CONTEMPORARY MATHEMATICS

505

Harmonic Analysis and Partial Differential Equations

8th International Conference on
Harmonic Analysis and Partial Differential Equations
June 16–20, 2008
El Escorial, Madrid, Spain

Patricio Cifuentes
José García-Cuerva
Gustavo Garrigós
Eugenio Hernández
José María Martell
Javier Parcet
Alberto Ruiz
Fernando Soria
José Luis Torrea
Ana Vargas
Editors

American Mathematical Society

CONTEMPORARY MATHEMATICS

505

Harmonic Analysis and Partial Differential Equations

8th International Conference on
Harmonic Analysis and Partial Differential Equations
June 16–20, 2008
El Escorial, Madrid, Spain

Patricio Cifuentes
José García-Cuerva
Gustavo Garrigós
Eugenio Hernández
José María Martell
Javier Parcet
Alberto Ruiz
Fernando Soria
José Luis Torrea
Ana Vargas
Editors

American Mathematical Society
Providence, Rhode Island

2000 *Mathematics Subject Classification.* Primary 35–XX, 42–XX, 47–XX, 52–XX, 53–XX, 58–XX, 65–XX, 94–XX, 30–XX, 26–XX.

Library of Congress Cataloging-in-Publication Data

International Conference on Harmonic Analysis and Partial Differential Equations (8th : 2008 : San Lorenzo de El Escorial, Spain)

Harmonic analysis and partial differential equations : proceedings of the 8th international conference, June 16–20, 2008, El Escorial, Madrid, Spain / Patricio Cifuentes... [et al.], editors.

p. cm. — (Contemporary mathematics ; v. 505)

Includes bibliographical references.

ISBN 978-0-8218-4770-1 (alk. paper)

1. Harmonic analysis—Congresses. 2. Differential equations, Partial—Congresses. I. Cifuentes, Patricio. II. Title.

QA403.C66 2008
515′.2433—dc22

2009036374

∞ The paper used in this book is acid-free and falls within the guidelines
established to ensure permanence and durability.
Visit the AMS home page at `http://www.ams.org/`

10 9 8 7 6 5 4 3 2 1 15 14 13 12 11 10

Introduction

This volume contains the Proceedings of the 8th International Conference on Harmonic Analysis and Partial Differential Equations, which took place in El Escorial, Madrid (Spain) during the week of June 16-20, 2008.

The celebration of this El Escorial 2008 Conference continues a tradition initiated in 1979 thanks to the leadership and enthusiasm of Professor Miguel de Guzmán. The purpose of that first Conference was to bring together the best mathematicians in the field and let them show the progress in the area to a wide audience of senior and – more importantly – young researchers. The success of El Escorial 1979 Conference, which counted among the main speakers Alberto Calderón, Ronald Coifman, Yves Meyer and Peter Jones, inspired a group of people in the Department of Mathematics of the Universidad Autónoma de Madrid to continue to hold an International Conference with the same aim, (almost) every four years. From that first Conference of 1979, another seven have been held in 1983, 1987, 1992, 1996, 2000, 2004 and 2008. The El Escorial Conferences have kept growing in size and impact and are by now a very valuable fixed point on the mathematical calendar taking place every olympic year.

The format of these Conferences has always been essentially the same, and it has established a model that counts with general approval within the mathematical community and has been adopted by many conferences in different fields of Mathematics. Four outstanding mathematicians in the field of Harmonic Analysis and Partial Differential Equations are invited to teach one mini-course each. These mini-courses are intended to present in three or four one-hour sessions the state of the art in some topic of current interest, assuming minimal background from the audience and reaching the level of present-day research in such a way as to be useful for young researchers seeking to join teams doing high quality, original work. Another important part of the Conference consists of some 15 to 20 invited one-hour lectures, which can be of a more specialized nature. Finally, a few sessions of short talks are scheduled to provide an opportunity for those participants who want to present their latest results.

These Proceedings contain the written versions of two of the four mini-courses given this time at the Conference, namely, that of Steve Hofmann on "Local T(b) Theorems and Applications in Partial Differential Equations"and the survey of Carlos E. Kenig about "The global behavior of solutions to critical nonlinear dispersive and wave equations". R. DeVore, who gave a mini-course at El Escorial 2008 on "The Mathematical Foundations of Compressed Sensing", has chosen to present in these Proceedings his paper on "Instance Optimal Decoding by Thresholding in Compressed Sensing", written in collaboration with Albert Cohen and Wolfgang Dahmen. Also in these Proceedings one can find the contributions of most of the other invited speakers. The topics of these Contributed Lectures cover a wide range of areas within Harmonic Analysis and Partial Differential Equations and illustrate well the fruitful interplay between the two subfields.

