ln N]$ elements. This gives a $C(k)[\log N]^2$ bound for the number of querries in the algorithm. One logarithm can be dropped by working adaptively.

2. There is also an adaptive version of the approximation result Theorem 4.2 which we shall not formulate because of time.

6 Constructing separating parititions

The last algorithm we have given begins with a set $\mathcal{A}$ of partitions that satisfy the **Partition Assumption**. It is important to know how large $\mathcal{A}$ needs to be for this assumption to hold. Indeed $\#(\mathcal{A})$ controls the size of the sets $\mathcal{P}$ and $\mathcal{Q}$ where we sample f. We shall begin by giving a constructive way to find sets $\mathcal{A}$ that satisfy the **Partition Assumption** that works in the case k is small. Later we shall give a probabilistic proof that for any k there are sets $\mathcal{A}$ which satisfy the **Partition Assumption** and have reasonable cardinality.

The problem of finding sets $\mathcal{A}$ satisfying (i) of our **Partition Assumption** is known in theoretical computer science as perfect hashing. Let us consider the case $k = 2$ which will be illustrative of how to do low dimensional constructions. Consider first the collection $\mathcal{B}$ of binary partitions $\mathbf{B}$ of $\{1, \dots, N\}$. Thus each $\mathbf{B}$ is determined by an integer $\nu \in \Lambda := \{1, \dots, \lceil \log_2 N \rceil\}$ and gives the two sets $B_0(\nu), B_1(\nu)$ where B_0 (respectively B_1) consists of all the integers in $\{1, \dots, N\}$ whose ν-th binary bit is 0 (respectively 1). The collection $\mathcal{B}$ will clearly satisfy (i) of the **Partition Assumption** but it will not satisfy (ii). To obtain (ii), we add some additional partitions to $\mathcal{B}$. Namely, for each $\nu, \nu' \in \{1, \dots, \lceil \log_2 N \rceil\}$, $\mu, \mu' = 0, 1$, we consider the new partition given by the two sets

$$[B_0(\nu) \cap B_\mu(\nu')] \cup [B_1(\nu) \cap B_{\mu'}(\nu')], \quad [B_0(\nu) \cap B_{\bar{\mu}}(\nu')] \cup [B_1(\nu) \cap B_{\bar{\mu}'}(\nu')], \tag{6.1}$$

where $\bar{\mu}$ denotes the complimentary index to μ (if $\mu = 0$ then $\bar{\mu} = 1$ and vice versa). When these partitions are added to the collection $\mathcal{B}$, we obtain a collection $\mathcal{A}$ which satisfies both (i) and (ii) of the **Partition Assumption**. To verify (ii), let j, j_1, j_2 be three distinct integers from $\{1, \dots, N\}$ and choose ν such that j is in one of the $B_i(\nu)$ (say $B_0(\nu)$) and j_1 is in the other (say $B_1(\nu)$). If j_2 is in $B_1(\nu)$ then we have the desired partition. In the other case, where both j, j_2 are in $B_0(\nu)$, we choose a second partition $B_0(\nu'), B_1(\nu')$ which separates j and j_2. In this case, we can without loss of generality assume that $j \in B_0(\nu) \cap B_0(\nu')$ and $j_2 \in B_0(\nu) \cap B_1(\nu')$. The integer j_1 will be in either $B_1(\nu) \cap B_0(\nu')$ or $B_1(\nu) \cap B_1(\nu')$. Hence one of the partitions in (6.1) will separate j from j_1, j_2. Notice that the collection $\mathcal{A}$ will have cardinality $\lceil \log_2 N \rceil + 4\lceil \log_2 N \rceil^2$.

Constructions of the above form become more unwieldly as k increases. Also they increase the power of $\log_2 N$. However, there are probabilistic arguments which show for any given k, the existence of a family $\mathcal{A}$ which satisfies the Partition Assumption and has favorable cardinality.

Suppose we are given N and $k < N$. We are interested in partitions $\mathbf{A}$ of $\Lambda := \{1, \dots, N\}$ consisting of k disjoint sets $A_1, \dots, A_k$. We view each A_i as a bucket which will have in it a collection of integers from Λ. We want to have sufficiently many partitions $\mathbf{A}$ to satisfy the Partition Assumption but we also want a control on the number of such partitions we shall need. We shall see that $2ke^k \lceil \ln N \rceil$ partitions will suffice.

To see this, we let ϕ denote the random variable which takes the values $\{1, \dots, k\}$ with the equal probability $1/k$. Alternatively, one can think of a draw from a box of balls labeled $1, \dots, k$. We randomly draw a ball, record its label, and then replace the ball into the box to repeat the experiment. We shall take qN independent draws ϕ_j of ϕ. We shall decide q later. Those variables define q random partitions $\mathbf{A} = \mathbf{A}(j) = \{A_1(j), \dots, A_k(j)\}$, $j = 1, \dots, q$, as

follows. For each $\mu \in \Lambda$, we place $\mu \in A_s(j)$ if and only if $\phi_{(j-1)N+\mu}(\omega) = s$. In other words, for each partition $\mathbf{A}(j)$ we run through the integers $1, \ldots, N$ in order and place the integer μ in the bucket $A_s = A_s(j)$ if the μ-th draw is s. Notice that for a given j, some of the sets $A_1(j), \ldots, A_k(j)$ may be empty.

If we are given $\mathbf{l} = \{l_1, \ldots, l_k\}$ then it is easy to see that the probability that for a given j, the parition $\mathbf{A}(j)$ separates the entries of $\mathbf{l}$ into distinct sets is $k!/k^k$. Indeed, the probability that l_i is in $A_i(j)$, for each $i = 1, \ldots, k$, is k^{-k}. But any permutation of the l_i will do as well and we have $k!$ of these. So the probability that a random partition does not separate a given $\mathbf{l}$ is $a := (1 - \frac{k!}{k^k})$. Therefore, if we have q independent partitions the probability that none of them separates $\mathbf{l}$ is a^q. There are $\binom{N}{k}$ k-tuples $\mathbf{l}$ when arranged in increasing order. Thus, if $\binom{N}{k}a^q < 1$ then a set of q random partitions will separate the coordinates of every $\mathbf{l}$ with positive probability.

