



Introduction to Compressed Sensing

With Coding Theoretic Perspective

This book is a course note developed for a graduate level course in Spring 2011, at GIST, Korea. The course aimed at introducing the topic of Compressed Sensing (CS). CS is considered as a new signal acquisition paradigm with which sample taking could be faster than what can be expected of the canonical approach. Namely, the number of signal samples sufficient to reproduce a given signal could be much smaller than the number of samples deemed sufficient under the Shannon Nyquist sampling theory. The CS theory is expected to influence many application areas with interruptive changes to their current practices in the years to come, including tomography, radars, communications, image and signal processing, and wireless sensor networks. In addition, we make note of the fact that the tenet of CS theory is equivalent to the parity-checking and syndrome decoding in the Channel Coding theory. On the one hand, this means that, wealth of information is available to solve the parity-check equation from Channel Coding theory which can be leveraged to understand the CS problem better; on the other hand, the new information being generated in the CS community can be utilized to provide new perspectives in advancing the Channel Coding theory.

Heung-No Lee

8/29/2011

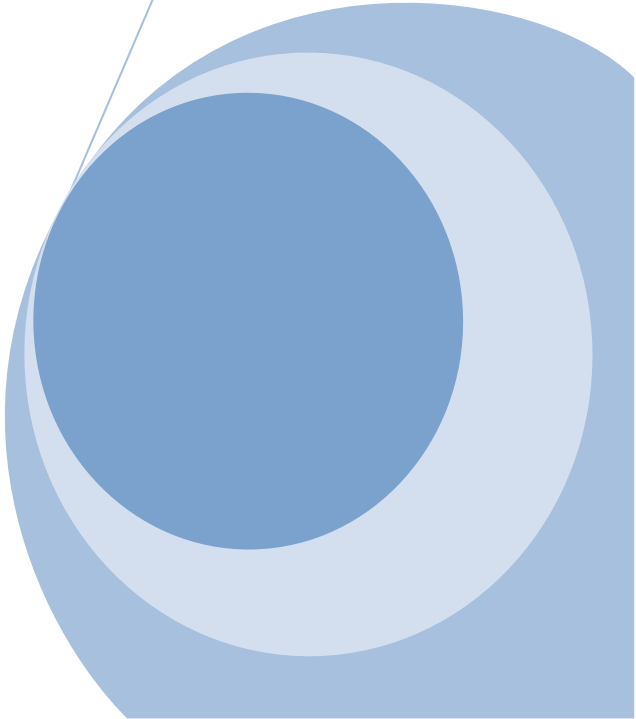


Table of Contents

Acknowledgement	4
Chapter I. Course Information	5
1. General Information.....	5
2. Course Syllabus	6
3. How to Cite this Note	8
4. Course Scope and Materials	8
5. The RICE University Repository.....	11
6. Applications.....	12
A. Single Pixel Cameras.....	12
B. Terahertz pulsed spectroscopic imaging.....	12
C. Other areas of applications	15
D. Summary of Applications	15
7. References.....	15
Chapter II. Compressed Sensing.....	16
1. Compressed Sensing, Compressive Sensing, Compressive Sampling.....	16
2. Pioneers of Compressed Sensing.....	16
3. Sampling Theorem and Dimensionality Reduction by Shannon.....	17
4. Compressed Sensing in a Nutshell	20
5. Compressed Sensing, explained with a little more care	22
A. Restricted Isometry Property	23
B. Incoherence Condition.....	24
C. Checking RIP is NP-hard.....	24
D. L0, L1, L2 norms and the null space	26
6. Summary.....	29
7. The Spark and The Singleton Bound.....	29
8. Matrix Design with Givens Rotations	30
9. Super Resolution Applications (Nano array filters and Nano lenses)	32
10. HW set #1	36
11. References for Chapter 2	38
Chapter III. Mathematics of Compressive Sensing.....	39
1. Uncertainty Principle and the L1 Uniqueness Proof	41
A. Representation via a Combined Dictionary.....	42
B. The Uncertainty Principle.....	44
C. The Theorem on the Uniqueness of the L0 Optimization	45
D. Uniqueness of the L1 Optimization.....	46
2. The Uniform Uncertainty Principle.....	52
A. Sufficient and Necessary Conditions for the Unique L0 Solution	53
B. Condition for the Unique L1 Solution.....	56
3. Uniqueness Proofs with Mutual Coherence μ	64
A. Comparison of $\theta_{K,K} + \theta_{K,2K}$ and $K\mu$	64
B. Connections to Other Results	66
4. Ensembles of Random Sensing Matrices	70
A. The $\log(N)$ factor for random ensembles.....	70
B. The $\log(N)$ factor, is it really needed for a random sensing matrix?	70
C. The $\log(N)$ factor, derived for binary K -sparse signals	72
5. Stable Recovery Property	74

6. The Chapter Problems	75
Chapter IV. Information Theoretic Consideration	79
1. K -sparse signal model	79
2. The Entropy of K -sparse signals	81
3. Mutual Information	83
4. Lower Bound on Probability of Sensing Error via Fano's Inequality	84
Chapter V. Sparse Recovery Algorithms	86
1. Linear Programming	86
2. Solving Linear Program via Lagrange Dual Interior Point Method	88
3. Solving Second Order Cone Programs with Log-Barrier Method	93
A. A log barrier function	94
B. The interior log barrier method for solving SOCPs	95
4. Homotopy Algorithms	99
A. Motivation	99
B. The Homotopy Problem	99
C. Two constraints from the subdifferential	102
D. The Homotopy Algorithm	103
E. Proof of the Theorem	107
5. Bayesian Compressive Sensing on the Graph Model	110
A. Iterative Compressive Sensing Algorithm	111
B. The Distribution of the Signal Value	112
C. The Distribution of the Signal Value in the Presence of Noise	115
D. Distribution of the Binary State Value	115
E. Useful Mathematical Identities for Log-Sum-Products	120
F. The Message Passing Algorithm	121
G. Donoho's message passing algorithm for compressed sensing	121
6. The Expander Graph Approach	122
7. Section References	123
8. Chapter Problems	124
Chapter VI. Recent Research Results	126
1. Deterministic Matrix Design	126
2. Coding Theoretic Approach	126
A. Compressed Sensing via Syndrome Decoding	126
B. The Sparks of Sensing Matrices over $GF(q)$	129
C. Signal Detection Algorithms	130
D. Simulation Results	132
E. Section Summary	133
F. Section References	133
3. Representation via Correlation vs. via Sparseness	134
Chapter VII. Review of Mathematical Results	135
1. Directional Derivatives, Subgradients, and Subdifferentials	135
2. Duality and Convex Optimization	139
3. Some Useful Linear Algebra Results	149
A. Some Facts on the Matrix Norm	149
B. Other useful matrix norm properties	149
Chapter VIII. References	151

Table of Figures

Figure 1: Fig. 2 of Shannon [1].....	19
Figure 2 : A chart from [15].	41
Figure 3 : A chart from [15].	42
Figure 4: An example of linear objective function $\min_x [1 \ -1]x$ s.t. $x^T x - 1 \leq 0$. Note that the optimal point is achieved at the perimeter of the feasible set.	89
Figure 5: The Concept of Placing Barriers At the Boundary of the Feasible Set	94
Figure 6: A log-barrier function	95
Figure 7: Variation of the penalized objective function as lambda is changing from 4 to 0.	100
Figure 8: Relation map of algorithms. From Donoho and Tsaig [18].	102
Figure 9: The Graph	112
Figure 10 : Gilbert-Varshamov Compressed Sensing Bounds for Sensing Matrices over $GF(q)$ and The Singleton Bound	129
Figure 11 : Simulation results of a $C(N=1200, d_s = 3, d_c = 6)$ code with different field sizes.	132
Figure 12: Hyperplane.....	135
Figure 13: Geometrical illustration of a subgradient of a convex function f . Note that the space is $n + 1$ dimensional.	136
Figure 14: Illustration of subgradients	137
Figure 15: Geometrical illustration of a subgradient of a convex function f . Note that the space is $n + 1$ dimensional.	138
Figure 16: The trajectory of $L(x, \lambda)$ with λ is varied from 0.4 to 4.	143
Figure 17: Illustration of the duality gap.....	143

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (Do-Yak Research Program, No. 2010-0017944)

This book of lecture notes has been developed under the generous support of National Research Foundation of Korea.

Contributors: There are numerous contributors to this book. They are the students and researchers in the INFONET lab. Students who have taken the course contributed providing feedbacks to the lecture materials, completing homework problems, and generating the best effort homework solution manuals. Some of the researchers in the lab have presented a summary of relevant papers at the INFONET journal club.

Chapter I. COURSE INFORMATION

1. General Information

Instructor: Heung-No Lee, Ph.D., Associate Professor, GIST, Korea.

Address: Gwangju Institute of Science and Technology, 261 Cheomdan Gwagiro, Gwangju, Republic of Korea.

Phone: (82) 62-715-2237, 3140.

E-mail: heungno@gist.ac.kr.

Home page: <http://infonet.gist.ac.kr/>

Open Source Policy

- The research trend is to moving towards “*reproducible researches.*”
- The homeworks and term projects submitted shall be reproducible as well.

2. Course Syllabus

I will use the following course schedule which will be posted at the course website at <http://infonet.gist.ac.kr/>.

Course Schedule

	Weekly Schedule	Remarks
1 st Week	GIST Entrance Ceremony	
2 nd Week 3/7, 3/9	Introduction to Compressed Sensing, Shannon Nyquist Sampling Theorem <ul style="list-style-type: none"> ● Richard Baraniuk, Compressive sensing. (IEEE Signal Processing Magazine, 24(4), pp. 118-121, July 2007) ● Justin Romberg, Imaging via compressive sampling. (IEEE Signal Processing Magazine, 25(2), pp. 14 - 20, March 2008) 	
3 rd Week 3/14, 3/16	Compressive Sensing Theory: L0, L1, L2 solutions	HW#1 Out
4 th Week 3/21, 23	Compressive Sensing Theory: L0 and L1 equivalence,	
5 th Week 3/28, 30	Compressive Sensing Theory: Generalized Uncertainty Principle, Sparse Representation, conditions for the unique L0 solution, and the unique L1 solution <ol style="list-style-type: none"> 1. D. Donoho and X. Huo, "Uncertainty Principles and Ideal Atomic Decomposition," IEEE Trans. on Info. Theory, vol.47, no.7, Nov. 2001. 2. M. Elad and A. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," IEEE Trans. Info. Theory, vol. 48, no. 9, Sept. 2002. 	HW#2 Out
6 th Week 4/4, 6	Compressive Sensing Theory: conditions for the L0 solution, and the unique L1 solution, the Candes-Tao's approach <ol style="list-style-type: none"> 1. Emmanuel Candès and Terence Tao, Decoding by linear programming. (IEEE Trans. on Information Theory, 51(12), pp. 4203 - 4215, December 2005) 2. Emmanuel Candès, Justin Romberg, and Terence Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. (IEEE Trans. on Information Theory, 52(2) pp. 489 - 509, February 2006) 	

	3. Emmanuel Candès and Terence Tao, “Near optimal signal recovery from random projections: Universal encoding strategies” (IEEE Trans. on Information Theory, 52(12), pp. 5406 - 5425, December 2006)	
7 th Week 4/11, 13	Sensing matrices and oversampling factors	HW#3 Out
8 th Week 4/18, 20	Stable Recovery	
9 th Week 4/25, 27	Midterm Exam	
10 th Week 5/2, 4	Recovery Algorithm I: Homotopy, LASSO, LARs, OMP	HW#4 Out
11 th Week 5/9, 11	Recovery Algorithm II: L1 minimization, Interior Point Methods, Log Barrier Methods	
12 th Week 5/16, 18	L1-Magic packages	HW#5 Out
13 th Week 5/23, 25	Bayesian Recovery Algorithm	
14 th Week 5/30, 6/1	Message Passing Algorithms: Support Set Recovery	HW#6 Out
15 th Week 6/6, 8	Memorial Day, Overview of the course	
16 th Week		Final Term Paper Due

3. How to Cite this Note

Please use the following line when you quote the materials in this book.

- Heung-No Lee, Introduction to Compressed Sensing (Lecture notes), Spring Semester, 2011.

4. Course Scope and Materials

The following materials will be discussed in class:

Please note that all of these papers are available just by clicking the link or at the RICE's Compressive Sensing web-site: <http://dsp.rice.edu/cs/>.

Introduction to Compressive Sensing

4. Richard Baraniuk, [Compressive sensing](#). (IEEE Signal Processing Magazine, 24(4), pp. 118-121, July 2007)
5. Justin Romberg, [Imaging via compressive sampling](#). (IEEE Signal Processing Magazine, 25(2), pp. 14 - 20, March 2008)
6. David Donoho and Yaakov Tsaig, [Extensions of compressed sensing](#). (Signal Processing, 86(3), pp. 533-548, March 2006)

Compressive Sensing Theory

7. Emmanuel Candès, Justin Romberg, and Terence Tao, [Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information](#). (IEEE Trans. on Information Theory, 52(2) pp. 489 - 509, February 2006)
8. Emmanuel Candès and Terence Tao, "Near optimal signal recovery from random projections: Universal encoding strategies" (IEEE Trans. on Info. Theory, 52(12), pp. 5406 - 5425, December 2006)
9. David Donoho, [Compressed sensing](#). (IEEE Trans. on Info. Theory, 52(4), pp. 1289 - 1306, April 2006)
10. D. Donoho and X. Huo, "Uncertainty Principles and Ideal Atomic Decomposition," IEEE Trans. on Info. Theory, vol.47, no.7, Nov. 2001.
11. M. Elad and A. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," IEEE Trans. Info. Theory, vol. 48, no. 9, Sept. 2002.
12. Scott S. Chen, D. Donoho, and M. Saunders, "Atomic Decomposition by Basis Pursuit," SIAM J. Sci. Comput. 20, pp. 33-61, vol.20, no.1, 1999.

Recovery Algorithms

13. David Donoho and Yaakov Tsaig, [Fast solution of L1-norm minimization problems when the solution may be sparse](#). (Stanford University Department of Statistics Technical Report 2006-18, 2006)
14. Jacob Mattingley and Stephen Boyd, “Real-time convex optimization in Signal Processing,” *IEEE Signal Processing Magazine*, pp.50 – 61, May, 2010.
 - (Source @ http://www.stanford.edu/~boyd/papers/rt_cvx_sig_proc.html).
 - Disciplined CVX Programming
 - The Robust Kalman Filtering example
15. Michael Zibulevsky and Michael Elad, “L1-L2 Optimization in Signal and Image Processing,” *IEEE Signal Processing Magazine*, pp. 76-88, May, 2010.
16. D. Donoho, A. Maleki, and A. Montanari, “Message passing algorithms for compressed sensing,” *PNAS*, Nov. 10, 2009.