The Proceedings of all the El Escorial Conferences have been published in different Mathematical journals of wide circulation. We consider the publication of the Proceedings an essential part of the Conference, the very final act and the starting point of the process to prepare the next El Escorial Conference. In this occasion, we want to thank the American Mathematical Society for its help in publishing the Proceedings of El Escorial 2008 in its "Contemporary Mathematics"series. It is also proper to thank the institutions that have helped financially with the organization of El Escorial 2008, namely, the Spanish Ministries of Education and Science, the Universidad Autónoma de Madrid, the Consejo Superior de Investigaciones Científicas, the project Consolider I-Math and the Real Sociedad Matemática Española. The next El Escorial Conference will be held in 2012.

The Organizing Committee
Madrid, July 2009

Titles in This Series

505 **Patricio Cifuentes, José García-Cuerva, Gustavo Garrigós, Eugenio Hernández, José María Martell, Javier Parcet, Alberto Ruiz, Fernando Soria, José Luis Torrea, and Ana Vargas, Editors,** Harmonic analysis and partial differential equations, 2010

504 **Christian Ausoni, Kathryn Hess, and Jérôme Scherer, Editors,** Alpine perspectives on algebraic topology, 2009

503 **Marcel de Jeu, Sergei Silvestrov, Christian Skau, and Jun Tomiyama, Editors,** Operator structures and dynamical systems, 2009

502 **Viviana Ene and Ezra Miller, Editors,** Combinatorial Aspects of Commutative Algebra, 2009

501 **Karel Dekimpe, Paul Igodt, and Alain Valette, Editors,** Discrete groups and geometric structures, 2009

500 **Philippe Briet, François Germinet, and Georgi Raikov, Editors,** Spectral and scattering theory for quantum magnetic systems, 2009

499 **Antonio Giambruno, César Polcino Milies, and Sudarshan K. Sehgal, Editors,** Groups, rings and group rings, 2009

498 **Nicolau C. Saldanha, Lawrence Conlon, Rémi Langevin, Takashi Tsuboi, and Pawel Walczak, Editors,** Foliations, geometry and topology, 2009

497 **Maarten Bergvelt, Gaywalee Yamskulna, and Wenhua Zhao, Editors,** Vertex operator algebras and related areas, 2009

496 **Daniel J. Bates, GianMario Besana, Sandra Di Rocco, and Charles W. Wampler, Editors,** Interactions of classical and numerical algebraic geometry, 2009

495 **G. L. Litvinov and S. N. Sergeev, Editors,** Tropical and idempotent mathematics, 2009

494 **Habib Ammari and Hyeonbae Kang, Editors,** Imaging microstructures: Mathematical and computational challenges, 2009

493 **Ricardo Baeza, Wai Kiu Chan, Detlev W. Hoffmann, and Rainer Schulze-Pillot, Editors,** Quadratic Forms—Algebra, Arithmetic, and Geometry, 2009

492 **Fernando Giráldez and Miguel A. Herrero, Editors,** Mathematics, Developmental Biology and Tumour Growth, 2009

491 **Carolyn S. Gordon, Juan Tirao, Jorge A. Vargas, and Joseph A. Wolf, Editors,** New developments in Lie theory and geometry, 2009

490 **Donald Babbitt, Vyjayanthi Chari, and Rita Fioresi, Editors,** Symmetry in mathematics and physics, 2009

489 **David Ginzburg, Erez Lapid, and David Soudry, Editors,** Automorphic Forms and L-functions II. Local aspects, 2009

488 **David Ginzburg, Erez Lapid, and David Soudry, Editors,** Automorphic forms and L-functions I. Global aspects, 2009

487 **Gilles Lachaud, Christophe Ritzenthaler, and Michael A. Tsfasman, Editors,** Arithmetic, geometry, cryptography and coding theory, 2009

486 **Frédéric Mynard and Elliott Pearl, Editors,** Beyond topology, 2009

485 **Idris Assani, Editor,** Ergodic theory, 2009

484 **Motoko Kotani, Hisashi Naito, and Tatsuya Tate, Editors,** Spectral analysis in geometry and number theory, 2009

483 **Vyacheslav Futorny, Victor Kac, Iryna Kashuba, and Efim Zelmanov, Editors,** Algebras, representations and applications, 2009

482 **Kazem Mahdavi and Deborah Koslover, Editors,** Advances in quantum computation, 2009

481 **Aydın Aytuna, Reinhold Meise, Tosun Terzioğlu, and Dietmar Vogt, Editors,** Functional analysis and complex analysis, 2009

480 **Nguyen Viet Dung, Franco Guerriero, Lakhdar Hammoudi, and Pramod Kanwar, Editors,** Rings, modules and representations, 2008