To see how large we need to take q we use Stirling's formula to find

$$\binom{N}{k}(1 - \frac{k!}{k^k})^q \leq \binom{N}{k}(1 - \frac{\sqrt{2\pi k}}{e^k})^q \leq N^k(1 - e^{-k})^q \tag{6.2}$$

If we take $q \approx 2ke^k \ln N$ and use the fact that $(1 - 1/x)^x \leq e^{-1}$ for $x > 1$, we see that the right side of (6.2) is

$$< N^k e^{-2k \ln N} \leq N^{-k}.$$

Thus, we see that if we take $q \geq 2ke^k \ln N$ partitions generated randomly, then with probability greater then $1 - N^{-k}$ the resulting set $\mathcal{A}_0$ will satisfy (i) of the **Partition Assumption**.

Now we look at random partitions that satisfy condition (ii). Observe that once we assigned i to a set, the probability that j_1 is assigned to a different set is $(k-1)/k$, so the probability that i is assigned to one set and all j_i's to other sets is $((k-1)/k)^k$. This means that the probability that for given numbers $i, j_1, \ldots, j_k$ the random partition fails (ii) is $1 - ((k-1)/k)^k$. So the probability that (ii) fails for q independent partitions and for all choices of $j_1, \ldots, j_k$ does not exceed (note we have $k \geq 2$)

$$\binom{N}{k}\left(1 - (\frac{k-1}{k})^k\right)^q \leq N^k(3/4)^{2ke^k \ln N} = N^{k(1+2e^k \ln(3/4))}. \tag{6.3}$$

One checks that this is at most $N^{-\beta k}$ with $\beta \geq .2$. Comparing this with the estimate for the probability of failure of (i) we see that with great probability $q \approx 2ke^k \ln N$ random partitions satisfy **Partition Assumption**.

7 A second model class for high dimensional functions

One of the weaknesses of the model classes studied above is that they are very coordinate biased. It would be desirable to have results that hold for general change of bases. Namely, we would like the set Φ to consist of all $k \times N$ matrices. We will indicate the beginnings of such a theory under development with Albert Cohen, Gerard Kerkyacharian, Dominique Picard, and Ingrid Daubechies. We will only discuss this theory for $k = 1$.

Consider a function defined for $x \in \Omega := [0,1]^N$ by

$$f(x) = g(a \cdot x),$$

where g is an unknown function on $[0,1]$ and $a = (a_1, \cdots, a_N)$ is an unknown vector with nonnegative coordinate values. We shall assume $\sum_{j=1}^N a_j = 1$ (this assumption can be weakened somewhat). We will also fix a value of $q < 1$ and assume that $\|a\|_{\ell_q} \leq M$ and try to see what quality of results we can obtain depending on s and q.

We are interested in recovering f from a small number of measurements. In order to get a convergence estimate we will assume that $g \in C^s$ for some $s > 1$. To simplify this presentation, we shall assume that $s = 2$. The arguments below generalize easily to $s > 1$.

Let $h := 1/m$ as before. For convenience of notation, we assume that $|g|_{C^2} = 1$. Let us first note that we can obtain the value of g at any point t by asking for the values of f at $t(1, \ldots, 1)$. We start by asking for the values at the base points of the previous section. Namely, we ask for the values $f(ih, \cdots, ih) = g(ih)$ for $i = 0, \cdots, m$. From this we reconstruct a piecewise linear interpolation $\hat{g}$ of g associated to the grid points $t_i := ih$. Then $\|g - \hat{g}\|_{C[0,1]} \leq h^2$ and $\|g' - \hat{g}'\|_{L_\infty[0,1]} \leq h$.

Our next goal is to find an approximation $\hat{a}$ to a by asking for further values of f. This is analogous to the padding points in the previous section. Let $A := \max_{0 \leq i,j \leq m} |g(t_i) - g(t_j)|$ and take a pair of points $t_i < t_j$ which assume A. Let $I_0 := [t_i, t_j]$. We ask for the value of g at the midpoint t of I_0. If $|g(t_i) - g(t)| \geq |g(t_j) - g(t)|$, we define I_1 as $[t_i, t]$, otherwise we define I_1 as $[t, t_j]$. In either case, the values of g at the two endpoints of I_1 differ by at least $A/2$. We continue in this way and define $I_2, \ldots, I_J$ where J is an integer which will be chosen shortly. We do impose now that $2^{-J} \leq h^4$. Now at the two endpoints of $I_J = [\alpha_0, \alpha_1]$, g has values that differ by at least $A2^{-J}$. So there is a point $\xi \in I_J$ where $|g'(\xi)| \geq A$. Let $\delta := |I_J| \leq 2^{-J} \leq h^4$. It follows that

$$|g'(\alpha_0)| \geq |g'(\xi)| - h^4 \geq A - h^4. \tag{7.1}$$

CASE 1: $A \leq h^2$**.** In this case, there is a constant c such that $g - c$ takes values at the points j/m which are in absolute value $\leq h^2/2$. From this and the fact that $|g|_{C^2} = 1$ gives that $\|g - c\|_{C(\Omega_0)} \leq Ch^2$ and $\|\hat{g} - c\|_{C(\Omega_0)} \leq Ch^2$. Hence regardless how we define $\hat{a}$ (as long as $\hat{a} \geq 0$ and $\|\hat{a}\|_{\ell_1} \leq 1$) we will have for $\hat{f} := \hat{g}(\hat{a})$,

$$\|f - \hat{f}\|_{C(\Omega)} \leq \|f - c\|_{C(\Omega)} + \|\hat{f} - c\|_{C(\Omega)} \leq \|g - c\|_{C(\Omega_0)} + \|\hat{g} - c\|_{C(\Omega_0)} \leq Ch^2. \tag{7.2}$$

CASE 2: $A > h^2$**.** In this case we proceed further to find a good approximation to a by asking more questions. Let Φ be an $n \times N$ Bernouli compressed sensing matrix and let r be one of its rows. If we ask for the value of f at the point $\alpha_0(1, \ldots, 1) + \delta r$, we receive the value of g at $\alpha_0 + \delta a \cdot r$. We want to use these to compute an approximation to $y_r := a \cdot r$. Observe that

$$|y_r| \leq \|a\|_{\ell_1^N} \|r\|_{\ell_\infty^N} \leq n^{-1/2}. \tag{7.3}$$

We have the following approximation to y_r:

$$\hat{y}_r := \frac{g(\alpha_0 + \delta y_r) - g(\alpha_0)}{g(\alpha_0 + \delta) - g(\alpha_0)} = \frac{\delta y_r g'(\alpha_0) + g''(\xi_1)(\delta y_r)^2/2}{\delta g'(\alpha_0) + g''(\xi_2)\delta^2/2} = y_r \frac{1 + \epsilon_1}{1 + \epsilon_2} = y_r + y_r \frac{\epsilon_1 - \epsilon_2}{1 + \epsilon_2}, \tag{7.4}$$

where

$$|\epsilon_1| = \frac{|g''(\xi_1)|\delta|y_r|}{2|g'(\alpha_0)|} \leq \frac{\delta n^{-1/2}}{2(A-h^4)} \tag{7.5}$$

and

$$|\epsilon_2| := \frac{|g''(\xi_2)|\delta}{2|g'(\alpha_0)|} \leq \frac{\delta}{2(A-h^4)}. \tag{7.6}$$