Connections to Shannon Theory/Coding Theory

17. The Shannon’s 1948 paper
18. The Rate Distortion Theory (Information Theory, Cover and Thomas)
19. Emmanuel Candès and Terence Tao, Decoding by linear programming. (*IEEE Trans. on Information Theory*, 51(12), pp. 4203 - 4215, December 2005)
20. Emmanuel Candès and Terence Tao, Error correction via linear programming. (Preprint, 2005)
21. Goyal, V.K., Fletcher, A.K., and Rangan, S.; , "Compressive Sampling and Lossy Compression," *Signal Processing Magazine, IEEE* , vol.25, no.2, pp.48-56, March 2008.
doi: 10.1109/MSP.2007.915001
URL: <http://www.ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4472243&isnumber=4472102>.
22. Pier Luigi Dragotti, Martin Vetterli, and Thierry Blu, [Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix](#). (*IEEE Trans. on Signal Processing*, 55(7), pp. 1741-1757, May 2007)
23. Gongguo Tang, Arye Nehorai, [Performance analysis for sparse support recovery](#). (Preprint, Nov 2009)
24. D. Baron, M.F. Duarte, and M.B. Wakin, “Distributed Compressive Sensing,” Dror Baron, Marco F. Duarte, Michael B. Wakin, Shriram Sarvotham, and Richard G.

- Baraniuk, [Distributed compressive sensing](#). (Preprint, 2005) [See also related [technical report](#) and conference publications: [Allerton 2005](#), [Asilomar 2005](#), [NIPS 2005](#), [IPSN 2006](#)]
25. Robert Calderbank and Sina Jafarpour, [Reed Muller Sensing Matrices and the Lasso](#) (Preprint, April 2010)
 26. Maxim Raginsky, Sina Jafarpour, Zachary Harmany, Roummel Marcia, Rebecca Willett, and Robert Calderbank, [Performance bounds for expander-based compressed sensing in Poisson noise](#). (Submitted to IEEE Transactions on Signal Processing, 2010)
 27. S. Jafarpour, X. Weiyu, B. Hassibi, and R. Calderbank, “Efficient and robust compressed sensing using optimized expander graphs,” *IEEE Information Theory*, vol. 55, no. 9, pp. 4299-4308, 2009.
 28. S. Sarvotham, D. Baron, and R. Baraniuk, “Measurements vs. Bits: Compressed Sensing meets Information Theory,” 44th Annual Allerton Conference, Sept. 27-29, 2006.
 29. M. Vetterli, P. Marziliano, T. Blu, “Sampling Signals with Finite Rate of Innovation,” *IEEE Trans. on Signal Processing*,” vol. 50, no.6, June, 2002.

5. The RICE University Repository

Rice University, U.S.A., maintains a nice list of resources for Compressed Sensing papers and software packages.

- [l1-Magic](#)
- [SparseLab](#)
- [GPSR](#)
- [L1 LS: Simple Matlab Solver for L1-Regularized Least Squares Problems](#)
- [sparsify](#)
- [MPTK: Matching Pursuit Toolkit](#) [See also related conference publication: [ICASSP 2006](#)]
- [Bayesian Compressive Sensing](#)
- [SPGL1: A solver for large scale sparse reconstruction](#)
- [sparseMRI](#)
- [FPC](#)
- [Chaining Pursuit](#)
- [Regularized OMP](#)
- [SPARCO: A toolbox for testing sparse reconstruction algorithms](#) [See also related [technical report](#)]
- [TwIST](#)
- [Compressed Sensing Codes](#)
- [Fast CS using SRM](#)
- [FPC_AS](#)
- [Fast Bayesian Matching Pursuit \(FBMP\)](#)
- [SL0](#)
- [Sparse recovery using sparse matrices](#)
- [PPPA](#)
- [Compressive sensing via belief propagation](#)
- [SpaRSA](#)
- [KF-CS: Kalman Filter based CS \(and other sequential CS algorithms\)](#)
- [Fast Bayesian CS with Laplace Priors](#)
- [YALL1](#)
- [TVAL3](#)
- [RecPF](#)

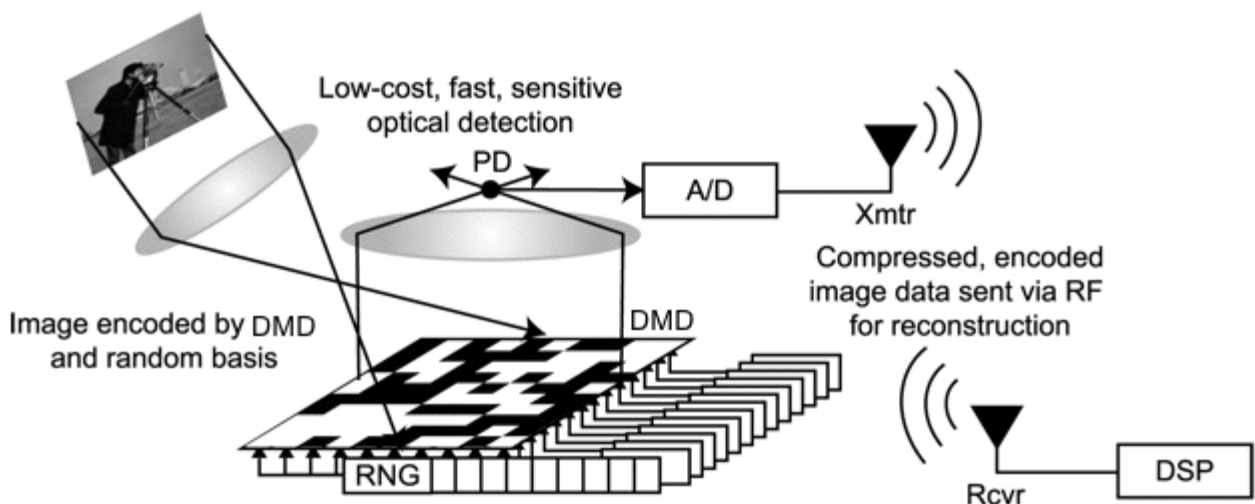
- [Basis Pursuit DeQuantization \(BPDQ\)](#)
- [k-t FOCUSS](#)
- [Sub-Nyquist sampling: The Modulated Wideband Converter](#)
- [Threshold-ISD](#)
- [A Sparse Learning Package](#)
- [Model-based Compressive Sensing Toolbox](#)
- [Sparse Modeling Software](#)
- [Spectral Compressive Sensing Toolbox](#)
- [CS-CHEST: A MATLAB Toolbox for Compressive Channel Estimation](#)
- [DictLearn: A MATLAB Implementation for Dictionary Learning](#)

6. Applications

There are many interesting application areas. Let us review several of them here.

A. Single Pixel Cameras

RICE University (Prof. Baraniuk's group) has applied the Compressed Sensing idea to a single pixel camera, and has shown that the Compressed Sensing idea is not only theoretical but also practical and feasible in real-world system. The following picture depicts the idea.



B. Terahertz pulsed spectroscopic imaging

- Terahertz waves ($0.3 - 10 \text{ THz}$, $10\text{-}330 \text{ cm}^{-1}$) penetrate common barrier materials such as clothes and plastics.

- Terahertz waves can be used in a non-destructive manner to reveal what is concealed such as weapons and explosives behind garments and plastic packages for security applications, or tumors and deceased cells inside human bodies for medical applications.
- Most Terahertz imaging systems use raster scanning with a focused Terahertz beam.
- It seems difficult to build a compact and sensitive multi-element Terahertz detector.
- Raster scanning a whole image scene in a pixel-by-pixel manner takes a large amount of time, say minutes or hours to acquire the total number of pixels for a certain resolution.
- Compressed sensing may provide a rescue for the Terahertz imaging system.
- Each sample in Compressed Sensing paradigm can provide a holistic view of the entire object. The resolution can be controlled by varying the number of holistic samples taken.
- There are a number of papers recently on this subject. Just to name a few, here they are [26][27].

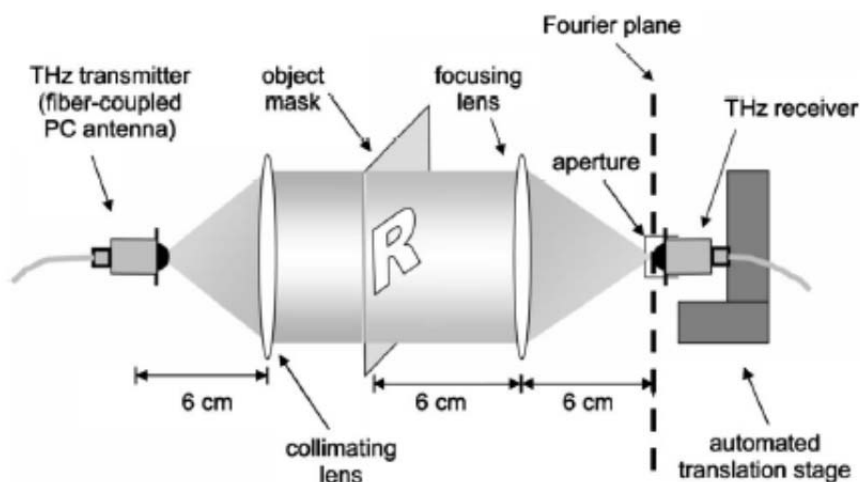


Fig. 1. THz Fourier imaging setup. An approximately collimated beam from the THz transmitter illuminates an object mask, placed one focal length away from the focusing lens. The THz receiver raster scans and samples the Fourier transform of the object on the focal plane.

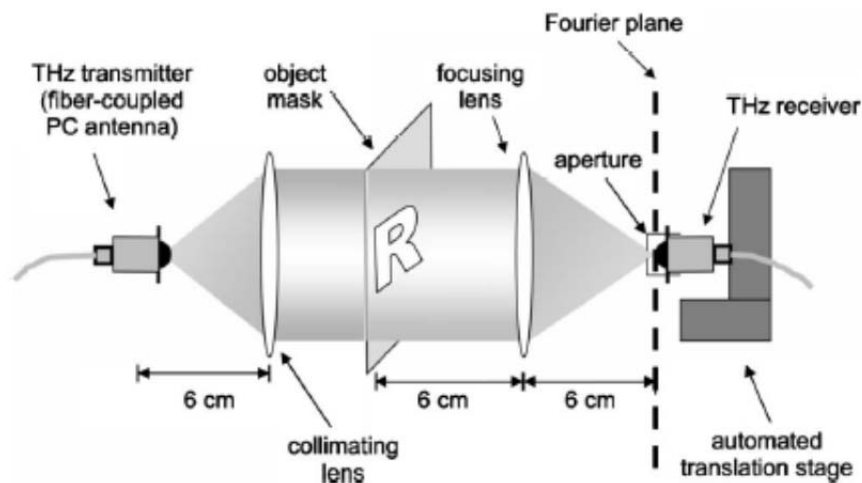


Fig. 1. THz Fourier imaging setup. An approximately collimated beam from the THz transmitter illuminates an object mask, placed one focal length away from the focusing lens. The THz receiver raster scans and samples the Fourier transform of the object on the focal plane.

© 2008 Optical Society of America

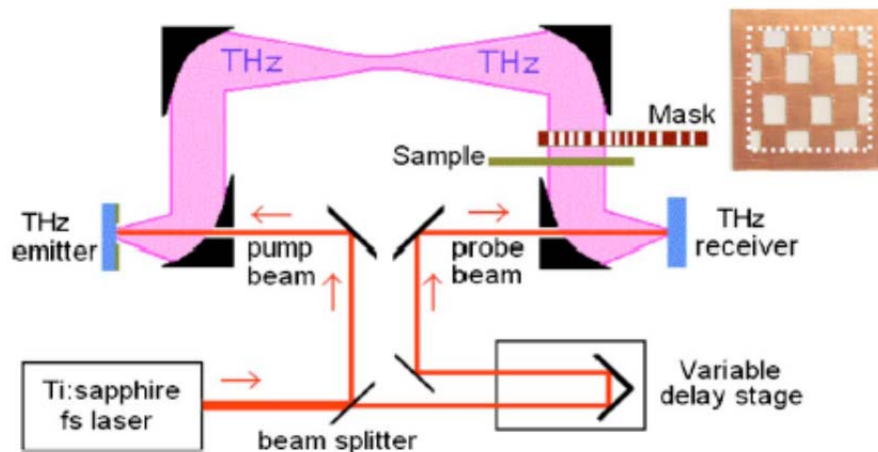


FIG. 1. (Color online) Experimental arrangement for terahertz pulsed imaging using compressive sampling. The inset shows one of 40 masks with the dotted line indicating the $40 \times 40 \text{ mm}^2$ imaging area. The copper pixels are opaque to terahertz radiation while the white pixels are transparent to terahertz radiation.



FIG. 3. Simulated results demonstrating the universality of the proposed mask. (a) Original image and reconstructed images using (b) 40 optimized masks, (c) 40 random masks, and (d) 120 random masks.

C. Other areas of applications

- Brain Computer Interface System with EEG Signal Classifications
- Detection of Images for Security Applications (See Yi Ma's paper)
- Ultrasound imaging system (One possible area)
- Super-Resolution Systems (See this at Chapter II.9)

D. Summary of Applications

One of the challenges is to bring down the cost of these nice technological gadgets. This is possible when low power and low cost signal acquisition and restoration technology are available!!! Compressive Sensing maybe is the way to meet this challenge!

7. References

- [1] David L. Donoho, "Compressed Sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.
- [2] David L. Donoho and Jared Tanner, "Precise Undersampling Theorems," *Proceedings of the IEEE*, vol. 98, pp. 913-924, May, 2010.
- [3] Richard Baraniuk, "Lecture Notes: Compressive Sensing," *IEEE Signal Processing Magazine*, p. 118-121, July, 2007.

Chapter II. COMPRESSED SENSING

1. Compressed Sensing, Compressive Sensing, Compressive Sampling

Compressed Sensing, Compressive Sensing, Compressive Sampling, they all mean the same in this note.

2. Pioneers of Compressed Sensing

“Stars” in the field of the Compressed Sensing include

- David Donoho (Stanford University, Statistics)
- Emmanuel Candes (Stanford University, Statistics)
- Richard Baraniuk (RICE University, ECE)

The claimed statement: Sub-Shannon Nyquist Rate Sampling is good enough for representing (sparse) signals.