Contents

Contemporary Mathematics
Volume 505, 2010

Instance Optimal Decoding by Thresholding in Compressed Sensing

Albert Cohen, Wolfgang Dahmen, and Ronald DeVore

ABSTRACT. Compressed Sensing seeks to capture a discrete signal $x \in \mathbb{R}^N$ with a small number n of linear measurements. The information captured about x from such measurements is given by the vector $y = \Phi x \in \mathbb{R}^n$ where Φ is an $n \times N$ matrix. The best matrices, from the viewpoint of capturing sparse or compressible signals, are generated by random processes, e.g. their entries are given by i.i.d. Bernoulli or Gaussian random variables. The information y holds about x is extracted by a decoder Δ mapping $\mathbb{R}^n$ into $\mathbb{R}^N$. Typical decoders are based on ℓ_1-minimization and greedy pursuit. The present paper studies the performance of decoders based on thresholding. For quite general random families of matrices Φ, decoders Δ are constructed which are instance-optimal in probability by which we mean the following. If x is any vector in $\mathbb{R}^N$, then with high probability applying Δ to $y = \Phi x$ gives a vector $\bar{x} := \Delta(y)$ such that $\|x - \bar{x}\| \le C_0 \sigma_k(x)_{\ell_2}$ for all $k \le an/\log N$ provided a is sufficiently small (depending on the probability of failure). Here $\sigma_k(x)_{\ell_2}$ is the error that results when x is approximated by the k sparse vector which equals x in its k largest coordinates and is otherwise zero. It is also shown that results of this type continue to hold even if the measurement vector y is corrupted by additive noise: $y = \Phi x + e$ where e is some noise vector. In this case $\sigma_k(x)_{\ell_2}$ is replaced by $\sigma_k(x)_{\ell_2} + \|e\|_{\ell_2}$.

1. Introduction

1.1. Background. The typical paradigm for acquiring a compressed representation of a discrete signal $x \in \mathbb{R}^N$, N large, is to choose an appropriate basis, compute all of the coefficients of x in this basis, and then retain only the k largest of these with $k < N$. Without loss of generality, we can assume that the appropriate basis is the canonical Kroenecker delta basis. If $S_k \subset \{1, \dots, N\}$ denotes a set of indices corresponding to k largest entries in x, then x_{S_k} is the compressed

2000 *Mathematics Subject Classification.* 94A12, 65C99, 68P30, 41A25, 15A52.
Key words and phrases. Compressed sensing, best k-term approximation, instance optimal decoders, thresholding, noisy measurements, random matrices.
This research was supported by the Office of Naval Research Contracts ONR-N00014-08-1-1113, ONR N00014-05-1-0715; the ARO/DoD Contracts W911NF-05-1-0227 and W911NF-07-1-0185; the NSF Grant DMS-0810869; the Leibniz-Programme of DFG; and the French-German PROCOPE contract 11418YB.

approximation to x. Here and throughout this paper, for a set T of indices, we denote by x_T the vector which is identical to x on T but is zero outside T.

For any ℓ_p norm, this approximation process is equivalent to best k-term approximation. Namely, if

$$\Sigma_k := \{z \in \mathbb{R}^N : \#(\mathrm{supp}(z)) \le k\}, \tag{1.1}$$

where $\mathrm{supp}(z)$ is the set of those indices corresponding to the nonzero entries in z, and if for any norm $\|\cdot\|_X$ on $\mathbb{R}^N$, we define

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

then $\|x - x_{S_k}\|_{\ell_p} = \|x_{S_k^c}\|_{\ell_p} = \sigma_k(x)_{\ell_p}$. That is, x_{S_k} is a best approximation to x from Σ_k. This approximation process should be considered as *adaptive* since the indices of those coefficients which are retained vary from one signal to another.

Since, in the end, we retain only k entries of x in the above compression paradigm, it seems wasteful to initially make N measurements. The theory of *compressed sensing* as formulated by Candes, Romberg and Tao [**8, 9**] and by Donoho [**14**], asks whether it is possible to actually make a number n of *non-adaptive* linear measurements, with n comparable to k, and still retain the necessary information about x in order to build a good compressed approximation. These measurements are represented by a vector

$$y = \Phi x, \tag{1.3}$$

of dimension $n < N$ where Φ is an $n \times N$ measurement matrix (called a CS matrix). To extract the information that the measurement vector y holds about x, one uses a decoder Δ which is a mapping from $\mathbb{R}^n$ into $\mathbb{R}^N$. The vector $x^* := \Delta(y) = \Delta(\Phi x)$ is our approximation to x extracted from the information y. In contrast to Φ, the operator Δ is allowed to be non-linear.

In recent years, considerable progress has been made in understanding the performance of various choices of the measurement matrices Φ and decoders Δ. Although not exclusively, by far most contributions focus on the ability of such an encoder-decoder pair (Φ, Δ) to recover a *sparse* signal. For example, a typical theorem says that there are pairs (Φ, Δ) such that whenever $x \in \Sigma_k$, with $k \le an/\log(N/k)$, then $x^* = x$.