Assuming that $m \geq 2$, we have $A - h^4 \geq 3h^2/4$ and $|\epsilon_2| \leq 2h^2/3 \leq 1/6$. Hence, $(1+\epsilon_2)^{-1} \leq 6/5$. This gives

$$|y_r - \hat{y}_r| \leq \frac{24}{15}\delta h^{-2}|y_r| \leq 2|y_r|\delta h^{-2}. \tag{7.7}$$

If we do the above for each row r of Φ then we will ask for n additional values of f and we will find an approximation $\hat{y}$ to $y = \Phi a$ satisfying

$$\|y - \hat{y}\|_{\ell_1^n} \leq 2\delta h^{-2}\|y\|_{\ell_1^n} \leq 2\sqrt{n}\delta h^{-2}, \tag{7.8}$$

where we have used that $\|y\|_{\ell_1^n} \leq \sqrt{n}$.

We now use ℓ_1-minimization to decode $\hat{y}$ resulting in a vector $\hat{a} \in \mathbb{R}^N$. The following lemma tells us the quality of the approximation of $\hat{a}$ as an approximation to a.

Lemma 7.1 *Under the assumptions of Case 2, for any $k = 1, 2, \ldots,$, by choosing $n \approx k\log(N/k)$ and J large enough so that $2^{-J} \leq h^4 n^{-3/2}[\log(N/n)]^{-1/2}$, we find*

$$\|a - \hat{a}\|_{\ell_1} \leq C\{\sigma_k(a)_{\ell_1} + h^2\}, \tag{7.9}$$

for an absolute constant C.

Proof: To start with, we know that

$$\|y - \hat{y}\|_{\ell_2^n} \leq \sqrt{n}\|y - \hat{y}\|_{\ell_1^n} \leq 2n\delta h^{-2} \leq 2n^{-1/2}h^2, \tag{7.10}$$

and

$$\|y - \hat{y}\|_{\ell_\infty^n} \leq [\sqrt{\log(N/n)}]^{-1}\{2n^{-1/2}\sqrt{\log(N/n)}\delta h^{-2}\} \leq 2[\sqrt{\log(N/n)}]^{-1}n^{-1/2}h^2. \tag{7.11}$$

Hence, from the **CBMP** that there is a vector z with $\Phi(z) = \hat{y} - y$ and

$$\|z\|_{\ell_1^N} \leq C_0 h^2, \tag{7.12}$$

where we have used that $\delta \leq 2^{-J}$ and the choice of J. Since $\Phi(a+z) = \Phi(\hat{a}) = \hat{y}$, we have from the ℓ_1 instance optimality in ℓ_1 that

$$\|a + z - \hat{a}\|_{\ell_1} \leq C\sigma_k(a+z)_{\ell_1^N} \leq C\{\sigma_k(a)_{\ell_1} + \|z\|_{\ell_1}\} \leq C\{\sigma_k(a)_{\ell_1} + h^2\}, \tag{7.13}$$

where we used $\|z\|_{\ell_1^N} \leq C_0 h^2$. □

Remark: We want the resulting $\hat{a}$ to have nonnegative coordinates and $\sum_{j=1}^N \hat{a}_j = 1$. We can always modify the $\hat{a}$ above to have this property without effecting the approximation error (7.9) because we know that a has these properties. So in going forward we assume $\hat{a}$ has these properties.

The following theorem now summarized the performance of the algorithm. We formulate the theorem for all $1 < s$ even though we are only proving it for $s = 2$.

Theorem 7.2 *If $f(x) = g(a \cdot x)$ with $\|a\|_{\ell_q^N} = M$ then the above algorithm uses*

$$\leq C(m + m^{\frac{2}{1/q-1}} \log N) \tag{7.14}$$

querries of f and returns an approximation $\hat{f}(x) := \hat{g}(\hat{a} \cdot x)$ which satisfies

$$\|f - \hat{f}\|_{C(\Omega)} \leq C_s\{|g|_{C^s} + \|a\|_{\ell_q^N}\}h^s, \quad h := m^{-1}. \tag{7.15}$$

Proof: As has been our convention, we give the proof for $s = 2$ under our assumption that $|g|_{C^s} = 1$ and $\|a\|_{\ell_q^N} \leq M$. We ask m question which determines $\hat{g}$ and we know that

$$\|g - \hat{g}\|_{C[0,1]} \leq h^2. \tag{7.16}$$

Next, given that $a \in \ell_q$, we can choose k as the smallest integer so that

$$k^{1-1/q} \leq m^{-2}. \tag{7.17}$$

Thus, $k \approx m^{\frac{2}{1/q-1}}$. We now define n to satisfy $n \approx k \log(N/n)$. Finally, we choose J so that

$$2^{-J} \leq m^{-4} n^{-3/2} [\log(N/n)]^{-1/2} \tag{7.18}$$

so that the assumptions of Lemma 7.1 apply. We will ask n questions using the compressed sensing matrix and arrive at the approximation

$$\|a - \hat{a}\|_{\ell_1^N} \leq C(h^2 + \sigma_k(a)_{\ell_1^N}) \leq C(h^2 + Mk^{-1/q+1}), \tag{7.19}$$

where the last inequality uses the fact that $\sigma_k(a)_{\ell_1^N} \leq \|a\|_{\ell_q^N} k^{1-1/q}$ as was proven in our first set of lecture notes.