Here’s what David L. Donoho has said:

- in his paper Compressed Sensing [4], “everyone now knows that most of the data we acquire “can be thrown away” with almost no perceptual loss—witness the broad success of lossy compression formats for sounds, images, and specialized technical data. The phenomenon of ubiquitous compressibility raises very natural questions: why go to so much effort to acquire **all** the data when **most** of what we get will be thrown away? Can we not just **directly measure** the part that will not end up being thrown away?”
- in another one of his paper [5], “The sampling theorem of Shannon-Nyquist-Kotelnikov-Whittaker has been of tremendous importance in engineering theory and practice. Straightforward and precise, it sets forth the number of measurements required to reconstruct any bandlimited signal. However, the sampling theorem is wrong! Not literally wrong, but psychologically wrong. More precisely, it engender[s] the psychological expectation that we need very large numbers of samples in situations where we need very few. We now give three simple examples which the reader can easily check, either on their own or

by visiting the website [Donoho's Sparse Lab web-site] that duplicates these examples.”

3. Sampling Theorem and Dimensionality Reduction by Shannon

Having seen the quotes from Compressed Sensing papers, it would be interesting to retrospect what Shannon has said in the past on the subject of the sampling theorem.

Here is what Shannon established in the late 1940s in one of his papers, [1], on the issue of sampling theorem. One thing we found interesting is that he also mentioned on the subject of dimensionality reduction when the sampling theorem is used in the context of representing messages.

The Theorem 1 of the paper [1] is stated below:

Theorem 1. (Shannon's sampling theorem [1]) *If a function $f(t)$ contains no frequencies higher than W cps [cycles per second], it is completely determined by giving its ordinates at a series of points spaced $\frac{1}{2W}$ seconds apart.* (See also Review of Sampling Theorem in Problem 6 on page 75)

In communications, it is often of interest to represent a function limited both in time and frequency, say a signal bandlimited to W cps starting at the zero frequency and time limited to the interval of T seconds. This is not possible in the strict sense due to the time-frequency equivalence of the Heisenberg uncertainty principle. But it becomes possible by making an adjustment that the signal has bandwidth W cps and very small values outside the interval T . Taking samples of such a signal at the speed of $2W$ samples per second is sufficient, the theorem states, for reproduction of the signal by interpolation using the sinc $\frac{\sin x}{x}$ kernel. This is not only an engineering approximation, but there exists rigorous forms of similar results by mathematicians, see Whittaker [2].

This theorem opens up the possibility of representing a continuous function $f(t)$ of a certain period T and a bandwidth W with a finite number of equally spaced samples. Namely,

a sequence of $2TW$ number of samples, each sample taken at $\frac{1}{2W}$ second apart, is sufficient for representing any signal with such time and bandwidth limitation.

Shannon in [1] then goes on to the topic of geometrical representation of the signals. Namely, he argues that the $2TW$ evenly spaced samples of a signal can be thought of as co-ordinates of a point in a space of $2TW$ dimensions. A continuous signal $f(t)$ corresponds to a point in this space.

In a similar way, one can associate a geometrical space with the set of possible messages. Suppose a speech signal for example of time duration T and bandlimited by W cps. This signal can also be represented by a set of samples of size $2TW$. Unlike the communications signals which we purposely generate and use as a means to carry digital information over a channel, the message bearing signals such as speech or television signals bear several points need close attention. The former type of signals would be designed to occupy the full dimension so that maximum amount of information is sent over the channel; for the latter case, signals could be grouped together for the purpose of representation and dimensionality reduction can be achieved. Namely, the apparent dimension $2TW$ of these signals can be reduced to $D \leq 2TW$. Shannon argues this dimensionality reduction idea in the following paragraphs:

- “Various different points may represent the same message, insofar as the final destination is concerned. For example, in the case of speech, the ear is insensitive to a certain amount of phase distortion. Messages differing only in the phases of their components sound the same. This may have the effect of reducing the number of essential dimensions in the message space. All the points which are equivalent for the destination can be grouped together and treated as one point. It may then require fewer numbers to specify one of these “equivalence classes” than to specify an arbitrary point. For example, in Fig. 2 we have a two-dimensional space, the set of points in a square. If all points on a circle are regarded as equivalent, it reduces to a one-dimensional space—a point can now be specified by one number, the radius of the circle.”

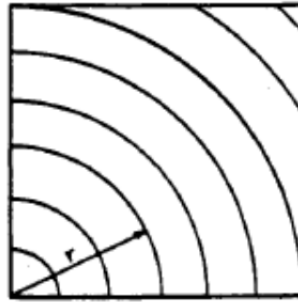


Fig. 2—Reduction of dimensionality through equivalence classes.

Figure 1: Fig. 2 of Shannon [1]

- “In the case of sounds, if the ear were completely insensitive to phase, then the number of dimensions would be reduced by one-half due to this cause alone. The sine and cosine components a_n and b_n for a given frequency would not need to be specified independently, but only $\sqrt{a_n^2 + b_n^2}$; that is, the total amplitude for this frequency. The reduction in frequency discrimination of the ear as frequency increases indicates that a further reduction in dimensionality occurs. The vocoder makes use to a considerable extent of these equivalences among speech sounds, in the first place by eliminating, to a large degree, phase information, and in the second place by lumping groups of frequencies together, particularly at the higher frequencies.”
- “In other types of communication there may not be any equivalence classes of this type. The final destination is sensitive to any change in the message within the full message space of $2TW$ dimensions. This appears to be the case in television transmission. A second point to be noted is that the information source may put certain restrictions on the actual messages. The space of $2TW$ dimensions contains a point for every function of time $f(t)$ limited to the band W and of duration T . The class of messages we wish to transmit may be only a small subset of these functions. For example, speech sounds must be produced by the human vocal system. If we are willing to forego the transmission of any other sounds, the effective dimensionality may be considerably decreased.”

These ideas of dimensionality reduction from the full $2TW$ dimension perhaps go hand in hand with the core idea of the compressed sensing theory, in particularly via the idea behind a sparse representation of a signal in a certain basis.

4. Compressed Sensing in a Nutshell

I would like to start the introduction to the theory of Compressive Sensing, based largely on the tutorial articles [6],[7] published in IEEE Signal Processing Magazine in 2007 and 2008 respectively. The aim here is to illustrate what constitutes the theory of Compressive Sensing. These articles are concise but contain the essential parts of the Compressive Sensing theory; thus, they shall serve as good starting materials for Electrical Engineering and Computer Science and Engineering majors.

Now, let us begin:

1. Need for new look at sampling
 - a Shannon-Nyquist sampling may lead to too many samples probably not all of these samples are necessary to reconstruct a given signal. Compression may become necessity prior to storage or transmission.
 - b In an imaging system, increasing the sampling rate is sometimes difficult.
2. Most signals are compressible signals
 - a Let a real-valued signal represented in a vector form, i.e.,

$$\mathbf{x} = \sum_{i=1}^N s_i \psi_i \quad \text{or} \quad \mathbf{x} = \psi \mathbf{s} \quad (1)$$

- o $N \times 1$ column vectors \mathbf{x} and \mathbf{s}
 - o An $N \times N$ sparsifying basis matrix ψ
 - o The signal \mathbf{x} is called *K-sparse* if it can be represented as a linear combination of only K basis vectors; only K elements of the vector \mathbf{s} are non-zero.
 - o The signal \mathbf{x} is called *compressible* if it contains a few elements with large values and many elements with small values.
3. Compression using the usual transformation based source coding (lossy)
 - o Uniform sample vector \mathbf{x} is obtained.
 - o Transform coefficients, via $\mathbf{s} = \psi^{-1} \mathbf{x}$, are found.
 - o K largest elements are taken; the rest are thrown off.
 - o Encode the K largest elements
 - o Inefficiency can be noted here.

4. The Compressive Sensing approach

- Directly acquire compressed samples without going through the intermediate stages
- Compressive measurements via linear projections

$$\mathbf{y} = \Phi \mathbf{x} = \Phi \psi \mathbf{s} = \Theta \mathbf{s} \quad (2)$$

- Here \mathbf{y} is an $M \times 1$ measurement vector, where $M < N$.
 - Φ , or the Θ , is an $M \times N$ measurement matrix.
 - A good measurement matrix preserves the information in \mathbf{x} .
 - A good recovery algorithm recovers \mathbf{x} .
5. In Compressive Sensing, there are two major tasks. Namely, they are
- Designing a good **measurement matrix**. Matrices with large compression effects and robustness against modeling errors are desired.
 - Designing a **good signal recovery algorithm**. Fast and robust algorithms are desired.

5. Compressed Sensing, explained with a little more care

- A K -sparse signal, $\mathbf{x} = \boldsymbol{\Psi}\mathbf{s}$, where there are K non-zero elements in \mathbf{s} .
 - The dimension is N .
 - $\boldsymbol{\Psi}$ is an orthonormal basis, i.e., $\boldsymbol{\Psi}^H\boldsymbol{\Psi} = \boldsymbol{\Psi}\boldsymbol{\Psi}^H = \mathbf{I}_N$, the identity matrix of size N , where the superscript $(\cdot)^H$ denotes the Hermitian transpose.
- An M by 1 measurement vector \mathbf{y} ,

$$\mathbf{y} = \boldsymbol{\Phi}\mathbf{x} = \boldsymbol{\Psi}\mathbf{s} = (\boldsymbol{\Phi}\boldsymbol{\Psi})\mathbf{s} =: \boldsymbol{\Theta}\mathbf{s} \quad (3)$$

The sensing matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$ are of size $M \times N$, where $M < N$.

- For a K -sparse signal \mathbf{x} , a minimization based on the L1 norm (L_1 norm) gives the unique solution \mathbf{x} under the condition that the sensing matrix $\boldsymbol{\Phi}$ is good.
- Surprising results:
 - M is closer to K than N as a sufficient condition for good signal recovery. Thus, there is a compression effect. It turns out that this is not so surprising from the perspective of channel coding, e.g., syndrome decoding. The dimension of syndrome is smaller than ambient dimension of the sparse error vector.
 - The L1 minimization solution gives a solution equivalent to the L0 solution which is the combinatorial solution, under a certain condition.
- Like I said in the previous lecture, the compressive sensing comes down to the following two problems.
 - The design of good sensing matrix
 - The design of good recovery algorithm
- Let us take a look at them one by one.
- Note that (3) appears to be an ill-conditioned system. There are more unknowns than the number of equations, $N > M$.
- But if \mathbf{x} is K -sparse and the locations of the K non-zero elements are known, then the problem can be solved provided $M \geq K$.

- Namely, we can form a simplified equation by deleting all those columns and elements corresponding to the zero-elements:

$$\mathbf{y} = \Theta_{\mathcal{K}} \mathbf{s}_{\mathcal{K}} \quad (4)$$

where \mathcal{K} is the support set which is the collection of indices corresponding to the non-zero elements of \mathbf{s} .

◆ Not always! Can you come up with a counter example?

- Equation (4) has the unique solution $\mathbf{s}_{\mathcal{K}}$ if the columns of $\Theta_{\mathcal{K}}$ are linearly independent. It can be found by

$$\mathbf{s}_{\mathcal{K}} = \left(\Theta_{\mathcal{K}}^T \Theta_{\mathcal{K}} \right)^{-1} \Theta_{\mathcal{K}}^T \mathbf{y} \quad (5)$$

Note that the inverse matrix exists since the columns are independent.

- Thus, once the support set is found, the problem is easy to solve provided the columns are independent.
- The support set \mathcal{K} can be any size K subset of the full index set $\{1, 2, 3, \dots, N\}$.
- The necessary and sufficient condition for (4) to be well conditioned is that for any K -sparse vector \mathbf{v} sharing the same K nonzero entries as \mathbf{s} , the sensing matrix should satisfy the following condition, for some $0 < \delta < 1$:

$$1 - \delta \leq \frac{\|\Theta \mathbf{v}\|_2}{\|\mathbf{v}\|_2} \leq 1 + \delta \quad (6)$$

- Θ should be length preserving for any K -sparse \mathbf{v} .
- It should be noted that if the condition (6) holds, then any K columns of Θ are linearly independent. Thus, the sufficient part has been proved.
- Let us wait until we study Candes and Tao's paper for the necessary part.
- At this point, our aim is to get familiar a little bit with this inequality (6). Later, this inequality will be used repeatedly under the name Restricted Isometry Property (RIP), which will be discussed next.

A. Restricted Isometry Property

A sufficient condition for a *stable* solution for both K -sparse and *compressible* signals is that the sensing matrix satisfies (6) for any arbitrary $3K$ -sparse vector \mathbf{v} .

- This statement is not obvious at the moment.
- We know that (6) for $1K$ -sparse vectors is a sufficient condition for (4) which is a simplified version of (3) under the assumption that the support set \mathcal{K} is known.
- The RIP for $3K$ -sparse vector is obviously a stricter condition than that with $1K$ -sparse vectors.
- It is natural to find a stricter condition, the RIP, for equation (3) in which the support set \mathcal{K} is also unknown, in addition to the values of $\mathbf{s}_{\mathcal{K}}$.

B. Incoherence Condition

The story related to incoherence condition is to say that the rows of Φ should be incoherent to the columns of Ψ . Why?

- What would happen if the rows of Φ are *coherent* to the columns of Ψ ?
- In the extreme case, we may select the M rows of Φ to be the first M columns of Ψ .
- Then, we have

$$\Phi\Psi = \Psi_{(1:M,:)}^T \Psi = \begin{pmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$$

- Note that it is easy to see that this matrix can never satisfy the RI condition.
- Another item in the story is that if i.i.d. Gaussian is used to construct the sensing matrix, it will be incoherent to any basis.