From both a theoretical and a practical perspective, it is highly desirable to have pairs (Φ, Δ) that are robust in the sense that they are effective even when the vector x is not assumed to be sparse. The question arises as to how we should measure the effectiveness of such an encoder-decoder pair (Φ, Δ) for non-sparse vectors. In [**6**] we have proposed to measure such performance in a metric $\|\cdot\|_X$ by the largest value of k for which

$$\|x - \Delta(\Phi x)\|_X \le C_0 \sigma_k(x)_X, \quad \forall x \in \mathbb{R}^N, \tag{1.4}$$

with C_0 a constant independent of k, n, N. We say that a pair (Φ, Δ) which satisfies property (1.4) is *instance-optimal* of order k with constant C_0. It was shown that this measure of performance heavily depends on the norm employed to measure error. Let us illustrate this by two contrasting results from [**6**]:

(i) If $\|\cdot\|_X$ is the ℓ_1-norm, it is possible to build encoding-decoding pairs (Φ, Δ) which are instance-optimal of order k with a suitable constant

C_0 whenever $n \geq ck\log(N/k)$ provided c and C_0 are sufficiently large. Moreover the decoder Δ can be taken as

$$\Delta(y) := \operatorname*{argmin}_{\Phi z = y} \|z\|_{\ell_1}. \tag{1.5}$$

Therefore, in order to obtain the accuracy of k-term approximation, the number n of non-adaptive measurements need only exceed the amount k of adaptive measurements by the small factor $c\log(N/k)$. We shall speak of the range of k which satisfy $k \leq an/\log(N/k)$ as the *large range* since it is the largest range of k for which instance-optimality can hold.

(ii) In the case $\|\cdot\|_X$ is the ℓ_2-norm, if (Φ, Δ) is any encoding-decoding pair which is instance-optimal of order $k = 1$ with a fixed constant C_0, then the number of measurement n is always larger than aN where $a > 0$ depends only on C_0. Therefore, the number of non-adaptive measurements has to be very large in order to compete with even one single adaptive measurement.

The matrices Φ which have the largest range of instance-optimality for ℓ_1 are all given by stochastic constructions. Namely, one creates an appropriate random family $\Phi(\omega)$ of $n \times N$ matrices on a probability space (Ω, ρ) and then shows that with high probability on the draw, the resulting matrix $\Phi = \Phi(\omega)$ will satisfy instance-optimality for the large range of k. There are no known deterministic constructions. The situation is even worse in the sense that given an $n \times N$ matrix Φ there is no simple method for checking its range of instance-optimality.

While the above results show that instance-optimality is not a viable concept in ℓ_2, it turns out that the situation is not as bleak as it seems. For example, a more optimistic result was established by Candes, Romberg and Tao in [**9**]. They show that if $n \geq ck\log(N/k)$ it is possible to build pairs (Φ, Δ) such that for all $x \in \mathbb{R}^N$,

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \frac{\sigma_k(x)_{\ell_1}}{\sqrt{k}}, \tag{1.6}$$

with the decoder again defined by (1.5). This implies in particular that k-sparse signals are exactly reconstructed and that signals x in the space weak ℓ_p (denoted by $w\ell_p$) with $\|x\|_{w\ell_p} \leq M$ for some $p < 1$ are reconstructed with accuracy $C_0 M k^{-s}$ with $s = 1/p - 1/2$. This bound is of the same order as the best estimate available on $\max\{\sigma_k(x)_{\ell_2} : \|x\|_{w\ell_p} \leq M\}$. Of course, this result still falls short of instance-optimality in ℓ_2 as it must.

The starting point of the present paper is the intriguing fact, that instance-optimality can be attained in ℓ_2 if one accepts a probabilistic statement. A first result in this direction, obtained by Cormode and Mutukrishnan in [**7**], shows how to construct random $n \times N$ matrices $\Phi(\omega)$ and a decoder $\Delta = \Delta(\omega)$, $\omega \in \Omega$, such that for any $x \in \mathbb{R}^N$,

$$\|x - \Delta(\Phi x)\|_{\ell_2} \leq C_0 \sigma_k(x)_{\ell_2} \tag{1.7}$$

holds with overwhelming probability (larger than $1 - \epsilon(n)$ where $\epsilon(n)$ tends rapidly to 0 as $n \to +\infty$) as long as $k \leq an/(\log N)^{5/2}$ with a suitably small. Note that this result says that given x, the set of $\omega \in \Omega$ for which (1.7) fails to hold has small measure. This set of failure will depend on x.

From our viewpoint, *instance-optimality in probability* is the proper formulation in ℓ_2. Indeed, even in the more favorable setting of ℓ_1, we can never put our hands on matrices Φ which have the large range of instance-optimality. We only know with high probability on the draw, in certain random constructions, that we can attain instance-optimality. So the situation in ℓ_2 is not that much different from that in ℓ_1.

The results in [**6**] pertaining to instance-optimality in probability asked two fundamental questions: (i) can we attain instance-optimality for the largest range of k, i.e. $k \leq an/\log(N/k)$, and (ii) what are the properties of random families that are needed to attain this performance. We showed that instance-optimality can be obtained in the probabilistic setting for the largest range of k, i.e. $k \leq an/\log(N/k)$ using quite general constructions of random matrices. Namely, we introduced two properties for a random matrix Φ which ensure instance-optimality in the above sense and then showed that these two properties hold for rather general constructions of random matrices (such as Gaussian and Bernoulli). However, one shortcoming of the results in [**6**] is that the decoder used in establishing instance-optimality was defined by minimizing $\|y - \Phi x\|_{\ell_2}$ over all k-sparse vectors, a task which cannot be achieved in any reasonable computational time.