In total, we have asked $m + J + n$ questions and from this and the definition of J and n one derives (7.14). To verify (7.15) we write $f(x) - \hat{f}(x) = g(a \cdot x) - g(\hat{a} \cdot x) + g(\hat{a} \cdot x) - \hat{g}(\hat{a} \cdot x)$, from which we derive

$$\|f - \hat{f}\|_{C(\Omega)} \leq \|g - \hat{g}\|_{C(\Omega)} + |g|_{Lip1} \|a - \hat{a}\|_{\ell_1^N} \leq C(h^2 + Mk^{-1/q+1}) \leq C(1 + M)h^2, \tag{7.20}$$

where we have used the definition of k. □

Remark: By using manifold widths, one can show that the above result cannot be improved. Although this is an interesting story, we will not have time to go into it in these lectures.

References

[1] R. DeVore, R. Howard and C. Micchelli, *Optimal non-linear approximation*, Manuscripta Math., **63** (1989), 469–478.

[2] R. DeVore, G. Kyriazis, D. Leviatan, and V. Tichomirov, *Wavelet compression and nonlinear n-widths*, Advances in Computational Math., **1**(1993), 197–214.

Course Notes (Paris 2009)

Ronald DeVore

June 24, 2009

Abstract

The following are notes on stochastic and parametric PDEs of the short course in Paris.

1 Lecture 4: Capturing Functions in Infinite Dimensions

Finally, we want to give an example where the problem is to recover a function of infinitely many variables. We will first show how such problems occur in the context of stochastic partial differential equations.

1.1 Elliptic equations: general principles

We consider the elliptic equation

$$-\nabla \cdot (a\nabla u) = f \quad \text{in} \quad D, \qquad u|_{\partial D} = 0, \tag{1.1}$$

in a bounded Lipschitz domain $D \subset \mathbb{R}^d$, where $f \in L_2(D)$. There is a rich theory for existence and uniqueness for equations of this form which we briefly recall.

Central to the theory of elliptic equations is the Sobolev space $V := H_0^1(D)$ (called the energy space) which is a Hilbert space equipped with the energy norm $\|v\|_V := \|\nabla v\|_{L^2(D)}$. The dual of V is $V^* = H^{-1}(D)$. The solution of (1.1) is defined in weak form as a function $u \in H_0^1(D)$ which satisfies

$$\int_D a(x)\nabla u(x) \cdot \nabla v(x) dx = \int_D f(x)v(x)dx, \ \text{ for all } \ v \in H_0^1(D), \tag{1.2}$$

where the gradient ∇ is taken with respect to the x variable. This formulation shows that the Lax-Milgram theory applies. In particular, a sufficient condition for the existence and uniqueness of the solution u is that a satisfies the ellipticity assumption

$$0 < a_{\min} \le a(x) \le a_{\max}, \quad x \in D. \tag{1.3}$$

Under this assumption, the solution satifies the estimate

$$\|u\|_V \le C_0 := \frac{\|f\|_{V^*}}{a_{\min}}. \tag{1.4}$$

The same theory applies even if a is complex valued. Now the lower ellipticity condition replaces a by $Re(a)$ in (1.3) and the upper condition is that $|a|$ is uniformly bounded.

There is also a general principal of perturbation for elliptic equations which shows to some extent the smooth dependence of the solution on the diffusion coefficient a. If $a, \tilde{a}$ are two such coefficients with the ellipticity constants $a_{\min}, \tilde{a}_{\min} \geq \mu > 0$, then the solutions u and $\tilde{u}$ with identical right side f will satisfy

$$\|u - \tilde{u}\|_{H_0^1(D)} \leq C(\mu)\|a - \tilde{a}\|_{L_\infty(D)}. \tag{1.5}$$

2 Stochastic equations

We are interested in the case of stochastic equations where $a = a(x, \omega)$ is now a real valued random field on some probability space (Ω, Σ, P) but f remains a deterministic function. The solution $u = u(x, \omega)$ is now a random field associated to the same probability space. Stochasticity describes the uncertainty in the diffusion coefficient a. In order to ensure uniform ellipticity, one assumes

Assumption S: *There exist constants* $0 < a_{\min} \leq a_{\max}$ *such that*

$$a_{\min} \leq a(x, \omega) \leq a_{\max}, \quad (x, \omega) \in D \times \Omega. \tag{2.1}$$

There are two general numerical approaches to stochastic elliptic PDEs: Monte-Carlo (MC) methods and deterministic methods.

Monte-Carlo (MC) methods: These methods approximate quantities such as the mean $(\overline{u}(x) := \mathbb{E}(u(x)) = \int_\Omega u(x, \omega) dP(\omega))$ or higher moments of u. One takes N independent draws of a and computes the solution u_i, $i = 1, \ldots, N$, corresponding to each of these draws and then uses the u_i to estimate the quantities of interest. For example, the average $\overline{u}_N := \frac{1}{N} \sum_{i=1}^N u_i$ gives an estimate in expectation

$$\mathbb{E}(\|\overline{u} - \overline{u}_N\|_V) \leq (\mathbb{E}(\|u\|_V^2))^{1/2} N^{-\frac{1}{2}} \tag{2.2}$$

i.e. Monte-Carlo approximations with N samples converge with rate $1/2$ provided that the solution u as a V-valued random function has finite second moments. Unfortunately, the rate $N^{-1/2}$ cannot be improved for MC.

In practice, the u_i are computed approximately by space discretization, for example by the finite element method. But we will leave this issue aside in this talk and instead focus on whether other (deterministic) methods could potentially outperform Monte-Carlo. Our benchmark is N which is the number of times we need to solve a corresponding elliptic equation.

Deterministic methods: These have been studied for several decades. In contrast to MC, these methods take advantage of the smooth dependence of u on a. We will consider the *spectral approach* which is based on the so-called Wiener generalized polynomial chaos expansion. The first step consists in representing a by a sequence of scalar random variables $(y_j)_{j \geq 1}$, usually

obtained through a decomposition of the oscillation $a - \overline{a}$ into an orthogonal basis $(\psi_j)_{j\geq 1}$ of $L_2(D)$:

$$a(x,\omega) = \overline{a}(x) + \sum_{j\geq 1} y_j(\omega)\psi_j(x). \tag{2.3}$$

Here, the reader can think of $\{\psi_j\}$ as his favorite basis, for example a wavelet basis or Fourier basis.

The solution is now viewed as a function $u(x,y)$ where $x \in D$ is the space variable and $y = (y_j)_{j\geq 1}$ is a vector of "stochastic variables", and the objective is to compute a numerical approximation to $u(x,y)$. Any such approximation would give us access to all information about the solution u. Note that

$$y_j := \|\psi_j\|_{L_2(D)}^{-2} \int_D (a - \overline{a})\psi_j, \quad j = 1, 2, \ldots . \tag{2.4}$$

Of course, for each draw $\omega \in \Omega$, y is just a sequence of real numbers. So in the end we can consider parametric problems for real sequences y.