C. Checking RIP is NP-hard.

- Deterministic approach
 - Checking the RI condition is an NP-hard problem.
 - Given a sensing matrix Θ , we should check all $\binom{N}{3K}$ possible combinations of $3K$ non-zero entries in the vector \mathbf{v} of length N .
 - This quickly becomes intractable for large N .
- Probabilistic approach

- Design Φ randomly and show that the RIP and the incoherence condition can be achieved with high probability.
- For example, let each element of Φ be i.i.d. Gaussian with zero mean and variance $1/M$.
- Gaussian sensing matrix Φ has two useful properties
 - With $N > M \geq cK \log(N/K)$ where c is a constant, the RIP is met with high probability; thus, the K -sparse signal can be recovered.
 - The matrix Φ is universal in the sense that $\Theta = \Phi\Psi$ will be i.i.d. Gaussian and thus have the RIP condition met with high probability for any choice of the orthonormal basis set Ψ .
- ◆ Let us check: First we note the columns of the sensing matrix Θ are mutually independent with each other, i.e.,

$$\begin{aligned}\mathbb{E}\Theta^H\Theta &= \mathbb{E}\Psi^H\Phi^H\Phi\Psi \\ &= \Psi^H\mathbb{E}(\Phi^H\Phi)\Psi \\ &= \Psi^H\mathbf{I}_N\Psi \\ &= \mathbf{I}_M\end{aligned}$$

- ◆ Second, we note that the rows of the sensing matrix Θ are mutually independent with each other as well, i.e.,

$$\begin{aligned}\mathbb{E}\Theta\Theta^H &= \mathbb{E}\Phi\Psi\Psi^H\Phi^H \\ &= \mathbb{E}(\Phi\Phi^H) \\ &= \frac{N}{M}\mathbf{I}_M\end{aligned}$$

- ◆ Third, each element in Θ is i.i.d. Gaussian.

D. L0, L1, L2 norms and the null space

- A signal reconstruction algorithm
 - Takes the input which is the measurement vector \mathbf{y}
 - Outputs the K -sparse vector \mathbf{x}
- The null space $\mathcal{N}(\Theta)$ of the sensing matrix Θ : The null space of Θ is defined as the collection of all vectors \mathbf{v} such that

$$\Theta \mathbf{v} = \mathbf{0} \quad (7)$$

- Namely, $\mathcal{N}(\Theta) = \{\mathbf{v} \mid \Theta \mathbf{v} = \mathbf{0} \text{ for any non zero } \mathbf{v} \in \mathbb{R}^N\}$. Since the dimension of Θ is $M \times N$, the dimension of the null space is at least $N - M$.
- Thus, there are infinitely many solution \mathbf{s}' to (3), $\mathbf{s}' := \mathbf{s} + \mathbf{v}$ where $\mathbf{v} \in \mathcal{N}(\Theta)$. That is, each \mathbf{s}' is the solution to $\mathbf{y} = \Theta \mathbf{s}' = \Theta(\mathbf{s} + \mathbf{v}) = \Theta \mathbf{s}$.
- Ex) Find the null space of the following matrix:

$$\Theta = \begin{pmatrix} 1 & 0 & \frac{1}{2} \\ 0 & 1 & \frac{1}{2} \end{pmatrix} \quad (8)$$

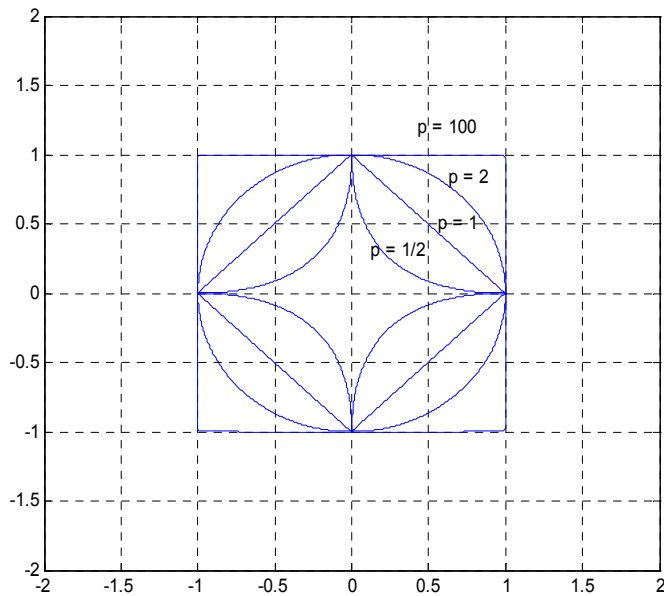
- We need a criterion to choose a solution uniquely.
- We will consider minimum L2 norm, minimum L1 norm, and minimum L0 norm criterion.

Example) Let $\mathbf{x} = [1 \ -1 \ 0]$ and $\mathbf{y} = [0.5 \ -0.5 \ 0.5]$. Whose is bigger in the sense of L0 norm? Whose L1 norm is bigger? Whose L2 norm is bigger?

- i. The L_p norm of \mathbf{x} is defined for $p > 0$; as

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^N |x_i|^p \right)^{\frac{1}{p}} \quad (9)$$

- ii. The unit circles with respect to different L_p norms, with $p = 1/2, 1, 2$, and $100(\infty)$.



- iii. The L_∞ norm is $\|\mathbf{x}\|_\infty = \max\{|x_1|, |x_2|, \dots, |x_N|\}$.
- iv. The L_0 norm is not well defined as a norm. Donoho uses it as a “norm” which counts the number of non-zero elements in a vector.

- The minimum L_2 norm solution:

$$\begin{aligned}\hat{\mathbf{s}} &= \arg \min \|\mathbf{s}'\|_2 \quad \text{s.t. } \mathbf{y} = \Theta \mathbf{s}' \\ &= \Theta^T (\Theta \Theta^T)^{-1} \mathbf{y}\end{aligned}\tag{10}$$

However, this conventional solution will give us a non-sparse solution and will not be appropriate. We will do a homework problem for this.

- The minimum L_0 norm solution:

$$\hat{\mathbf{s}} = \arg \min \|\mathbf{s}'\|_0 \quad \text{s.t. } \mathbf{y} = \Theta \mathbf{s}'\tag{11}$$

The L_0 norm of a vector is the number of non-zero elements in the vector by definition. This involves combinatorial search, finding all $\binom{N}{K}$ possible support sets. This is an NP-complete problem.

- The minimum L_1 norm solution: The biggest surprise in compressive sensing comes from this. Namely, the L_1 norm solution coincides with the L_0 solution provided the

RIP condition is met.

$$\hat{\mathbf{s}} = \arg \min \|\mathbf{s}'\|_1 \quad \text{s.t. } \mathbf{y} = \mathbf{\Theta} \mathbf{s}' \quad (12)$$

We will have to spend some time to prove this statement in this course.

i. Example of L1 norm.

■ Example) Consider the following underdetermined problem:

$$\mathbf{y} = \begin{bmatrix} 1 & -2 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad (13)$$

Let $\mathbf{y} = -2$.

- i. Find the L0 solution
- ii. Find the L2 solution.
- iii. Find the L1 solution.

6. Summary

The bottom line is that

1. The L1 norm minimization solution is the L0 norm solution under the condition that the RIP is met.
2. A randomly generated i.i.d. Gaussian measurement matrix Φ with dimension $N > M \geq cK \log(N/K)$ satisfies the RIP condition with high probability.

Having said this, it feels like that we have already solved both of the problems related to Compressive Sensing. Namely, we know how to design a good sensing matrix as well as a good recovery algorithm: use an i.i.d. Gaussian sensing matrix, and apply the L1 norm minimization to obtain a signal recovery algorithm. This is true in some sense. But it should rather be the beginning of a new field, I hope since we need new ideas and new applications. Current issues of interest may include the following:

1. Design of sensing matrix with *deterministic* performance guarantee
2. *Faster* signal recovery algorithms
3. Application of the Compressive Sensing theory to solve practical problems: channel coding problems, super-resolution problems, sparse representation problems, image compression problems, etc.
4. Finite field results: If we wanted to use Compressed Sensing for compression purpose, it would be perhaps better off if we have used the parity check matrices of the channel codes. The syndrome vectors are packed in its vector space, $\text{GF}(q)^M$, while the error vectors are widely spread out in its vector space, $\text{GF}(q)^N$. More discussion on this shall be needed. See those sections Chapter VI.2 for the Coding Theoretic Approach, and 0 for the Bayesian Recovery methods.

7. The Spark and The Singleton Bound

Definition. The **spark** of a matrix \mathbf{A} is the smallest number n such that there exists a set of n columns in \mathbf{A} which are linearly dependent, i.e.,

$$\text{spark}(\mathbf{A}) := \min_{\mathbf{x} \neq \mathbf{0}} \|\mathbf{x}\|_0 \quad \text{s.t. } \mathbf{A}\mathbf{x} = \mathbf{0}. \quad (14)$$

Prove/disprove questions.

- The rank of an $[M \times N]$ matrix \mathbf{A} , $M < N$, can be larger than M .
- The rank of matrix \mathbf{A} is the minimum of the column and the row dimension.
- The rank of an $[M \times N]$ matrix \mathbf{A} , $M < N$, is the smallest number D such that *all* sets of $D + 1$ columns in \mathbf{A} are linearly dependent, and D can be larger than or equal to M .

- **(The Singleton Bound)** The highest spark of an $[M \times N]$ matrix \mathbf{A} is less than or equal to $M + 1$.

8. Matrix Design with Givens Rotations

- Exercise Problem 1: Let us design a sensing matrix \mathbf{A} starting with the following initial matrix:

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & \\ & & & \end{bmatrix}$$

- ◆ A matrix with spark = 3, let us aim to design.
- ◆ Any idea? Any systematic method?
- ◆ Let us use the Givens rotation matrix, $G = \begin{pmatrix} \cos(a) & \sin(a) \\ -\sin(a) & \cos(a) \end{pmatrix}$. This matrix will turn its input vector by angle a in the clock wise direction. For example, let us use $a = \pi/2$. Then, we obtain the sensing matrix \mathbf{A} as the concatenation of the two, the two by two identity matrix \mathbf{I} and G , $\mathbf{A} = [\mathbf{I}; G]$.

- Exercise Problem 2: Design a matrix \mathbf{A} starting with the following initial matrix:

$$\begin{bmatrix} 1 & & & \\ & 1 & & \\ & & & \\ & & & 1 \end{bmatrix}$$

- ◆ A matrix with spark = 3
- ◆ A matrix with spark = 4
- ◆ Can we use the Givens rotation matrix again?
- ◆ Using $G_1 = \begin{pmatrix} c(a) & s(a) & 0 \\ -s(a) & c(a) & 0 \\ 0 & 0 & 1 \end{pmatrix}$ with $a = \pi/2$, we have the following rotation

$$\text{result } \underbrace{\begin{pmatrix} 1 & & \\ & 1 & \\ & & 1 \end{pmatrix}}_{=: \mathbf{A}_1} \underbrace{\begin{pmatrix} c(a) & s(a) & 0 \\ -s(a) & c(a) & 0 \\ 0 & 0 & 1 \end{pmatrix}}_{G_{12}} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \\ & & 1 \end{pmatrix} =: \mathbf{A}_1 G_{12}$$

- ◆ Using $G_{23} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & c(a) & s(a) \\ 0 & -s(a) & c(a) \end{pmatrix}$ with $a = \pi/2$, we have the following rotation

$$\text{result } \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & \\ & & 1 \end{pmatrix} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & c(a) & s(a) \\ 0 & -s(a) & c(a) \end{pmatrix}}_{G_{23}} = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{\sqrt{2}} & \frac{1}{2} & \frac{1}{2} \\ & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \mathbf{A}_1 G_{12} G_{23}.$$

- ◆ We note that the spark, when we form the 3 x 6 matrix with this result, is only 3.
 ◆ Why?
 ◆ So, let us do it one more time. This time, let us do a rotation between the axis 1

and the axis 3. That is, using $G_{13} = \begin{pmatrix} c(a) & 0 & s(a) \\ 0 & 1 & 0 \\ -s(a) & 0 & c(a) \end{pmatrix}$ with $a = \pi/4$, we have

the following rotation result

$$\begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{2} & \frac{1}{2} \\ -\frac{1}{\sqrt{2}} & \frac{1}{2} & \frac{1}{2} \\ & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \underbrace{\begin{pmatrix} c(a) & 0 & s(a) \\ 0 & 1 & 0 \\ -s(a) & 0 & c(a) \end{pmatrix}}_{G_{13}} = \begin{pmatrix} 0.1464 & 0.5 & 0.8536 \\ -0.8536 & 0.5 & -0.1464 \\ -0.5 & -\frac{1}{\sqrt{2}} & 0.5 \end{pmatrix} = \mathbf{A}_1 G_{12} G_{23} G_{13}.$$

- ◆ Now, let us form the sensing matrix \mathbf{A} using the results obtain so far

$$\text{◆ } \mathbf{A} = \begin{bmatrix} 1 & 0.1464 & 0.5 & 0.8536 \\ & 1 & -0.8536 & 0.5 & -0.1464 \\ & & 1 & -0.5 & -\frac{1}{\sqrt{2}} & 0.5 \end{bmatrix}. \text{ We note that the spark of this matrix}$$

is 4.

- ◆ We have not aimed at optimizing this process yet. But from this simple procedure, it can be seen that we can easily obtain a matrix with spark 4, and thus the Singleton bound has been achieved. That is, the angle we have chosen at each rotation is simply $\frac{\pi}{4}$. Was this an optimal choice? Probably not. This should be in the list of possible future research topics. For higher dimensional cases, the

best way to find the rotations is an open problem, and a good research direction.

- ◆ Another possible direction to look at relevant works is in the space-time block code design literature.
 - See under “Unitary Space-Time Constellation” designs. There the aim is to design a codebook which is collection of rectangular matrices; thus, their problem is a little different. The distances between any pair of codewords have been maximized.
 - See the paper published in 2010 by Ashikhmin and Calderbank [8].

9. Super Resolution Applications (Nano array filters and Nano lenses)

An interesting application area of compressive sensing could be *filter array based spectrometers*. See the paper “on the estimation of target spectrum for filter array based spectrometers,” C.C. Chang and H.N. Lee, OPTICS EXPRESS, vol. 16, no.2, 2008.

In this application, the inverse problem in the compressed sensing theory can be used in a little bit different flavor. Super-resolution perspective, rather than compression, can be emphasized solving the under-determined system of equations. In compressed sensing, we aim to recover the uncompressed samples, an $N \times 1$ vector, \mathbf{x} from the compressed samples, an $M \times 1$ vector $M < N$. The emphasis was put on the effect of compression, since $M < N$. This can be viewed differently as we put the emphasis on the fact that N is bigger than M when the number of samples M is given and fixed in a particular system. The view point is that we started off with a fixed number of M measurement samples, we aim to improve its resolution by a factor of N/M . In the paper referred above, we have used a non-negative least squares algorithm (NNLS). Back then, we did not notice the compressive sensing algorithms nor the sparse representation problem, and thus we did not make use of the L1 minimization algorithms. Somehow, however, we were able to pull out and use the best algorithm, the NNLS. Since the intensity of light is non-negative, this was the obvious choice.