1.2. Objectives. In the present paper, we shall be interested in which practical decoders can be used with a general random family so as to give a sensing system which has instance-optimality in probability for ℓ_2 for the largest range of k. The first result in this direction was given by Wojtasczcek [**23**] who has shown that ℓ_1-minimization can be used with Gaussian random matrices to attain instance-optimality for this large range of k. This result was recently generalized in [**12**] to arbitrary random families in which the entries of the matrix are generated by independent draws of a sub-Gaussian random variable. This result includes Bernoulli matrices whose entries take the values $\pm 1/\sqrt{n}$.

The problem of decoding in compressed sensing, as well as for more general inverse problems, is a very active area of research. In addition to ℓ_1-minimization and its efficient implementation, several alternatives have been suggested as being possibly more efficient. These include decoding based on greedy procedures such as Orthogonal Matching Pursuit (OMP) (see [**15, 19, 20, 21**]) as well as decoding through weighted least squares [**11**]. Some of the pertinent issues in analyzing a decoding method is the efficiency of the method (number of computations) and the required storage needed.

Concerning efficiency, Gilbert and Tropp [**15**] have proposed to use a greedy procedure, known as Orthogonal Matching Pursuit (OMP) algorithm, in order to define $\Delta(y)$. The greedy algorithm identifies a set of Λ of column indices which can be used to decode y. Taking zero as an initial guess, successive approximations to y are formed by orthogonally projecting the measurement vector y onto the span of certain incrementally selected columns ϕ_j of Φ. In each step, the current set of columns is expanded by one further column that maximizes the modulus of the inner product with the current residual. The following striking result was proved in [**15**] for a probabilistic setting for general random matrices which include the Bernoulli and Gaussian families: if $n \geq ck \log N$ with c sufficiently large, then for any k sparse vector x, the OMP algorithm returns exactly $x^k = x$ after k iterations, with probability greater than $1 - N^{-b}$ where b can be made arbitrarily large by taking c large enough.

Decoders like OMP are of high interest because of their efficiency. The above result of Gilbert and Tropp remains as the only general statement about OMP in the probabilistic setting. A significant breakthrough on decoding using greedy pursuit was given in the paper of Needell and Vershynin [**19**] (see also their followup [**20**]) where they showed the advantage of adjoining a batch of coordinates at each iteration rather than just one coordinate as in OMP. They show that such algorithms can deterministically capture sparse vectors for a slightly smaller range than the large range of k.

The present paper examines decoders based on thresholding and asks whether such algorithms can be used as decoders to yield ℓ_2 instance-optimality in probability for general families of random matrices. We will describe in Section 6 a greedy thresholding scheme, referred to as **SThresh**, and prove that it gives instance-optimality in probability in ℓ_2 for the large range of k. This algorithm adds a batch of coordinates at each iteration and then uses a thinning procedure to possibly remove some of them at later iterations. Conceptually, one thinks in terms of a bucket holding all of the coordinates to be used in the construction of x. In the analysis of such algorithms it is important to not allow more than a multiple of k coordinates to gather in the bucket. The thinning is used for this purpose.

While preparing this paper, we became aware of the work of Needell and Tropp [**21**] in which they develop a deterministic algorithm (called COSAMP) which has features similar to ours. In fact, we have employed some of the ideas of that paper in our analysis. This will be discussed in more detail after we give a precise description of our algorithm.

While the benchmark of instance-optimality covers the case of an input signal x which is a perturbation of a sparse signal, it is not quite appropriate for dealing with possible noise in the measurements. By this we mean that instead of measuring Φx, our measurement vector y is of the form

$$y = \Phi x + e, \tag{1.8}$$

with $e \in \mathbb{R}^n$ a noise vector. **SThresh** will also perform well in this noisy setting. Stability under noisy measurements has been also established for COSAMP ([**21**]) as well as for schemes based on ℓ_1-regularization [**9**]. While this latter strategy requires a-priori knowledge about the noise level, this is not the case for COSAMP and the schemes developed in this paper.

A brief overview of our paper is the following. In the next section, we introduce the probabilistic properties we will require of our random families. In §3, we introduce a deterministic algorithm based on thresholding and analyze its performance. This algorithm is then used as a basic step in the greedy decoding algorithm for stochastic families in §4. In this section, we prove that the stochastic decoding algorithm gives instance optimality in probability. As we have noted above, a key step in this decoding is a thinning of the indices placed into the bucket. It is an intriguing question whether this thinning is actually necessary. This leads us to consider an algorithm without thinning. We introduce such an algorithm in §6 and we show in §7 that almost gives instance-optimality in probability for ℓ_2 for the large range of k. The results for that algorithm are weaker than the thinning algorithms in two ways. First they require the addition of a small term ϵ to $\sigma_k(x)_{\ell_2}$ and secondly the range of k is slightly smaller than the large range. Finally, we append in §8 the proof that random matrices whose columns are uniformly distributed vectors on the unit sphere satisfy the properties which are used in the analysis of both algorithms. These properties are known to hold for matrices whose entries are i.i.d. draws from Gaussian or Bernoulli random variables.