Up to a renormalization of the basis functions ψ_j, we may assume without loss of generality that for all $j \geq 1$ the random variables y_j are such that $\|y_j\|_{L_\infty(\Omega)} = 1$. Up to a change of the definition of a on a set of measure zero in Ω this is equivalent to

$$\sup_{\omega\in\Omega} |y_j(\omega)| = 1. \tag{2.5}$$

The vector y is thus a point in the infinite dimensional cube

$$U := [-1,1]^{\mathbb{N}},$$

i.e. the unit ball of $\ell^\infty(\mathbb{N})$. Hence in going further, we assume that the ψ_j has been so normalized.

3 Parametric elliptic equations

Now, we make a major but illuminating change in our point of view. Rather than view the vectors y of interest as only those that arise as realizations of the stochastic process, we instead admit any $y \in U$ as being viable. For any such $y \in U$, we define

$$a(x,y) = \overline{a}(x) + \sum_{j\geq 1} y_j\psi_j(x). \tag{3.1}$$

and further define $u(x,y)$ as the solution to the (parametric) elliptic equation

$$-\nabla_x(a(x,y)\nabla_x u(x,y)) = f(x), \quad x \in D, \tag{3.2}$$

with boundary condition

$$u(x,y) = 0, \quad x \in \partial D.$$

In order to guarantee uniform ellipticity for $(x,y) \in D \times U$, we use

Assumption P: *We assume that there are constants* $a_{\min}$ *and* $a_{\max}$ *such that*

$$0 < a_{\min} \leq a(x,y) \leq a_{\max} < +\infty, \quad (x,y) \in D \times U.$$

This assumption is very close to the assumption imposed in the stochastic setting and in fact implies the stochastic ellipticity assumption. The only difference in these two conditions is that the set of y for which we require ellipticity may be larger in **Assumption P**.

Remark: *Notice that solving the parametric problem* (3.2) *will certainly give the solution to the stochastic problem. However, we may be solving* (3.2) *for values of* y *which do not arise from draws of* $\omega \in \Omega$.

4 Compressible representations of u

We would like to efficiently capture the function $u(x,y)$ for all $(x,y) \in D \times U$. This would in turn allow us to solve the stochastic problem as well. The goal of our work with Albert Cohen and Chris Schwab is to show that under very mild conditions on a, the solution $u(x,y)$ can be efficiently represented by a polynomial expansion in y with coefficients in V. To describe this, we introduce the standard multivariate notation. We let $\mathcal{F}$ be the set of all sequences $\nu = (\nu_1, \nu_2, \ldots)$ such that ν has finite support and each entry in ν is a nonnegative integer. So $|\nu| = \sum_{j \geq 1} |\nu_j|$ is always finite. If $\alpha = (\alpha_j)_{j \geq 1}$ is a sequence of positive numbers, we define for all $\nu \in \mathcal{F}$

$$\alpha^\nu := \prod_{j \geq 1} \alpha_j^{\nu_j}.$$

We also use the following sequence b throughout

$$b = (b_j)_{j=1}^\infty, \qquad b_j := \frac{\|\psi_j\|_{L_\infty(D)}}{a_{\min}}. \tag{4.1}$$

The theorems that follow which guarantee a sparse representation of $u(x,y)$ will assume some decay for the sequence (b_j). Let us observe that such decay conditions follow from very moderate assumptions on the smoothness of a.

Remark: *Recall that the normalization of* ψ_j *is determined by the stochastic problem and the fact that we recast it in the parameter domain* U. *It is easy to see that even very minimal smoothness conditions on* $a(x,\omega)$ *will result in decay of* (b_j). *Let us show how this goes only in the simple case of a one-dimensional Fourier expansion,*

$$a(x,\omega) = \overline{a}(x) + \sum_{k \in \mathbb{Z}} \hat{a}(k,\omega) e^{i2\pi kx}.$$

It is known that if the function $a(\cdot,\omega) - \overline{a}$ *is in* $\mathrm{Lip}(s, L^1)$ *for some* $s > 1$, *then its Fourier coefficients satisfy the decay estimate*

$$|a(k,\omega)| \leq C|k|^{-s}, \qquad |k| \geq 1,$$

with C depending on the $\mathrm{Lip}(s, L^1)$*-norm of* $a(\cdot,\omega) - \overline{a}$*. Assuming that this norm is bounded independently of* ω *and returning to* (2.4) *gives that*

$$\|\psi_j\|_{L_\infty(D)} \leq Cj^{-s}, \qquad j = 1, 2, ... \tag{4.2}$$

when we reindex the basis. Therefore ℓ^p *summability of the sequence* $(b_j)_{j\geq 1}$ *is ensured when* $s > \frac{1}{p}$.

We now know that mild smoothness condtions on a translate into decay conditions on the (b_j) so that a has a compressible decomposition on U:

$$a(x, y) = \overline{a}(x) + \sum_{j\geq 1} y_j \psi_j(x), \tag{4.3}$$

where $(\|\psi_j\|_{L_\infty(D)})$ is in ℓ_p. However, we are not so much interested in approximating a which we know but rather the solution $u(x, y)$. We are therefore interested in seeing where these decay conditions on (b_j) translate into compressible representation of $u(x, y)$. This is indeed the case.

In [1], we showed the following theorem.

Theorem 4.1 *If* (i) $\sum_{j\geq 1} b_j < 1$*, and* (ii) $(b_j) \in \ell_p$ *for some* $p < 1$*, then*

$$u(x, y) = \sum_{\nu\in\mathcal{F}} c_\nu(x) y^\nu, \tag{4.4}$$

where the functions $c_\nu(x)$ *are in* V *and* $(\|c_\nu\|_V) \in \ell_p$ *for the same value of* p.

Remark: *This theorem shows that the compressibility of* a *translates into the same compressibility of* u.