The NNLS algorithm has the intriguing power of achieving parsimonious representation. When one compares the algorithm with the L1 minimization routines, one finds that the NNLS algorithm is superior to many of these L1 minimization routines. It is an iterative algorithm. At each iteration, an active set is managed. An active element is selected by correlating the residual vector with the columns of the “sensing” matrix. Lawson and Hanson [30] have shown that the algorithm always converges.

The problem of super resolution for nano array filter is summarized here:

- The objective is to obtain a miniature spectrometer which is small and portable.
- A nano array filter is cheap and incapable as an individual filter. Each filter is not sharp, but blurs the input image. Namely, it does not only pass a specific wavelength but also others at different wavelengths. In addition, a CCD camera array is

expensive.

- The aim is to obtain the maximum resolution with a small number of photoelectric sensors, a resolution beyond the *Nyquist* Rate.
- The output of a particular filter is a convolution of the light inputs at different wavelengths.
- For the success of a miniature spectrometer, it is desired to have a smaller number of photoelectric sensor arrays.
- The question is to ask “is it possible to have a spatial resolution finer than the spatial resolution obtained by the fixed number of sensors in the array?”

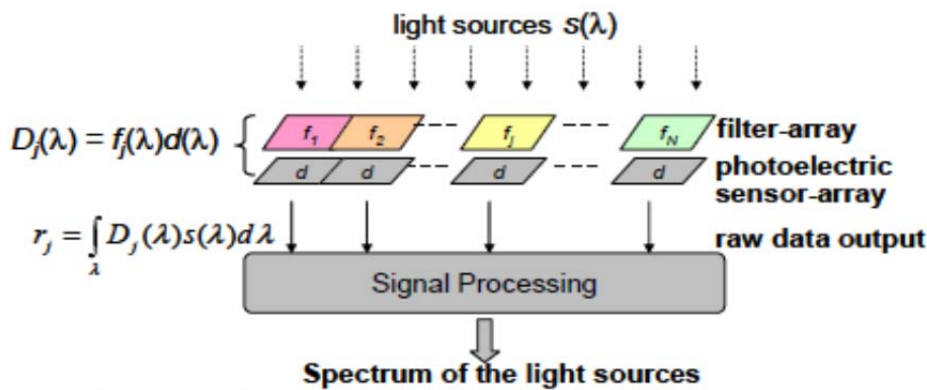


Fig. 1. System structure of the filter-array based spectrometers

Namely, we have

$$\mathbf{r} = \mathbf{H}\mathbf{s} + \mathbf{n}$$

where \mathbf{r} is the $M \times 1$ observation vector, \mathbf{H} is the $M \times N$ diffusion (convolution) matrix, \mathbf{s} is the sparse vector, and \mathbf{n} is the AWGN vector. Note here that M stands for the Nyquist rate spatial sampling; the aim is to see if the spatial resolution of the image can be improved, i.e., $N > M$. That is, we would like to see if beyond the Nyquist rate resolution can be achieved by exploiting the prior knowledge that the image is compressible.

Murky Lenses Make Sharper Images!

A similar application can be found in the Nanolens area. See one article at New Scientist (July 2nd, 2011), “Murky lenses make sharper images,” written by MacGregor Campbell. It introduces a recent breakthrough in super resolution experiment in which a scattering lense, a rat skin or a layer of a peel of egg’s interior shell, can be used to improve resolution of an optical image. It includes the discussion of a technical paper entitled as “Scattering Lens Resolves Sub-100 nm Structures with Visible Light,” by Putten et. al., published in Physical Review Letters, May 13th, 2011 (DOI: 10.1103/PhysRevLett.106.193905). In the second article, they use a layer of Gallium Phosphide topped by a slab of Porous layer for creating a

scattering lense.

Here is the logic:

- When the wavefront of a coherent light source propagates through a scattering medium, the rays are scattered in all directions, each with different phases and intensities.
- The scattered rays are collected with an object lens. These rays are out of phase with each other, obviously.
- These phase differences can be measured and fed back to the source.
- A spatial modulator can then be used at the source to modulate the phases of transmitted light so that the collected rays at the image collector can remain coherent with each other.

The following pictures are taken from the Putten et. al.'s paper:

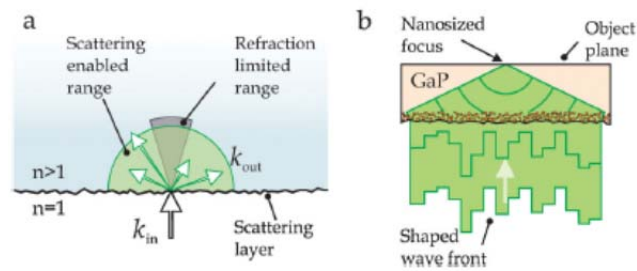


FIG. 1 (color). (a) Principle of light coupling to high transversal k vectors into a high-index material. Without scattering refraction would strongly limit the angular range to which light could be coupled. By exploiting strong scattering at the interface, incident light k_{in} is coupled to all outgoing angles k_{out} in the high-index material. (b) Schematic of a HIRES lens that uses light scattering to achieve a high optical resolution. This HIRES lens consists of a slab of gallium phosphide (GaP) on top of a strongly scattering porous layer. By controlling the incident wave front, a small focus is made in the object plane of the HIRES lens.

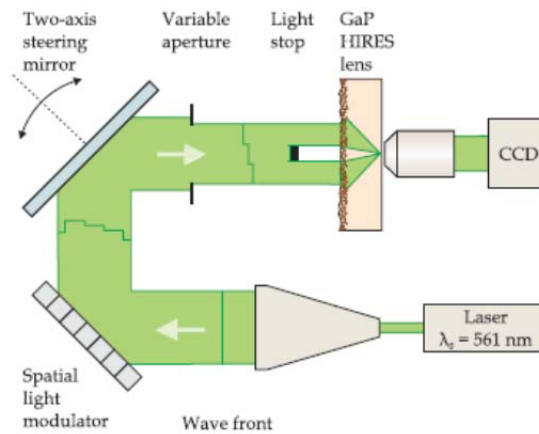


FIG. 2 (color). Overview of the setup. A $\lambda_0 = 561$ nm laser beam is expanded and illuminates a phase-only spatial light modulator. The modulated reflected beam is first imaged onto a two-axis steering mirror and then onto the porous surface of the GaP HIRES lens. A variable aperture controls the extent of the illuminated area and a light stop places the setup in a dark field configuration by blocking the center of the light beam. We image the object plane onto a CCD camera using an oil immersion microscope objective.

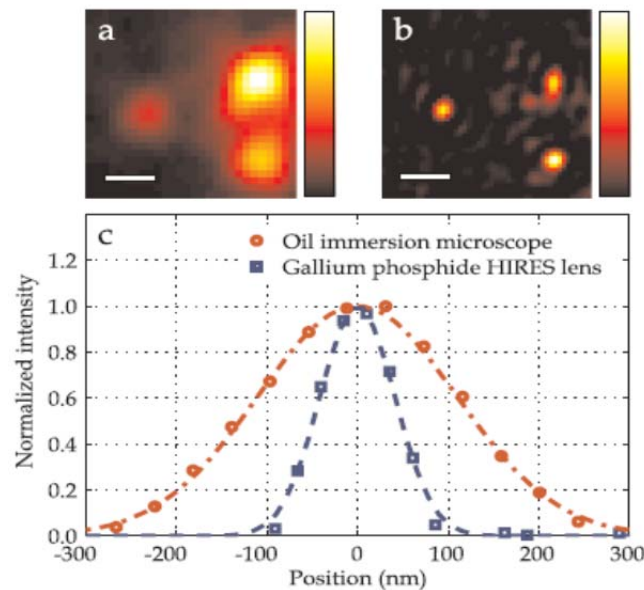


FIG. 3 (color). Experimental imaging demonstration with a GaP HIRES lens. (a) A reference image taken with conventional oil immersion microscope (NA = 1.49). The image shows a blurred collection of gold nanoparticles. The scale bar represents 300 nm. (b) A high-resolution image acquired with our GaP HIRES lens. The image was obtained by scanning a small focus over the objects while monitoring the amount of scattered light and deconvoluted with Eq. (1) [25]. (c) A vertical cross section through the center of the left sphere in (a) and (b) shows the increase in resolution. The dashed lines are Gaussian fits to the data points.

10. HW set #1

1. (L_p norms; L0 and L1 solutions) Let us learn the difference between various L_p norms.
 - a Write the definition of the L0 norm.
 - b Use the MATLAB and draw the boundaries of the L_p balls, where $p = 2, 1,$ and $1/2$ in the 2 dimensional space.
 - c Now find the solutions of the following equation with respect to L0, L1, and L1/2 minimization, $y = [1 \ 2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ for $y = 2$. Write the expression of the right null space of the matrix $[1 \ 2]$, and draw it on a same picture together with the L_p balls.
 - d Depict your L_p solutions in your picture and explain the procedure how you have obtained your solution. See [Fig 3] of [7] for reference.
 - e Find the L0 norm solution. Is it unique?
 - f Find the L1 norm solution. Is it unique? Does it coincide with the L0 norm solution? If not, when does it?

2. (RIP, Spark, Rank) Let \mathbf{A} be an arbitrary $[3 \times 6]$ real valued matrix.
 - a Find the maximum row dimension of the matrix. Give an example matrix \mathbf{A} that achieves the maximum row dimension.
 - b Find the maximum column dimension of the matrix. Give an example matrix \mathbf{A} .
 - c Design a $[3 \times 6]$ sensing matrix every set of three columns of which is linearly independent.
 - d What is the rank of the matrix of c?
 - e What is the Spark of the matrix of c?

3. (RIP + L0 recovery) Let a $[4 \times 8]$ matrix \mathbf{A} have Spark = 5.
 - a Prove/Disprove that the matrix \mathbf{A} satisfies the lower bound part of the RIP condition for any 4-sparse signals.
 - b Prove/Disprove that the solution to $\mathbf{y} = \mathbf{A}\mathbf{x}$ has the unique L0 solution for any 2-sparse \mathbf{x} .

4. (Underdetermined problem) What would happen when $M \geq N$ in (3)?
 - a (Prove/Disprove) The solution always exists and unique.
 - b (Prove/Disprove) The L1 solution (12) and the L2 solution (10) coincide, when

the unique solution exists.

5. (DCT, Noiselet, Wavelet, Fourier Transform) Let us figure out what they are and do small examples with them.
 - a Obtain the general expression of the DCT transformation. Give a small example of DCT (say 4×4 or 8×8).
 - b Do the same with a Noiselet transformation. See Reference [8] of [7].
 - c Do the same with a Wavelet transformation (the Harr wavelet).

11. References for Chapter 2

- [4] David L. Donoho, “Compressed Sensing,” *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.
- [5] David L. Donoho and Jared Tanner, “Precise Undersampling Theorems,” *Proceedings of the IEEE*, vol. 98, pp. 913-924, May, 2010.
- [6] Richard Baraniuk, “Lecture Notes: Compressive Sensing,” *IEEE Signal Processing Magazine*, p. 118-121, July, 2007.
- [7] Justin Romberg, “Imaging via compressive sampling,” *IEEE Signal Processing Magazine*, 25(2), pp. 14 - 20, March 2008.
- [8] Ashikhmin, A. and Calderbank, A.R., “Grassmannian Packings from Operator Reed-Muller Codes,” *IEEE Trans. Information Theory*, vol. 56, Issue: 11, pp. 5689-5714, 2010.

Chapter III. MATHEMATICS OF COMPRESSIVE SENSING

In this chapter, we aim to introduce a selection of mathematical results in compressed sensing theory. Before we do this, we summarize what we learned in previous chapters:

- Natural signals can be represented as sparse signals in a certain basis. We say that a signal is K -sparse if only K non-zero elements are needed to describe a signal.
- Sparse signals can be compressively sampled, meaning that the number M of samples needed for perfect reconstruction is less than the number N of Shannon-Nyquist samples. In our notation, the relation between the number of samples are $K < M < N$.
- The reconstruction of the signal is done by the L1 minimization, rather than the usual L2 minimization.

This line of thoughts is considered pioneering. Obviously, there are many questions we would like to ask. Among them, the first batch could be:

- How to obtain the $[M \times N]$ sensing matrix?
- How small M can we choose given K and N ?
- How sparse the signal has to be at a given M ?
- When will the L1 convex relaxation solution attains the L0 solution?

These questions are related with each other, we will find answers to them as we explore. Being able to answer them would mean that we have understood the core results of the compressed sensing theory. After we have conquered them, we may explore further questions dealing with more practical issues such as:

- The input output model, $y = Fx$, is too simple. In practice, there is always measurement noise. What would happen to the L1 minimization signal recovery then? Would it still be possible to recover the signal reliably?
- What would happen if there is a model mismatch? That is, suppose that the signal is not exactly a K -sparse signal, what kind of results do we expect under such an assumption?

In this chapter, we will explore these issues.

We now begin our discussion with the topic of uncertainty principle and sparse representation.

The first section presents the result of Donoho et al., sufficient conditions for L0 unique and L1 unique solutions, which are given for a special case when there are two orthonormal bases used for sparse representation of a given signal y . In compressive sensing approach, this is a special case where the sensing matrix dimension is given by $M \times 2M$.

We then give the sufficient conditions by Candes and Tao which are for general $M \times N$ sensing matrices. These results are however given by the so-called RIP constants which are tight but difficult to evaluate.

In Section 3, we give our novel results on L0 and L1 uniqueness theorems for the general $M \times N$ sensing matrices. Novelty lies in the fact that these theorems are given in terms of mutual coherence which is easy to calculate.

1. Uncertainty Principle and the L1 Uniqueness Proof

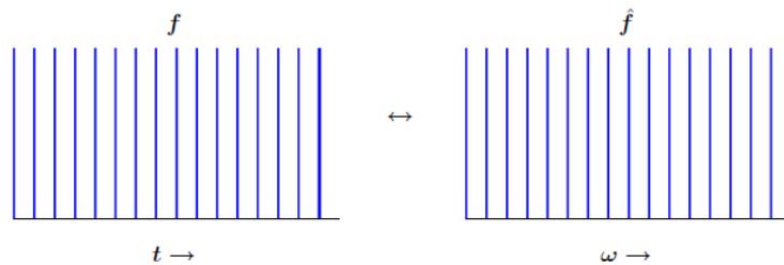
In this section, we aim to introduce the subjects of uncertainty principle and L1 uniqueness first studied by Donoho and Stark in 1989, and later by others including Donoho and Huo, Elad and Bruckstein, Griebonbal and Nielson, etc. This subject is very interesting due to a number of reasons. Just to name a few. First, the uncertainty principle provides the fundamental law of signal resolution for sparse signal representation. The spirit is that a parsimonious representation is of value: a small sparsity representation, being able to represent a signal in a parsimonious way, is a figure of merit. But be aware that there is a limit. The uncertainty principle (UP) gives that limit. Second, it has intriguing relation to the L1 uniqueness proof which is one of the key issues in Compressed Sensing.