While a lot of progress has been made on understanding the performance of greedy algorithms for decoding in compressed sensing, there remain fundamental unsettled questions. The most prominent is whether the original OMP algorithm can indeed give instance optimality in probability for ℓ_2 for the large range of k.

2. The Setting

As we have already mentioned, one of our goals is to derive results that hold for general random families. In this section, we state general properties of random families which will be used as assumptions in our theorems.

We consider random $n \times N$ matrices $\Phi = \Phi(\omega)$, on a probability space (Ω, ρ). We denote the entries in Φ by $\phi_{i,j}$, $1 \le i \le n$, $1 \le j \le N$ and denote the j-th column of Φ by ϕ_j, $j = 1, \dots, N$. One of the main properties needed of random families for compressed sensing is that given any $x \in \mathbb{R}^N$, with high probability Φx has norm comparable to that of x. We formulate this in

P1: For any $x \in \mathbb{R}^N$ and $\delta > 0$, there is a set $\Omega_1(x, \delta) \subset \Omega$ such that

$$|\|\Phi x\|_{\ell_2}^2 - \|x\|_{\ell_2}^2|\text{æ} \le \delta \|x\|_{\ell_2}^2, \quad \omega \in \Omega_1(x, \delta), \tag{2.1}$$

and

$$\rho(\Omega_1^c(x, \delta)) \le b_1 e^{-c_1 n \delta^2}, \tag{2.2}$$

where b_1 and c_1 are absolute constants.

An important *consequence* of property **P1**, often used in compressed sensing, is the following *Restricted Isometry Property* (**RIP**), as formulated by Candes and Tao [**8**]:

RIP(k, η): *An $n \times N$ matrix Φ_0 is said to satisfy the Restricted Isometry Property of order m with constant $\eta \in (0, 1)$, if*

$$(1 - \eta)\|x\|^2 \le \|\Phi_0 x\|^2 \le (1 + \eta)\|x\|^2, \ \ x \in \Sigma_m. \tag{2.3}$$

It was shown in [**3**] that **P1** implies **RIP**. More precisely, their analysis gives the following fact.

PROPOSITION 2.1. *Whenever the random family $\mathbf{\Phi} = \{\Phi = \Phi(\omega) : \omega \in \Omega\}$ of $n \times N$ matrices satisfies* **P1**, *then for each $\eta \in (0, 1)$ there exists a subset $\Omega_0(m, \eta, \Phi) \subset \Omega$ with*

$$\rho(\Omega_0(m, \eta, \Phi)^c) \le b_1 e^{-\frac{c_1 n \eta^2}{4} + m\left[\log(eN/m) + \log(12/\eta)\right]} \tag{2.4}$$

where b_1, c_1 are the constants from **P1**, *such that for each draw $\omega \in \Omega_0(m, \eta, \mathbf{\Phi})$ the matrix $\Phi(\omega) \in \mathbf{\Phi}$ satisfies* **RIP**(m, η) *(order m with constant η). In particular, given η, if a is chosen suitably small (depending on η) then with high probability $\mathbf{\Phi}$ will satisfy* **RIP**(m, η) *as long as $m \le an/\log(N/m)$, i.e for the large range of m.*

3. A deterministic thresholding algorithm

In this section, we shall introduce a deterministic thresholding algorithm. Later, we shall embed this algorithm into the probabilistic setting and show that the corresponding probabilistic algorithm has ℓ_2 instance optimality in probability.

We continue to denote by k the envisaged range of instance optimality. We shall assume throughout this section that Φ is an $n \times N$ compressed sensing matrix that satisfies the $\mathbf{RIP}(m, \eta)$ where $m \geq 3k$ is an integer which will be specified later. For the validity of the theorems that follow, there will also be a restriction that η is sufficiently close to 0.

3.1. Description of the thresholding algorithm and main result. In this section, we shall describe our thresholding algorithm. The algorithm starts with an input vector $y \in \mathbb{R}^n$ and generates a set Λ of at most k indices. The input vector y is either $y = \Phi x$ in the noiseless case or $y = \Phi x + e$ in the presence of noise e in the measurements. The output of the algorithm is a vector x^* which is an approximation to x determined by the noisy information y.

We now describe our thresholding algorithm for decoding an input vector $v \in \mathbb{R}^n$ of either type:

$\mathbf{DThresh}[v, k, \delta] \to x^*$

(i) Fix a thresholding parameter $\delta > 0$. Choose the sparsity index k, let $r^0 := v$, $x^0 := 0$, and set $j = 0$, $\Lambda_0 = \bar{\Lambda}_0 = \emptyset$.

(ii) If $j = k$ stop and set $x^* := x^j$.