Let us say a few words about the proof of this theorem before giving it a hard looking over. To prove the theorem, we need to give estimates for $\|c_\nu\|_V$. This is not too hard. For a fixed $y \in U$, we know that for all $v \in V$

$$\int_D a(x, y)\nabla u(x, y)\nabla v(x)dx = \int_D f(x)v(x)dx.$$

Differentiating this identity with respect to the variable y_j gives

$$\int_D a(x, y)\nabla\partial_{y_j} u(x, y)\nabla v(x)dx + \int_D \psi_j(x)\nabla u(x, y)\nabla v(x)dx = 0. \tag{4.5}$$

We claim that more generally for every $v \in V$ holds

$$\int_D a(x, y)\nabla\partial_y^\nu u(x, y)\nabla v(x)dx + \sum_{\{j:\,\nu_j\neq 0\}} \nu_j \int_D \psi_j(x)\nabla\partial_y^{\nu-e_j} u(x, y)\nabla v(x)dx = 0, \tag{4.6}$$

where e_j is the Kronecker sequence with value 1 at position j and 0 elsewhere. (4.6) is proved by induction on $|\nu|$ using the same idea as used in deriving (4.5). From (4.6) it is not difficult to prove

$$\|\partial_y^\nu u(\cdot, y)\|_V \leq C_0 \sum_{\{j:\,\nu_j\neq 0\}} \nu_j b_j(|\nu| - 1)!b^{\nu-e_j} = C_0\Big(\sum_{\{j:\,\nu_j\neq 0\}} \nu_j\Big)(|\nu| - 1)!b^\nu = C_0|\nu|!b^\nu, \quad \nu \in \mathcal{F}.$$

One now proves the representation (4.4) with $c_\nu(x) := \frac{D^\nu u(x,0)}{\nu!}$ (see [1] for details).

While there is certainly a beauty in the above theorem, we are not here to praise it but rather to point out its deficiencies. This centers around the relevance of the assumption (i) (we have already argued (ii) is very reasonable). The motivation for (i) lies in the ellipticity assumption that must be imposed for $a(x,y)$. From this condition, we know that

$$\overline{a}(x) + \sum_{j\geq 1} y_j\psi_j(x) \geq \overline{a}(x) - \sum_{j\geq 1}\|\psi_j\|_{L_\infty(D)} \geq a_{\min} - \sum_{j\geq 1}\|\psi_j\|_{L_\infty(D)} > 0. \tag{4.7}$$

In other words, we do get the ellipticity condition we want. However, if the functions ψ_j are not global, this condition is much too strong and ellipticity can be guaranteed by the potentially much weaker condition

$$\sum_{j\geq 1}|\psi_j(x)| \leq \overline{a}(x) - a_{\min}, \quad x \in D. \tag{4.8}$$

which is actually equivalent to the lower inequality in **Assumption P**. Moving to this weaker condition is important in applications where we use bases such as a wavelet basis whose power lies in part in the fact that they are local. It therefore becomes an important question as to whether the above theorem can be generalized to hold under the weaker assumption (4.8). We shall indeed show this is the case but to do so we need to make another conceptual step and consider parametric problems with complex parameters.

5 Complex parametric problems

Let us recall that the problem before us is to hopefully prove Theorem 4.1 under the weaker **Assumption P** whose lower inequality is equivalent to (4.8). We shall do this by considering the extension of the parametric problem to complex parameters.

We want to go from the real variables y to the complex variables z and so we now define $a(x,z)$ as in (3.1) but with z now complex. If we are on any domain $\tilde{U}$ of complex sequences $z = (z_1, \ldots, z_n, \ldots)$ for which the series (3.1) defining $a(x,z)$ converges and satisfies

$$0 < r \leq Re(a(x,z)) \leq |a(x,z)| \leq R < \infty, \tag{5.1}$$

then the solution $u(x,z)$ to the elliptic equation with diffusion coefficient $a(x,z)$ is unique and satisfies

$$\sup_{z\in\tilde{U}}\|u(\cdot,z)\|_V \leq \frac{\|f\|_{V^*}}{r}. \tag{5.2}$$

Let us see that **Assumption P** allows us to conclude the validity of (5.1) for analytic domains that are significantly larger than U. Indeed, let us take $r := a_{\min}/2$ and take any sequence $\rho = (\rho_j)_{j\geq 1}$ where $\rho_j > 0$ and $\max(\rho_j, 1) = 1 + \epsilon_j$ with

$$\sum_{j\geq 1}\epsilon_j|\psi_j(x)| \leq r. \tag{5.3}$$

Then **Assumption P** (see (4.8)) shows that (5.1) will be satisfied for any $z = (z_j)$ provided that

$$|Re(z_j)| \leq \rho_j, \quad |z_j| \leq R, \quad j \geq 1$$

for some constant $R > 0$. In particular, if $\epsilon_j = 0$ for $j > J$, where J is arbitrary but fixed, then we obtain uniform ellipticity on the polydisk

$$U_\rho := \otimes_{j\geq 1}\{|z_j| \leq \rho_j\}, \quad \rho := (\rho_j)_{j\geq 1}. \tag{5.4}$$

5.1 Bounds for $\|c_\nu\|_V$ under Assumption P

As advertised, we shall work under the weaker condition **Assumption P** and prove that Theorem 4.1 remains valid. That is, we have the following

Theorem 5.1 *Suppose $a(x, y)$ satisfies the* **Assumption P** *and in addition $(b_j) \in \ell_p$, for some $p < 1$. Then,*

$$u(x, y) = \sum_{\nu\in\mathcal{F}} c_\nu(x)y^\nu, \quad y \in U, \tag{5.5}$$

where $(\|c_\nu\|_V)_{\nu\in\mathcal{F}} \in \ell_p$ for this same value of p.

Sketch of Proof:

1. Fix any $J \geq 1$ and define $u_J(z_1, \ldots, z_J) := u(z_1, \ldots, z_J, 0, 0, \ldots)$ which is a V valued function of J complex variables. Fix any such ρ of the above form for which $\rho_j = 1$, $j > J$. We know that we have uniform ellipticity on the polydisc U_ρ (see (5.3) and (5.4)). This gives that $\|u_J(x, z)\|_V \leq \frac{\|f\|_{V^*}}{r}$ uniformly for $z \in U_\rho$. The next step is to show that u_J is holomorphic (i.e. infinitely differentiable in U_ρ). This is proved in the same manner that we have estimated $\|\partial_y^\nu u(\cdot, y)\|_V$ using the real method.

2. From the holomorphy above, we have that $u_J(x, z)$ is a V-valued analytic function in the polydisc U_ρ. Let $\mathcal{F}_J$ denote the sequences in $\mathcal{F}$ which are supported on $\{1, \ldots, J\}$. The Cauchy formula gives that for any $\nu \in \mathcal{F}_J$, we have for $c_\nu(x) := \frac{\partial_y^\nu u(x,0)}{\nu!}$, that

$$\|c_\nu\|_V \leq C_r \prod_{j>0} \rho_j^{-\nu_j} = C_r\rho^{-\nu}, \tag{5.6}$$

with the convention that $\rho_j^{-\nu_j} = 1$ if $\nu_j = 0$.