There is nice review presentation made by Romberg [15] available in the Internet. A quick reference to this material, we find that the research history of this subject goes as follows:

- i. Discrete Uncertainty Principle for \mathbb{C} : $|T| + |\Omega| \geq 2\sqrt{N}$ (Donoho and Stark 1989)

Dirac Comb

The discrete uncertainty principle is exact.



- \sqrt{N} spikes spaced \sqrt{N} apart
- Invariant under Fourier transform ($f = \hat{f}$)
- $|T| + |\Omega| = 2\sqrt{N}$

Figure 2 : A chart from [15].

- ii. Generalization to pairs of bases
 - A. Donoho and Huo 2001 [13]
 - B. Elad and Bruckstein 2002 [14]
- iii. Sharp Uncertainty Principle by Tao 2004
 - A. The block length N is prime.
 - B. $|T| + |\Omega| \geq N$ (much more relaxed result)

Sharp Uncertainty Principle: N Prime

[Tao (2004)]

- For N prime, evenly spaced signals like the Dirac comb are impossible.
- The uncertainty principle is *significantly* more relaxed:

$$|\text{supp } f| + |\text{supp } \hat{f}| > N$$

- Key: minors of the Fourier matrix have full rank

$$A = R_\Omega F R_T^*$$

R_T, R_Ω are restriction operators.

- Compare to general UP: N vs. $2\sqrt{N}$

Figure 3 : A chart from [15].

In this sub-section, our discussion will be based primarily on Donoho and Huo 2001 [13] and Elad and Brucstein 2002 [14]. In the next sub-sections, we will have a chance to discuss Tao's UP when we consider papers by Candes, Romberg, and Tao [8][9][10].

In quantum mechanics, Heigenberg's uncertainty principle states that the momentum Δp and the position Δx of a particle, say of an electron, cannot be simultaneously determined precisely. A little more specifically, it is the multiplicative uncertainty relation given by $\Delta p \Delta x \geq h$ where h is the Planck's constant. If one aims to measure the position precisely, for example, the momentum cannot be determined precisely, and vice versa. In Signal Processing, there is a similar uncertainty relation between time and frequency resolution of a signal.

A. Representation via a Combined Dictionary

In the sparse representation theory by Donoho *et al.*, it is of interest to determine if a signal can be sparsely represented in an arbitrary pair of two different orthonormal bases A and B , simultaneously. A signal is uniquely representable by each basis. That is, each basis spans the vector space to which the signal belongs, i.e., $x \in \mathbb{R}^M$. An M dimensional column vector x is uniquely represented by an M dimensional basis A , as well as by B . In practice, we can find that one representation is more *effective* than the other. Effective here is to mean that the virtue is in the smallness of the sparsity: the sparser the representation is, the more it is effective. On the one hand, for example, using a wavelet basis would give a better time-localization result and thus it would be more suitable when the signals of interest are impulse-like ones. On the other hand, for rhythmic signals with high frequency contents, use of the Fourier basis would be more effective. Since one can encounter any type of signal in practice, it is interesting to see if there is any benefit to seek sparse signal representation in a combined dictionary, using two or more bases simultaneously. The combined dictionary can be constructed easily: concatenate the bases, matrix by matrix. For example, a dictionary matrix can be given by $D = [A; B]$, a $[M \times 2M]$ matrix. Thus, the problem becomes:

- (P1) Find the most sparse representation s ,
 given a signal $x = Ds$, using the dictionary $D = [A; B]$. (15)

The L1 minimization can again be used as the tool to draw the sparsest solution from the under-determined system of equations, i.e., $x = Ds$. One can expect that while one representation may give a poorer result, the other representation might give the more effective solution.

This study would be very interesting for the cases where the signal of interest is complex exhibiting multiple distinctive features each of which is identifiable in different bases such as time, frequency, and space. For example, EEG signals exhibit distinctive features in time, frequency, and space, and the most suitable basis for each is known.

Going back to our discussion of effective solution, the key questions we would like to ask then include:

- (Q1) Would the combined representation provide a result more effective than what can be expected of a single best representation? What is the benefit of the dual representation of the same signal in the combined framework?
- (DI) Obviously, parsimonious representation result can be obtained as we have discussed above with the example of rhythmic and impulse like signals. Say, A is a basis effective for impulse like signals and B for rhythmic signals. Suppose x_1 is a rhythmical signal. Then, its representations with respect to A and B are $x_1 = As_1$ and $x_1 = Bs_2$ respectively, and s_2 shall be more parsimonious solution than s_1 . What about when we choose to use the combined representation, i.e., $x_1 = [A; B] \begin{pmatrix} s_1 \\ s_2 \end{pmatrix} = Ds$? If we have an algorithm that seeks parsimonious representation, that algorithm will enforce the s_1 part to be zero and utilize the s_2 part for parsimonious solution. Thus, the solution of such an algorithm would take a form of $\begin{pmatrix} 0 \\ s_2 \end{pmatrix}$. Thus, we shall expect a nice result, i.e., $\|s\|_0 \approx \|s_2\|_0 \ll \|s_1\|_0$. If this is true, clearly there are benefits of using the combined representation: (i) the combined representation gives us the parsimonious solution, at the same level that we perfectly knew that the signal is rhythmical and used the right basis B ; (ii) the combined approach gives the benefit that a single framework can be used to deal with different classes of signals.
- (Q2) Would the idea of concatenating not only two but more bases in the dictionary produce more effective representation, overall?

- (D2) The matrix D will become fatter and fatter as we include more bases; the system of equations becomes more and more underdetermined. With more bases covered, the dictionary may be able to deal with more diverse classes of signals. But on the negative side, this may impose a burden on the uniqueness of the L1 solution. Given the number of bases in the dictionary, the level of sparsity that can be dealt with by the system equation (15) must be limited. It makes perfect sense to say that the uniqueness of the representation via the combined system should depend completely on the choice of both the two bases as well as the sparsity of the signal. When the two bases are selected in such a way that they are uncorrelated with each other, the system equation (15) can produce a one-to-one correspondent map between s and x as long as the sparsity level of s is small enough. If we let the sparsity level grow, there will be a certain point beyond which the unique representation becomes impossible. Thus, it would make perfect sense to avoid inclusion of bases that are correlated with each other. When there is a certain amount of correlation unavoidable, one may choose to remove them as much as possible, via one of the dimensionality reduction techniques such as the principal component analysis (PCA) and common spatial pattern (CSP) analysis, and put them into the dictionary.

We believe that the answers to these questions are all positive, while further research shall be followed to prove them. A combined representation framework could serve as a universal approach to represent a diverse group of signals each with distinct characteristics. Then, we ask the following more specific question.

- (Q3) Under what condition(s), would the L1 solution, or the L0 solution of (15), be unique?
- (D3) The level of under-determinedness, the number of bases in the dictionary *and the level of cross correlation between them*, must have bearings on the level of sparsity which (P1) in (15) can obtain. Namely, the level of concentration (P1) can achieve depends up on the dictionary itself, in particular the cross-correlation level.

For this very interesting problem, Donoho and Huo give the following results.

B. The Uncertainty Principle

Let the time domain signal x , $(x_t)_{t=0}^{M-1}$, has the sparsity K_t under transformation $x = I s_1$, where I is the identity, and its Fourier transform \tilde{x} , $(\tilde{x}_w)_{w=0}^{M-1}$, has the sparsity K_w under

$\tilde{x} = Fx$. Then, the two sparsity parameters should satisfy

$$K_t K_w \geq M \geq \frac{1}{\mu^2} \quad (16)$$

where μ is the maximum correlation of the two bases A and B , they are I and F in this particular example, in the dictionary,

$$\mu := \max_{i,j} \left\{ \left| \langle a_i, b_j \rangle \right| \right\} \quad (17)$$

and using the fact that $\mu \geq \frac{1}{\sqrt{M}}$ (See HW#2 for proving this) and that the arithmetic mean is larger than or equal to the geometric mean, it is also given by,

$$K_t + K_w \geq 2\sqrt{M} \geq \frac{2}{\mu}. \quad (18)$$

By this theorem, we can see that any signal of interest cannot be sparsely represented in both domains simultaneously. If the sparsity is in one domain, say K_t , is given and fixed, then, the sparsity level obtainable in the other domain shall be limited, i.e., $K_w \geq \frac{2}{\mu} - K_t$. We will sketch the proof here while leaving the exact proof as a problem in Homework set #2.

C. The Theorem on the Uniqueness of the L0 Optimization

In this section, we will show that the UP can be utilized to prove that the L0 solution is unique if $\|s\| < \frac{1}{\mu}$. If the solution is not unique given $\|s\| < \frac{1}{\mu}$, then the UP will be violated. In other words, the condition $\|s\| < \frac{1}{\mu}$ is a sufficient condition such that the L0 solution of (15) is unique.

The line of thoughts goes as follows:

- ① If $\|s\| < \frac{1}{\mu}$ and $x = Ds$ (15), then the L0 solution is unique. We aim to prove this statement. We will use the proof-by-contradiction routine (Hint: p and $\sim q$ leads to a contradiction, thus if p , then q).
- ② Let $\|s\| =: K_s < \frac{1}{\mu}$ and $x = Ds$; suppose there are two distinct non-zero K_s -sparse solutions $s_1, s_2 \in \mathbb{R}^{2M}$, i.e., $x = Ds_1, x = Ds_2$ respectively.
- ③ We have $D \underbrace{(s_1 - s_2)}_{=: d} = 0$. Let $d := (s_1 - s_2) \in \mathbb{R}^{2M}$.
- ④ On the sparsity of d , we can say,

$$\|d\|_0 \leq 2K_s. \quad (19)$$

⑤ Since $K_s < \frac{1}{\mu}$, (19) leads to

$$\|d\|_0 < \frac{2}{\mu} + 2. \quad (20)$$

⑥ But one should note that (20) is a contradiction since it violates the UP. Let us see it below.

⑦ Let us consider the L0 norm of d . Note that we can divide d into the top part with M elements, and the bottom part with the rest M elements. Let us denote them as d_{top} and d_{bottom} respectively. That is, $0 = A d_{top}$ and $0 = B d_{bottom}$. Then, we can write

$$\|d\|_0 = \|d_{top}\|_0 + \|d_{bottom}\|_0 \quad (21)$$

⑧ From the UP, we must have

$$\|d\|_0 = \|d_{top}\|_0 + \|d_{bottom}\|_0 \geq \frac{2}{\mu}. \quad (22)$$

⑨ The statement is now proved.

D. Uniqueness of the L1 Optimization

We now aim to find the sufficient condition for L1 uniqueness. We note that this is the central result and perhaps the most intriguing part in the Compressed Sensing and the sparse representation theory. The best resource for understanding this includes Donoho and Huo [13] and Elad and Bruckstein [14]. Note that the latter is an extension of the former. The latter has obtained a sufficient condition tighter than that of Donoho and Huo by running an optimization routine.

In this lecture note, we will generally follow their approaches but attempt to show along the way that these previous results can be further improved, at any place possible. The final result is thus a tighter sufficient condition, which is novel.

The L0 optimization is given by

$$s_{t_0} := \arg \min \|s\|_0 = \sum_{i=0}^{2M-1} s_j^0 \quad \text{s.t. } x = D s = D s_0 \quad (23)$$

whereas the L1 optimization is given by

$$s_{l_1} := \arg \min \|s\|_1 = \sum_{j=0}^{2M-1} |s_j| \quad \text{s.t. } x = Ds = Ds_0 \quad (24)$$

We first assume the sufficient condition, $\|s_0\|_0 \leq \frac{1}{\mu}$; then the L0 solution s_{l_0} is unique and it should be the case $s_{l_0} = s_0$, the true solution. We then aim to find a sufficient condition for the L1 uniqueness. In other words, find the sufficient condition under which the L1 solution is unique, and thus $s_{l_1} = s_{l_0}$?

Namely, we would like to show that $\|s\|_1 \geq \|s_{l_0}\|_1$ for any $s \in \mathbb{R}^{2M}$ satisfying $x = Ds$ if this sufficient condition is met.

Here are three possible lines of thoughts:

- ① The direct proof: Let us have $\|s_0\|_0 < \frac{1}{\mu}$. Show if $s \in \mathbb{R}^{2M}$ and $x = Ds$, implies $\|s\|_1 \geq \|s_{l_0}\|_1$ if the sufficient condition is met. That is, follow along the line of showing “ $\|s\|_1 \geq \|s_{l_0}\|_1$ for any feasible solution $s \in \mathbb{R}^{2M}$ and $x = Ds$,” and attempt to draw the condition which makes the statement inside “ ” to be true.
- ② The contra positive proof (if p , then q = if $\sim q$, then $\sim p$): Let us have $\|s_0\|_0 < \frac{1}{\mu}$. Find the sufficient condition so that $\|s\|_1 \not\geq \|s_{l_0}\|_1$ implies $x \neq Ds$.
- ③ The proof-by-contraction way: Let us have $\|s_0\|_0 < \frac{1}{\mu}$. Find a sufficient condition under which if $s \in \mathbb{R}^{2M}$ is a feasible solution, i.e., $x = Ds$, but its L1 norm is not greater than nor equal to the L1 norm of the L0 solution, i.e., $\|s\|_1 \not\geq \|s_{l_0}\|_1$, then it leads to a contradiction.

Let us follow the first line here:

- ① Let $\|s_0\|_0 = K \leq \frac{1}{\mu}$. Show that if $s \in \mathbb{R}^{2M}$ and $x = Ds = [A; B] \begin{pmatrix} s_A \\ s_B \end{pmatrix}$, implies $\|s\|_1 \geq \|s_{l_0}\|_1$. Recall $\mu := \max_{i,j} \{ \langle a_i, b_j \rangle \}$ where $(a_i)_{i=0}^{M-1}$ and $(b_i)_{i=0}^{M-1}$ are the columns of the orthonormal matrices A and B respectively.
- ② Since s is a feasible solution and s_0 is the true K -sparse solution, they both satisfy

$$x = Ds = Ds_0 \quad (25)$$

③ Then, we have $D(s - s_0) = 0$, which leads to $[A; B]d = 0$ where we let $d := s - s_0$.