(iii) Given Λ_j calculate the residual $r^j := v - \Phi x^j$ for the input vector v and define

$$\tilde{\Lambda}_{j+1} := \{i \in \{1, \ldots, N\} : |\langle r^j, \phi_i \rangle| \geq \frac{\delta \|r^j\|}{\sqrt{k}}\}$$

If $\tilde{\Lambda}_{j+1} = \emptyset$, stop and output $\Lambda^* = \Lambda_j$ and $x^* := x^j$.
Otherwise set $\bar{\Lambda}_{j+1} := \Lambda_j \cup \tilde{\Lambda}_{j+1}$.

(iv) Compute $\hat{x}(\bar{\Lambda}_{j+1})$ (according to (5.13)) as

$$\hat{x}(\bar{\Lambda}_{j+1}) = \mathrm{argmin}_{\mathrm{supp}(z) \subseteq \bar{\Lambda}_{j+1}} \|\Phi z - v\|,$$

and define Λ_{j+1} as the set of indices $\nu \in \bar{\Lambda}_{j+1}$ corresponding to the k largest (in absolute value) entries in $\hat{x}(\bar{\Lambda}_{j+1})$. Let $x^{j+1} := \hat{x}(\bar{\Lambda}_{j+1})_{\Lambda_{j+1}}$, $j + 1 \to j$ and return to (ii).

Step (iv) is a thinning step which prevents the bucket of indices to get too large so that in our analysis $\mathbf{RIP}(\eta, m)$ will turn out to remain applicable for a fixed suitable multiple m of k.

Perhaps a few remarks concerning a comparison with COSAMP are in order. In both schemes any a priori knowledge about the noise level is *not* needed but the envisaged sparsity range k appears as a parameter in the scheme. This is in contrast to ℓ_1-regularization in [**9**] which, however, does seem to require a priori knowledge about the noise level. Of course, one can take k as the largest value for which the scheme can be shown to perform well. The subsequent analysis will show that this is indeed the case for the maximal range.

While **DThresh** as well as COSAMP are based on thresholding, COSAMP from the very beginning always works with least squares projections of size $2k$. In the above scheme the sets of active indices Λ_j are allowed to grow and, in fact, the scheme may terminate before they ever reach size k.

The following theorem summarizes the convergence properties of **DThresh**.

THEOREM 3.1. *Assume that $\delta, \eta \leq 1/32$ and that the matrix Φ satisfies* **RIP**(m, η) *with $m \geq \lceil k(1 + \frac{3}{2\delta^2})\rceil$. Then for any $x \in \mathbb{R}^N$ and $y = \Phi x + e$ the output x^* of* **DThresh**$[y, k, \delta]$ *has the following properties:*

(i) *If in addition $x \in \Sigma_k$, then the output x^* satisfies*

$$\|x - x^*\| \leq 90\|e\|. \tag{3.1}$$

(ii) *If $x \in \mathbb{R}^N$ and x_{S_k} is its best approximation from Σ_k, i.e. the indices in S_k identify the k largest terms (in absolute value) in x, then*

$$\|x - x^*\| \leq 90[\|\Phi(x - x_{S_k})\| + \|e\|]. \tag{3.2}$$

(iii) *For arbitrary $x \in \mathbb{R}^N$, one has*

$$\|x - x^*\| \leq 90\Big((1+\eta)^{1/2}\Big(\frac{\sigma_k(x)_{\ell_1^N}}{\sqrt{k}} + \sigma_k(x)_{\ell_2^N}\Big) + \|e\|\Big). \tag{3.3}$$

We postpone the proof of Theorem 3.1 to §5 and explain first its ramifications in the stochastic setting.

4. Thresholding in the stochastic setting

Let us now assume that $\mathbf{\Phi} = \{\Phi(\omega) : \omega \in \Omega\}$ is a random family of matrices which satisfy **P1**. As we have shown in Proposition 2.1, with high probability on the draw (see (2.4)), $\Phi(\omega)$ will satisfy **RIP**(m, η), m a fixed multiple of k, for the large range of k, with constant a depending on that multiple and on η. We shall use the following stochastic version **SThresh** of the thresholding algorithm which differs from **DThresh** only in the initialization step (i).

SThresh$[v, k, \delta] \to x^*$

(i) Fix a thresholding parameter $\delta > 0$ and the sparsity index k. Given any signal $x \in \mathbb{R}^N$ take a random draw $\Phi = \Phi(\omega)$ and consider as input the measurement vector $v = \Phi x + e \in \mathbb{R}^n$ where e is a noise vector. Let $r^0 := v$, and set $j = 0$, $\Lambda_0 = \bar{\Lambda}_0 = \emptyset$.

(ii) If $j = k$ stop and set $x^* := x^j$.