3. An important remark is that we have at our disposal the ability to choose the sequence ρ (as long as it satisfies (5.3)). We will use this option to choose for each ν a sequence tailored to ν. Namely, we first choose J_0 large enough such that

$$\sum_{j>J_0} b_j \leq 1/16. \tag{5.7}$$

Such a J_0 exists because $b \in \ell_p \subset \ell_1$. Note that we may always assume, up to some reindexing, that the sequence b is non-increasing. We split $\mathbb{N}$ into the two sets

$$E := \{0 < j \leq J_0\} \text{ and } F := \{j > J_0\}.$$

For $j \in E$, we set

$$\rho_j := \kappa > 1,$$

with κ chosen so that

$$(\kappa-1)\sum_{j\le J_0}\|\psi_j\|_{L_\infty(D)}\le\frac{a_{\min}}{4}. \tag{5.8}$$

For $j\in F$ we set

$$\rho_j:=\frac{\nu_j}{4|\nu_F|b_j},$$

where we use the notation ν_F for the restriction of ν to a set F and $|\cdot|$ denotes the ℓ^1 norm so that $|\nu_F|:=\sum_{j>J_0}\nu_j$. With such choices, we have for all $x\in D$, and for the sequence $\epsilon_j:=\max(\rho_j,1)-1$,

$$\begin{aligned}\sum_{j>0}\epsilon_j|\psi_j(x)| &\le(\kappa-1)\sum_{j\le J_0}|\psi_j(x)|+\sum_{j>J_0}|\rho_j||\psi_j(x)|\\ &\le\frac{a_{\min}}{4}+\sum_{j>J_0}\frac{\nu_j a_{\min}|\psi_j(x)|}{4|\nu_F|\|\psi_j\|_{L_\infty(D)}}\\ &\le\frac{a_{\min}}{4}+\frac{a_{\min}}{4}=\frac{a_{\min}}{2}=r\end{aligned} \tag{5.9}$$

which proves that (5.3) holds and $u_J(x,z)$ is analytic on U_ρ. Introducing the notation

$$\eta:=\frac{1}{\kappa}<1\ \text{ and }\ d_j:=4b_j$$

we have from (5.6)

$$\|c_\nu\|_V\le C_r\Big(\prod_{j\in E}\eta^{\nu_j}\Big)\Big(\prod_{j\in F}\Big(\frac{|\nu_F|d_j}{\nu_j}\Big)^{\nu_j}\Big). \tag{5.10}$$

This is the bound we want for the $\|c_\nu\|_V$ from which we can derive the theorem.

4. Let us now discuss how one proves that the bound (5.10) implies $(\|c_\nu\|)_{\nu\in\mathcal{F}}$ is in ℓ_p. We denote by $d=(d_j)_{j\in F}$ the sequence of the d_j indexed only by the $j\in F$. Note that by assumption (5.7), we have

$$\|d\|_{\ell_1}=\sum_{j>J_0}d_j\le\frac{1}{4}. \tag{5.11}$$

The estimate (5.10) has the general form

$$\|c_\nu\|_V\le C_r\alpha(\nu_E)\,\beta(\nu_F). \tag{5.12}$$

Such a general form allows to perform a factorization in the estimate the ℓ_p norms of the $\|c_\nu\|_V$: introducing $\mathcal{F}_E$ and $\mathcal{F}_F$ the sequences of positive integers with finite support indexed by E and F respectively, we can write

$$\begin{aligned}\textstyle\sum_{\nu\in\mathcal{F}}\|c_\nu\|_V^p &\le C_r^p\textstyle\sum_{\nu\in\mathcal{F}}\alpha(\nu_E)^p\beta(\nu_F)^p\\ &=C_r^p\Big(\textstyle\sum_{\nu\in\mathcal{F}_E}\alpha(\nu)^p\Big)\Big(\sum_{\nu\in\mathcal{F}_F}\beta(\nu)^p\Big)\\ &:=C_r^p\,A_E\,A_F\end{aligned}$$

In our particular setting, the first factor A_E is easily estimated, again by factorization since we have

$$\begin{aligned}A_E &=\textstyle\sum_{\nu\in\mathcal{F}_E}\alpha(\nu)^p\\ &=\textstyle\sum_{\nu\in\mathcal{F}_E}\prod_{j\in E}\eta^{p\nu_j}\\ &=\textstyle\prod_{j\in E}(\sum_{n\ge0}\eta^{np})\end{aligned}$$

and therefore

$$A_E \leq \Big(\frac{1}{1-\eta^p}\Big)^{J_0} < +\infty. \tag{5.13}$$

The second factor A_F requires a bit more work because of the appearance of the factorial factors. This bound is given in the appendix.

5. Finally, we need to show that $u(x,y) = \sum_{\nu\in\mathcal{F}} c_\nu(x)y^\nu$. The analysis that has just been given shows that $u_J(x,y) = \sum_{\nu\in\mathcal{F}_J} c_\nu(x)y^\nu$. We arrive at the representation of u from the fact that $\sup_{y\in U} \|u(\cdot,y) - u_J(\cdot,y)\|_V \to 0$, $J\to\infty$ because of the stability result (1.5).

6 Concluding remarks

6.1 Approximation of u

If a satisfies **Assumption P** and in addition the sequence $(b_j) \in \ell_p$ for some $p < 1$, then we have proven that $(\|c_\nu\|_V)_{\nu\in\mathcal{F}}$ is also in ℓ_p and

$$u(x,y) = \sum_{\nu\in\mathcal{F}} c_\nu(x)y^\nu. \tag{6.1}$$

Hence there is a polynomial space $\mathcal{P}_{\Lambda_N}$ spanned by z^ν, $\nu\in\Lambda_N$, with $\#(\Lambda_N)\leq N$, such that

$$\|u(x,y) - \sum_{\nu\in\Lambda_N} c_\nu(x)y^\nu\|_V \leq CN^{1-1/p}. \tag{6.2}$$

The sets Λ_N can be taken to have a lot of structure. For example, they can be taken as nested as N increases. Also, one can prove that for a given Λ_N, whenever $\nu\in\Lambda_N$ then all $\mu\leq\nu$ can be taken in Λ_N.