④ Similar to what we have done in the unique L0 optimization, we can divide d into the top part with the first M elements, and the bottom part with the rest M elements. Let us denote them as d_A and d_B respectively. That is, $d := \begin{pmatrix} d_A \\ d_B \end{pmatrix}$. Then, from (25) we have

$$0 = Ad_A + Bd_B \quad (26)$$

⑤ Thus, any feasible solution can be written as $s = s_0 + d$. To show $\|s\|_1 \geq \|s_0\|_1$, we need to show

$$\sum_{i=0}^{2M-1} |s_{0,i} + d_i| - \sum_{i \in \mathcal{K}} |s_{0,i}| \geq 0 \quad (27)$$

⑥ Decomposing the first part into two mutually exclusive index sets, the support set \mathcal{K} and the non support set \mathcal{K}^c , i.e., $\mathcal{K} \cup \mathcal{K}^c = \{0, 1, 2, \dots, 2M-1\}$, we have

$$\sum_{i \in \mathcal{K}^c} |d_i| + \sum_{i \in \mathcal{K}} (|s_{0,i} + d_i| - |s_{0,i}|) \geq 0 \quad (28)$$

⑦ Let us use a lower bound such that

$$|s_{0,i} + d_i| - |s_{0,i}| \geq -|d_i| \quad (29)$$

⑧ Use (29) in L.H.S. of (28) and have the result greater than or equal to zero, i.e.,

$$\sum_{i \in \mathcal{K}^c} |d_i| - \sum_{i \in \mathcal{K}} |d_i| \geq 0 \quad (30)$$

⑨ Now, add $2 \sum_{i \in \mathcal{K}} |d_i|$ on both sides to get $\sum_{i=0}^{2M-1} |d_i| \geq 2 \sum_{i \in \mathcal{K}} |d_i|$. Namely,

$$\frac{\sum_{i \in \mathcal{K}} |d_i|}{\sum_{j=0}^{2M-1} |d_j|} \leq \frac{1}{2} \quad (31)$$

- ⑩ From the discussion so far, we note that if (31) holds for any null vector $d \in \mathcal{N}(D) := \{d \mid Dd = 0 \text{ for any } d \in \mathbb{R}^{2M}\}$, it implies that the L1 norm of any feasible solution is larger than that of the L0 solution, $\|s\|_1 \geq \|s_0\|_1$ (Recall that that was our objective in this section).
- ⑪ Now let us consider (31) a little more carefully. We first write it in the following form:

$$\sum_{i \in \mathcal{K}} \frac{|d_i|}{\sum_j |d_j|} \leq \frac{1}{2} \quad (32)$$

- ⑫ Consider each ratio first, $\frac{|d_i|}{\sum_j |d_j|}$, and aim to express the denominator as a multiple of the numerators $|d_i|$, i.e., $\sum_j |d_j| \geq |d_i| \tilde{K}$.

- ⑬ Then, we can write

$$\sum_{i \in \mathcal{K}} \frac{|d_i|}{\sum_j |d_j|} \leq \sum_{i \in \mathcal{K}} \frac{|d_i|}{|d_i| \tilde{K}} = \frac{K}{\tilde{K}} \quad (33)$$

- ⑭ And then, we can let it be smaller than or equal to 1/2 by making sure that $K \leq \frac{1}{2} \tilde{K}$. To do that, let us find out how large \tilde{K} can get.
- ⑮ Let us focus on a single index $i \in \mathcal{K}$. Without loss of generality, Let us assume it belongs to the top index set, i.e., $i \in [0, M-1]$. From (26), we note $Ad_A = -Bd_B$. Then, we have, $d_A = -A^T B d_B$ (Hint: they are orthogonal matrices). Since $|d_i|$ is the i -th element of d_A , we have

$$\begin{aligned}
|d_i| &= \left| -\left(A^T B\right)_{\text{row } i} d_B \right| \\
&\leq \left| (\mu \mu \cdots \mu) d_B \right| \\
&= \mu \|d_B\|_1
\end{aligned} \tag{34}$$

Thus, we get

$$\|d_B\|_1 \geq \frac{|d_i|}{\mu}. \tag{35}$$

⑩ Let us focus In addition, we have a trivial bound

$$\|d_A\|_1 \geq |d_i|. \tag{36}$$

⑪ Adding the two inequalities (35) and (36) we have

$$\|d_A\|_1 + \|d_B\|_1 \geq |d_i| + \frac{|d_i|}{\mu} \tag{37}$$

⑫ Thus, we note, $\|d\|_1 \geq |d_i| \left(1 + \frac{1}{\mu}\right)$. Using it to replace the L1 norm $\sum_j |d_j|$ in (33), we have

$$\sum_{i \in \mathcal{K}} \frac{|d_i|}{\sum_j |d_j|} \leq \sum_{i \in \mathcal{K}} \frac{|d_i|}{|d_i| \left(1 + \frac{1}{\mu}\right)} = \frac{K}{\left(1 + \frac{1}{\mu}\right)}. \tag{38}$$

⑬ At this point, we take $\tilde{K} = \left(1 + \frac{1}{\mu}\right)$, and let $\frac{K}{\tilde{K}} \leq \frac{1}{2}$, which leads to Donoho's result:

$$K \leq \frac{1}{2} \left(1 + \frac{1}{\mu}\right). \tag{39}$$

The result (39) implies that if the sufficient condition, $K \leq \frac{1}{2} \left(1 + \frac{1}{\mu}\right)$, is met, then the L1 solution is the L0 solution, and thus the exact solution.

The development so far implies that if $K \leq \frac{1}{\mu}$, the L0 solution is unique, and if $K \leq \frac{1}{2}\left(1 + \frac{1}{\mu}\right)$ ($< \frac{1}{\mu}$), then the L1 solution is the unique L0 solution. We note that

$$\frac{1}{2}\left(1 + \frac{1}{\mu}\right) < \frac{1}{\mu}, \quad (40)$$

and thus this may indicate that there is a price we have to pay using L1 optimization. However, since it is not a necessary condition, there is a room to explore further.

Elad and Bruckstein have attempted to narrow down this gap, via solving an optimization problem, and obtained a better bound

$$\frac{1}{2}\left(1 + \frac{1}{\mu}\right) < \frac{0.9142}{\mu} < \frac{1}{\mu}. \quad (41)$$

That optimization problem is recasting the problem in the following manner:

$$\begin{aligned} & \text{minimize } \frac{1}{2} \sum_{j=0}^{2M-1} |d_j| - \sum_{i \in \mathcal{K}} |d_i| \\ & \text{subject to } [A; B]d = 0 \end{aligned} \quad (42)$$

To understand this approach, let us recapitulate the Donoho's approach which was to find the sufficient condition on the sparsity $K \leq \frac{1}{2}\left(1 + \frac{1}{\mu}\right)$ such that if is met, any null space vector d , i.e., $[A; B]d = 0$ (25), satisfies the inequality $\frac{\sum_{i \in \mathcal{K}} |d_i|}{\sum_{j=0}^{2M-1} |d_j|} \leq \frac{1}{2}$. We can write the inequality in the following form, i.e., $\frac{1}{2} \sum_{j=0}^{2M-1} |d_j| - \sum_{i \in \mathcal{K}} |d_i| \geq 0$. This means, we note, that as long as the L.H.S. remains positive, the inequality $\frac{\sum_{i \in \mathcal{K}} |d_i|}{\sum_{j=0}^{2M-1} |d_j|} \leq \frac{1}{2}$ is satisfied. They aimed at minimizing the L.H.S. while not letting it go into the negative domain. For this, they included a couple of nice ideas added to (42).

2. The Uniform Uncertainty Principle

Candes and Tao have introduced the notion of uniform uncertainty principle (UUP) in [8]. It aims to define an $[M \times N]$ sensing matrix which obeys a “restricted isometry property (RIP).”

(Restricted Isometry Property) Candes and Tao [9] define that the K -restricted isometry constant δ_K of the $[M \times N]$ sensing matrix F is the smallest quantity such that

$$1 - \delta_K \leq \frac{\|Fv\|_2^2}{\|v\|_2^2} \leq 1 + \delta_K \quad (43)$$

for any K -sparse vector v sharing the same K nonzero entries as the K -sparse signal x . The interpretation of this property should be clear: Any K or smaller column collection of the sensing matrix F should behave like a unitary transformation, i.e., a length preserving transformation. This would make sense when the constant is small and certainly it should not be equal to 1, i.e., $\delta_K < 1$.

Remarks on Further Research

We note from the Coding Theory that designing a code, or equivalently a parity-check matrix of a code, which guarantees a certain minimum distance, which is referred to as the spark in CS theory, is generally considered as a challenging problem, especially for a large N . For a short block N , the problem is less difficult to handle and there are some codes that guarantee a certain minimum distance (a spark), such as the BCH and the Reed Solomon codes. In the approach of Candes and Tao, this difficulty is dealt with by defining a class of matrices which satisfy (43). Thus, the RIP condition (43), or the UUP, provides an *asylum* for the theory of Compressive Sensing. For those matrices satisfying the RIP, many good things are guaranteed to happen such as the L1 recovery is perfect. But it is an NP-complete problem to check if a matrix satisfies the RIP condition.

In both the Coding Theory and the Compressed Sensing, the method of random construction has been used. If we obtain a sensing matrix randomly, with high probability the matrix satisfies the RIP. For a coding problem, if we obtain a parity check matrix randomly, the code comes with a certain minimum distance with a high probability. Further discussion of this aspect is a good research direction to which we have some preliminary results. See Chapter VI.2 for our discussion on this direction.

A. Sufficient and Necessary Conditions for the Unique L0 Solution

Now we would like to ask for the condition that the L0 solution be unique under the setting that $y = Fx$ where F is an $[M \times N]$ sensing matrix and x is a K -sparse signal. Let us fix K and find the minimum M for good recovery. Let us consider the L0 recovery first followed by the L1 recovery.

- We first note that M should be at least larger than or equal to K , i.e., $M \geq K$. Why?

- Second, we note that for the unique L0 recovery we need $M \geq 2K$. Let us prove this now.

There are two approaches we can think of.

a) For a general sensing matrix F , we may use the spark S of F .

Lemma 1. $S > 2K$ is the sufficient and necessary condition for unique recovery of a K sparse signal under L0 minimization. In addition, if $S > 2K$, $M \geq 2K$; the converse does not hold.

Proof: Let the spark S of the matrix be strictly greater than $2K$, i.e., $S > 2K$. Suppose two distinct K -sparse solutions, say x_1 and x_2 . Then, $y = Ax_1$ and $y = Ax_2$; thus, $A(x_1 - x_2) = 0$. This implies $x_d := x_1 - x_2$ is in the null space of A . Note that from the definition x_d is a sparse vector whose sparsity is $2K$ at maximum. Thus, there exists a set of $2K$ columns of A that are linearly dependent. This contradicts $S > 2K$.

The Singleton bound dictates that $S \leq M + 1$. Thus, $M \geq 2K$ is a necessary condition for unique L0 solution.

Conversely, if $S \leq 2K$, then there exists two distinct K -sparse solutions, say x_1 and x_2 , to a same observation. Let us assume the worst case $S = 2K$. Since the spark of A is $2K$, there exists a $2K$ -sparse non-zero vector x_d , i.e., $Ax_d = 0$. We can express

$x_d := x_1 - x_2$ as the difference of two distinct K -sparse signals. Then, we can write $A(x_1 - x_2) = 0$ from which we obtain $y = Ax_1$ and $y = Ax_2$, two distinct solutions to the same observation y .

In addition, $M \geq 2K$ does not necessarily enforce $S > 2K$. There exists an $M \times 2K$ matrix whose spark is smaller than $2K$.

b) Now, let's go with a direction little bit different from the one given above (by Candes, Romberg and Tao [10])

Lemma 2. Let an $M \times N$ sensing matrix F be selected out from the $N \times N$ Fourier transform matrix where N is a prime number. M rows are selected randomly to construct the $M \times N$ sensing matrix. In this setting, $M \geq 2K$ is the sufficient and necessary condition for unique recovery of a K sparse signal under L_0 minimization. If $M \geq 2K$, then $S > 2K$.

The proof goes as follows: Let $M \geq 2K$. Suppose there are two distinct K sparse solutions $y = Ax_1$ and $y = Ax_2$ with support \mathcal{K}_1 and \mathcal{K}_2 respectively. Then, $A(x_1 - x_2) = 0$. The difference of the two $x_d := x_1 - x_2$ is a signal in the null space of A whose support is $\mathcal{K}_1 \cap \mathcal{K}_2$ and sparsity $|\mathcal{K}_1 \cap \mathcal{K}_2|$ can be as large as up to $2K$. Suppose the worst case that it is $2K$. Note that Ax_d is a linear transformation via the $2K$ columns of A into the space of dimension $M \geq 2K$ from the space of dimension $2K$. This transformation thus is injective; thus, for a non-zero input vector $x_d \neq 0$, the output vector cannot be zero, i.e. $Ax_d \neq 0$, which contradicts the supposition.

Conversely, if $M < 2K$, we can show that, there exist two distinct K -sparse solutions, x_1 and x_2 to $y = Ax$. Namely, the transformation $M \times 2K$ $A_{\mathcal{K}_1 \cap \mathcal{K}_2}$ cannot but be surjective. Thus, there exists a non-zero x_d such that $Ax_d = 0$.

This problem is a special case of a), our L_1 uniqueness done with the spark. The sensing matrix is selected as an $M < N$ number of randomly selected rows of the

$N \times N$ Fourier transform matrix with prime N . Because of the structure given by the Fourier transformation and prime N , the spark S of such an $M \times N$ matrix is always at its maximum $M + 1$, $S = M + 1$, achieving the Singleton bound. That is, if $M \geq 2K$, then $S > 2K$. Thus, a short proof for Lemma 2 is to use both Lemma 1 and this result.

(Fourier matrix with non prime N) Candes, Romberg and Tao [10] includes a discussion that for a non-prime N , the result does not hold any longer. This is because there exists nontrivial subgroups of \mathbb{Z}_N with addition modulo N . If the set of row indices and the support set are not the subgroups of \mathbb{Z}_N , the results still hold. This led to probabilistic argument when the support set and the measurement set are selected randomly. In such a setting, they argue that, with probability close to 1, the results hold.

- Now, let us have it expressed with the RIP constant.

Lemma 3. If a sensing matrix fulfills the RIP with its constant satisfying $\delta_{2K} < 1$, then L_0 minimization solution x_1 is the unique and exact solution to $y = Ax$, i.e., $x_1 = x$, where x is K -sparse.

The proof is very similar to what is given above with the proof with spark, and thus omitted.