(iii) Given Λ_j calculate the residual $r^j := v - \Phi x^j$ for the input vector v and define

$$\tilde{\Lambda}_{j+1} := \{i \in \{1, \dots, N\} : |\langle r^j, \phi_i\rangle| \geq \frac{\delta\|r^j\|}{\sqrt{k}}\}$$

If $\tilde{\Lambda}_{j+1} = \emptyset$, stop and output $\Lambda^* = \Lambda_j$ and $x^* := x^j$. Otherwise set $\bar{\Lambda}_{j+1} := \Lambda_j \cup \tilde{\Lambda}_{j+1}$.

(iv) Compute $\hat{x}(\bar{\Lambda}_{j+1})$ (according to (5.13)) as

$$\hat{x}(\bar{\Lambda}_{j+1}) = \operatorname{argmin}_{\operatorname{supp}(z) \subseteq \bar{\Lambda}_{j+1}} \|\Phi z - v\|,$$

and define Λ_{j+1} as the set of indices $\nu \in \bar{\Lambda}_{j+1}$ corresponding to the k largest (in absolute value) entries in $\hat{x}(\bar{\Lambda}_{j+1})$. Let $x^{j+1} := \hat{x}(\bar{\Lambda}_{j+1})_{\Lambda_{j+1}}$, $j + 1 \to j$ and return to (ii).

Notice that the output $x^* = x^*(\omega)$ is stochastic. From the analysis of the previous section, we can deduce the following theorem.

THEOREM 4.1. *Assume that $\delta \le 1/63$ in* **SThresh** *and that the family* $\mathbf{\Phi}$ *of stochastic matrices $\Phi(\omega)$ has property* **P1**. *Then, for any $x \in \mathbb{R}^N$ there exists a subset $\Omega(x)$ of Ω with*

$$\rho(\Omega(x)^c) \le 2b_1 e^{-c_1 n/8\cdot 63^2}, \tag{4.1}$$

such that for any $\omega \in \Omega(x)$ and measurements of the form $y = \Phi(\omega)x + e$, with $e \in \mathbb{R}^n$ a noise vector, the output x^* of **SThresh**$[y, \delta, k]$ *satisfies*

$$\|x - x^*\| \le C\sigma_k(x) + 90\|e\|, \quad k \le an/\log(N/n), \tag{4.2}$$

with $C \le 92$ and a depending only on δ, c_1 and the bound on η.

In particular, when $e = 0$ this algorithm is instance-optimal in probability in ℓ_2 for the large range of k.

Proof: Fixing $\eta = 1/63$ and $m = \lceil (1 + \frac{3}{2\delta^2})k \rceil$ we know by Proposition 2.1 that there exists a set $\Omega_0 \subset \Omega$ such that for $\omega \in \Omega_0$ the matrix $\Phi = \Phi(\omega)$ satisfies $\mathbf{RIP}(m, 1/63)$ and

$$\rho(\Omega_0^c) \le b_1 e^{-\frac{c_1 n}{4\cdot 63^2} + m\left[\log 756 + \log(eN/m)\right]}. \tag{4.3}$$

Thus, as long as $N \ge 756m/e$ it suffices to have $2m\log(eN/m) \le c_1 n/8 \cdot 63^2$, to ensure that

$$\rho(\Omega_0^c) \le b_1 e^{-\frac{c_1 n}{8\cdot 63^2}}, \quad \text{whenever} \quad k \le an/\log(N/k) \tag{4.4}$$

provided a is sufficiently large. Thus, we infer from Theorem 3.1 (ii) that

$$\|x - x^*\| \le 90(\|\Phi(x - x_{S_k})\| + \|e\|) \tag{4.5}$$

holds for every $\omega \in \Omega_0$. Now, by Property **P1**, there exists a subset $\Omega_1(x^c_{S_k}, 1/63)$ with complement

$$\rho(\Omega_1(x^c_{S_k}, 1/63)^c) \le b_1 e^{-c_1 n/63^2},$$

such that $\|\Phi(x - x_{S_k})\| \le 1.013\|x - x_{S_k}\|$ which ensures the validity of (4.2) with $\Omega(x) := \Omega_0 \cap \Omega_1(x^c_{S_k}, 1/63)$. □

5. Proof of Theorem 3.1

We begin by collecting a few prerequisites.

5.1. Consequences of RIP. Let us first record some simple results that follow from the $\mathbf{RIP}(m, \eta)$ assumption. Most of the results we state in this subsection can be found in [**19**] but we include their simple proofs for completeness of the present paper.

LEMMA 5.1. *For any $I \subset \{1, \dots, N\}$ with $\#(I) \le m$ we have*

$$\|\Phi_I^*\|^2 = \|\Phi_I\|^2 \le (1 + \eta). \tag{5.1}$$

Proof: The equality in (5.1) holds because the norm of a matrix and its conjugate transpose are identical (this follows for example from the fact that $\|A\| = \sup_{\|x\|=1, \|y\|=1} y^t Ax$). The upper inequality follows from the $\mathbf{RIP}(m, \eta)$ assumption because for any $x \in \mathbb{R}^N$, supported in I one has $\|\Phi_I x\| = \|\Phi x_I\| \le (1+\eta)^{1/2}\|x_I\| = (1 + \eta)^{1/2}\|x\|$. □