The fact that $u(x,y)$ can be well approximated by certain polynomials (with coefficients in V) can be thought of as a regularity theorem on the structure of the solution.

6.2 Numerical algorithms

We have not discussed any numerical algorithms based on the above results. If we know the corresponding space of polynomials then algorithms can be developed based on space discretization for example by using finite element methods (see [1]).

6.3 The role of the underlying probability measure:

The main point of the above analysis is to show that under very mild smoothness conditions on $a(x,\omega)$, the solution to all of the stochastic problems that arise have a very sparse representation. This was done by moving to a parametric setting. One could ask where the original probability measure has gone? The influence of this measure is felt in the assumption that the renormalized bases $\|\psi_j\|_{L_\infty(D)}$ have a mildly fast decay. This assumption will hold if the $a(x,\omega)$ all have mild smoothness but this smoothness is required uniformly over the draws $\omega\in\Omega$. It is clear that such a theory could be developed with the weaker assumption that only with high probability the $a(x,\omega)$ have the required smoothness.

Once one has the required smoothness, the role of the probability measure is not felt since we actually show there is a polynomial space of dimension N such that for any ω the solution $u(x,\omega)$ can be approximated by V valued polynomials to accuracy $O(N^{1-1/p})$ in the V norm.

6.4 Comparison with Monte Carlo

We have begun with a discussion of the disadvantages of Monte Carlo. Have we overcome these? Well, Monte Carlo applies in a different setting where one receives N independent draws of a. For our theory to have an impact, we would have to be able to ask directed questions. It is reasonable to expect that asking N directed questions we should be able to recover $u(x,\omega)$ to accuracy $O(N^{1-1/p})$ in the V norm, uniformly in $\omega \in \Omega$.

7 Appendix: Completion of bound for A_F

In our particular setting for $\nu \in \mathcal{F}$, we have

$$\beta(\nu) := \prod_{j\in F} \Big(\frac{|\nu_F| d_j}{\nu_j}\Big)^{\nu_j} = \frac{|\nu_F|^{|\nu_F|}}{\prod_{j\in F} \nu_j^{\nu_j}} d^{\nu_F}, \tag{7.1}$$

where we have used the notation $d^{\nu_F} = \prod_{j\in F} d_j^{\nu_j}$ and the convention that $0^0 = 1$. We first transform the quantities of the form n^n into $n!$ by using Stirling type estimates: for all $n > 0$, we have

$$\frac{n! e^n}{e\sqrt{n}} \leq n^n \leq \frac{n! e^n}{\sqrt{2\pi}\sqrt{n}}. \tag{7.2}$$

The right hand side is actually equivalent to n^n as $n \to +\infty$, and the left hand side is easy to prove by passing to the logarithm. In addition, according to our convention

$$0^0 = 0! = 1. \tag{7.3}$$

Using the right inequality in (7.2) without even using the factor $\sqrt{n}$, we bound by above the quantity $|\nu_F|^{|\nu_F|}$ in (7.1) by

$$|\nu_F|^{|\nu_F|} \leq \frac{|\nu_F|! e^{|\nu_F|}}{\sqrt{2\pi}}.$$

On the other hand, using the left inequality in (7.2) as well as (7.3), we bound by below the quantity $\prod_{j\in F} \nu_j^{\nu_j}$ in (7.1) by

$$\prod_{j\in F} \nu_j^{\nu_j} \geq \frac{\nu_F! e^{|\nu_F|}}{\prod_{j\in F} \max\{1, e\sqrt{\nu_j}\}}.$$

Injecting these estimates in (7.1), this allows to bound $\beta(\nu)$ as follows

$$\beta(\nu) \leq \frac{|\nu_F|!}{\nu_F!} d^{\nu_F} \prod_{j\in F} \max\{1, e\sqrt{\nu_j}\}. \tag{7.4}$$

We now use a lemma from [1], that says that since $\|d\|_{\ell^1} = \frac{1}{4} < 1$ and since $\|d\|_{\ell^p} < +\infty$, we may factorize d as

$$d_j = \gamma_j \delta_j,$$

where γ and δ are positive sequences such that

$$\|\gamma\|_{\ell^1} < 1, \ \ \|\delta\|_{\ell^\infty} < 1 \ \text{ and } \ \|\delta\|_{\ell^q} < \infty, \ \ \text{with } \ q := \frac{p}{1-p}.$$

Then, using this factorization in (7.4) and Hölder's inequality, we obtain

$$A_F \quad = \sum_{\nu\in\mathcal{F}_F} \beta(\nu)^p \le \sum_{\nu\in\mathcal{F}_F} \Big(\tfrac{|\nu_F|!}{\nu_F!} d^{\nu_F} \prod_{j\in F} \max\{1, e\sqrt{\nu_j}\} \Big)^p \tag{7.5}$$

$$\le \Big(\sum_{\nu\in\mathcal{F}_F} \tfrac{|\nu_F|!}{\nu_F!} \gamma^\nu \Big)^p \Big(\sum_{\nu\in\mathcal{F}_F} \prod_{j\in F} \max\{1, (e\sqrt{\nu_j})^q\} \delta_j^{q\nu_j} \Big)^{1-p} \tag{7.6}$$

We know that the first factor is bounded since $\|\gamma\|_{\ell^1} < 1$, and we actually have

$$\sum_{\nu\in\mathcal{F}_F} \frac{|\nu_F|!}{\nu_F!} \gamma^\nu = \frac{1}{1-\|\gamma\|_{\ell^1}}.$$

The second factor can be computed by factorization

$$\sum_{\nu\in\mathcal{F}_F} \prod_{j\in F} \max\{1, (e\sqrt{\nu_j})^q\} \delta_j^{q\nu_j} = \prod_{j\in F} \Big(\sum_{n\ge 0} \max\{1, (e\sqrt{n})^q\} \delta_j^{qn} \Big)$$

Since $\|\delta\|_{\ell^\infty} < 1$, it is not very difficult to prove that the sums in $n \ge 0$ that appear in the product all converge and that one can bound them by

$$\sum_{n\ge 0} \max\{1, (e\sqrt{n})^q\} \delta_j^{qn} \le 1 + C\delta_j^q,$$

where the constant C is independent of j. Since $\delta \in \ell^q$, this second factor is also bounded. This concludes the sketch of the proof of the theorem.

References

[1] A. Cohen, R. DeVore and C. Schwab, *Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs*, preprint.