B. Condition for the Unique L1 Solution

Here we would like to find the condition for the L1 solution to be identical to the L0 solution. We have learned that we need at least $M \geq 2K$ for the unique L0 solution. Expressing this as a factor of K , i.e., $M = O(\lambda K)$, we would expect that $\lambda > 2$ for the L1 solution to be unique. We also note that for the unique L0 solution, we need the RIP satisfied for δ_{2K} . For the unique L1 solution, we would expect that, the condition should get stricter, i.e., the RIP for δ_{3K} .

Candes and Tao [9] define another constant $\theta_{K,K'}$, the *restricted orthogonality constant*, to be the smallest quantity such that

$$\theta_{K,K'} \geq \frac{|\langle Fv, Fv' \rangle|}{\|v\|_2 \|v'\|_2} \quad (44)$$

for all K -sparse v and K' -sparse vector v' where the two supports sets are disjoint, $\mathcal{K} \cap \mathcal{K}' = \emptyset$.

Theorem 1. (Theorem 1.4 of Candes and Tao [9]) Let $K \geq 1$. Let the sensing matrix F be a matrix such that $\delta_K + \theta_{K,K} + \theta_{K,2K} < 1$. Then the L1 solution is the L0 solution.

Let us consider the proof of this theorem.

- First, note that the L0 solution is unique. Why?
- Let us call the L0 solution x .
- Let us call the L1 solution d .
- By definition, the L1 solution is the minimum L1 vector s.t. $y = Fd$; thus

$$\|d\|_1 \leq \|x\|_1. \quad (45)$$

- To show $d = x$, therefore, it suffices to show $\|d\|_1 \geq \|x\|_1$.
- We call the columns of F as f_j , i.e., $F = (f_1 \ f_2 \ \cdots \ f_N)$.
- This problem will be solved by duality.
- (**Lagrange duality**) Consider an optimization problem where we aim to minimize $J(d)$ subject to $Fd = z$, and assume that J is differentiable. Then, from the Lagrange multiplier optimality condition, i.e.,

$$\nabla J(d) - F^T v = 0 \quad (46)$$

we can find the Lagrange multiplier v . For d to be the solution, we set up the Lagrange multiplier v to satisfy $\nabla J(d) = F^T v$.

- In the L1 minimization case, the function J is given by $J(x) = \sum_j |x_j|$. The usual gradient vector in this case is not well defined since the function of absolute value is not differentiable at zero. But we note that the function J is at least a convex

function. In such a case, the notion of the *subgradient* can be used, in place of the usual gradient. See Chapter VII.1 for reference on the *subgradient* and the *subdifferential*. Namely, a subgradient of a convex function J is equal to the usual gradient at a point where the function is differentiable; but at a point where the function is not differentiable, it is any vector whose affine extension, the extended line along the vector, can be used as a linear *under-estimate* of the function at any point along the line. This is best explained with pictures. See Figure 13 and Figure 14. As can be seen in those pictures, there may exist many such subgradients at any non differentiable point of a function. A collection of all possible subgradients at a point x is called the *subdifferential* of the function, denoted as $\partial J(x)$. Namely, one can understand the subgradient as an extended version of the gradient. A subgradient $\nabla J(x)$ of a convex function J is thus the usual gradient vector at each point x ; at each point x where the function is not differentiable, it is a vector in the subdifferential of the function. It works as long as the function J is convex.

- **(Subgradient $\nabla J(x) \in \partial J(x)$)** The subgradient of $J(x) = \sum_j |x_j|$ is now well understood and thus can be used. First, note that the function J is not differentiable at the origin. But, secondly, we note that the function J is convex since it is sum of convex functions, the absolute value functions.
 - i. The absolute value function is convex, obviously. The absolute value function has its subdifferential as shown in Figure 14. Let $f : \mathbb{R} \rightarrow [0, \infty)$ be the function taking the absolute value of its argument. The function is -1 whenever its argument is negative; +1 whenever positive; it is the interval $[-1, 1]$ when its argument is 0.
 - ii. Thus, we can use a subgradient of J at x for (46), i.e., $(\nabla J(x))_j = \text{sgn}(x_j)$ for $j \in \mathcal{K}$ and $(\nabla J(x))_j \in [-1, 1]$ for $j \notin \mathcal{K}$. From (46), we have

$$(F^T v)_j = \text{sgn}(x_j) \quad \text{for } j \in \mathcal{K} \quad (47)$$

and

$$\left| (F^T v)_j \right| \leq 1, \quad \text{for } j \notin \mathcal{K} \quad (48)$$

- Note that $(F^T v)_j = \langle v, f_j \rangle$.
- Now note that if we find the Lagrange multiplier vector v that satisfies the following two properties, we can show that the minimum L0 solution x is the minimum L1 solution.
 - i. $\langle v, f_j \rangle = \text{sgn}(x_j)$ for all $j \in \mathcal{K}$, and
 - ii. $|\langle v, f_j \rangle| < 1$ for all $j \notin \mathcal{K}$,

where $\text{sgn}(x_j)$ is the sign of x_j and it is zero when x_j is zero.

- That is, by making the supposition that the Lagrange multiplier vector v satisfying the two properties exists, we aim to show first that the L1 norm of the L1

minimization solution d , will satisfy $\|d\|_1 \geq \|x\|_1$. This result can then be combined with the trivial convexity result such that $\|d\|_1 \leq \|x\|_1$ by the definition of L1 minimization, to produce the desired result $\|d\|_1 = \|x\|_1$. Thus, $d = x$.

$$\begin{aligned}
\|d\|_1 &= \sum_{j \in \mathcal{K}} |d_j - x_j + x_j| + \sum_{j \notin \mathcal{K}} |d_j| \\
&\geq \sum_{j \in \mathcal{K}} \text{sgn}(x_j)(x_j + (d_j - x_j)) + \sum_{j \notin \mathcal{K}} d_j \langle v, f_j \rangle \\
&= \sum_{j \in \mathcal{K}} |x_j| + \sum_{j \in \mathcal{K}} (d_j - x_j) \langle v, f_j \rangle + \sum_{j \notin \mathcal{K}} (d_j - 0) \langle v, f_j \rangle \\
&= \sum_{j \in \mathcal{K}} |x_j| + \left\langle v, \sum_{j \in \{1, 2, \dots, N\}} d_j f_j - \sum_{j \in \mathcal{K}} x_j f_j \right\rangle \\
&= \sum_{j \in \mathcal{K}} |x_j| + \langle v, y - y \rangle \\
&= \sum_{j \in \mathcal{K}} |x_j| \\
&= \|x\|_1
\end{aligned} \tag{49}$$

- Note that we have not used the condition in the theorem yet. Now, we need to show if we can construct such a vector v provided the sufficient condition of the theorem is satisfied. Namely, if the sensing matrix F is a matrix such that $\delta_K + \theta_{K,K} + \theta_{K,2K} < 1$, then we should be able to find such a vector v satisfying the two properties and thus the L1 solution is the L0 solution, i.e., $d = x$.

Candes and Tao have proved the theorem by proving the following two lemmas.

Lemma 4. (Lemma 2.1 of [9], Dual Sparse Reconstruction Property, L2 version) Let $K, K' \geq 1$ be such that $\delta_K < 1$, and x be a K -sparse real vector supported on \mathcal{K} . Then, there exists a vector $v \in \mathcal{H}$, the column space of F , such that $\langle v, f_j \rangle_H = x_j$ for all $j \in \mathcal{K}$. In addition, there is an “exceptional set” $\mathcal{E} \subset \{1, 2, \dots, N\}$ which is disjoint from \mathcal{K} , of size at most

$$|\mathcal{E}| \leq K' \tag{50}$$

and with the properties

$$\left| \langle v, f_j \rangle \right| \leq \frac{\theta_{K,K'}}{(1 - \delta_K) \sqrt{K'}} \|x\| \quad \text{for all } j \notin \mathcal{K} \cup \mathcal{E} \tag{51}$$

and

$$\left(\sum_{j \in \mathcal{E}} |\langle v, f_j \rangle|^2 \right)^{1/2} \leq \frac{\theta_{K,K}}{(1-\delta_K)} \|x\|. \quad (52)$$

Lemma 5. (Lemma 2.2 of [9] Dual Reconstruction Property, ell- ∞ version) Let $K \geq 1$ be such that $\delta_K + \theta_{K,2K} < 1$, and x be a K -sparse real vector supported on \mathcal{K} . Then, there exists a vector $v \in \mathcal{H}$, the column space of F , such that

$$\langle v, f_j \rangle = x_j \quad \text{for all } j \in \mathcal{K}. \quad (53)$$

In addition, v obeys

$$\left| \langle v, f_j \rangle \right| \leq \frac{\theta_{K,K}}{(1-\delta_K - \theta_{K,2K})} \quad \text{for all } j \notin \mathcal{K} \quad (54)$$

- We note that the result of Lemma 3 is what we aimed to find. Namely, one can find such a vector $v \in \mathcal{H}$ with the two properties.
- The inequality in (54) is in fact $\left| \langle v, f_j \rangle \right| \leq \frac{\theta_{K,K}}{(1-\delta_K - \theta_{K,2K})\sqrt{K}} \|x\|$ used by Candes and Tao to give the property *ii*. Note that they added a normalization constraint such that $\|x\| = \sqrt{K}$ at the start of the proof of Lemma 3 (See the first line in the Proof of Lemma 2.2 on page 4214), and thus it becomes (54).
- In order to enforce $\left| \langle v, f_j \rangle \right| \leq 1$, we need to have $\frac{\theta_{K,K}}{(1-\delta_K - \theta_{K,2K})} \leq 1$. Thus, the sufficient condition has been derived, $\theta_{K,K} \leq 1 - \delta_K - \theta_{K,2K}$.

The proofs of Candes and Tao on the two lemmas are rather tedious. Let us see if we can obtain a result in our own way and compare our result with that of Candes and Tao.

Again, here we aim to show that a vector $v \in \mathcal{H}$, the column space of F , can be found such that

$$\langle v, f_j \rangle = \text{sgn}(x_j) \quad \text{for all } j \in \mathcal{K}. \quad (55)$$

And, v must obey

$$\left| \langle v, f_j \rangle \right| \leq \frac{\theta_{K,K}}{(1-\delta_K - \theta_{K,2K})} \leq 1 \quad \text{for all } j \notin \mathcal{K} \quad (56)$$

provided that $\theta_{K,K} \leq 1 - \delta_K - \theta_{K,2K}$. In addition, do not forget to add $y = Fx$ and $\|x\|_0 = K$.

- Note that v is an $M \times 1$ vector, the same dimension as the columns $\{f_j\}$ of F .
- From $y = Fx$, we have

$$\begin{aligned} y &= Fx \\ &= F_{\mathcal{K}}x_{\mathcal{K}} \end{aligned} \quad (57)$$

- Now, let us decompose y into two $M \times 1$ parts.

$$y = v_1 + v_2 \quad (58)$$

- We can choose the first one $v_1 \in \mathcal{H}$ by

$$v_1 = F_{\mathcal{K}}(F_{\mathcal{K}}^T F_{\mathcal{K}})^{-1} \text{sgn}(x_{\mathcal{K}}) \quad (59)$$

Note that $\delta_{\mathcal{K}} + \theta_{\mathcal{K},\mathcal{K}} + \theta_{\mathcal{K},2\mathcal{K}} < 1$ implies that $\delta_{\mathcal{K}} < 1$. Thus, the inverse exists. Observe that $v_1 \in \mathcal{H}$ is a general $M \times 1$ vector, not a sparse vector.

- Then, from (58), we have

$$v_2 = y - v_1 \quad (60)$$

- Then, it is trivial to show that such a vector v_1 satisfies the first property (55).

■ From (59), we have $F_{\mathcal{K}}^T v_{\mathcal{K}} = \text{sgn}(x_{\mathcal{K}})$.

- Now, we only need to show that it satisfies the second property (56) as well, provided the sufficient condition $\delta_{\mathcal{K}} + \theta_{\mathcal{K},2\mathcal{K}} + \theta_{\mathcal{K},\mathcal{K}} \leq 1$, or perhaps a less tight condition we can find, holds.

■ Let us now find a sufficient condition such that $|\langle v_1, f_j \rangle| \leq 1$ for all $j \notin \mathcal{K}$.

$$|\langle v_1, f_j \rangle| = \left| \left[F_{\mathcal{K}^c}^T F_{\mathcal{K}} \right]_{j\text{th row}} (F_{\mathcal{K}}^T F_{\mathcal{K}})^{-1} \text{sgn}(x_{\mathcal{K}}) \right| \quad (61)$$

- Note that from the definition of restricted isometry constant $\delta_{\mathcal{K}}$, we have for $\forall v_{\mathcal{K}} \neq 0$, by letting $\|v\|_2 = 1$.

$$1 - \delta_{\mathcal{K}} \leq v_{\mathcal{K}}^T F_{\mathcal{K}}^T F_{\mathcal{K}} v_{\mathcal{K}} \leq 1 + \delta_{\mathcal{K}} \quad (62)$$

Let the minimum and the maximum eigenvalue of the symmetric matrix $F_{\mathcal{K}}^T F_{\mathcal{K}}$ be $\lambda_{\min}(F_{\mathcal{K}}^T F_{\mathcal{K}})$ and $\lambda_{\max}(F_{\mathcal{K}}^T F_{\mathcal{K}})$ respectively. In fact, the norm of the symmetric matrix $F_{\mathcal{K}}^T F_{\mathcal{K}}$ is the square root of the maximum eigenvalue of the symmetric matrix. See (178).

- Then, we have other inequality

$$\lambda_{\min}(F_{\mathcal{K}}^T F_{\mathcal{K}}) \leq \|F_{\mathcal{K}}^T F_{\mathcal{K}}\|_2 := \lambda_{\max}(F_{\mathcal{K}}^T F_{\mathcal{K}}) \quad (63)$$

- Then, we can say that $1 - \delta_K \leq \lambda_{\min}(F_{\mathcal{K}}^T F_{\mathcal{K}})$ and $\lambda_{\max}(F_{\mathcal{K}}^T F_{\mathcal{K}}) \leq 1 + \delta_K$ from the definition of RI constant.
- See also the discussion that these inequalities stay hold in the limit of large M and for a fixed support set \mathcal{K} of size K . Candes and Tao on page 4209 of [9] discuss this issue. In addition, $\lambda_{\min}(F_{\mathcal{K}}^T F_{\mathcal{K}})$ converges to $(1 - \sqrt{\frac{K}{M}})^2$, and $\lambda_{\max}(F_{\mathcal{K}}^T F_{\mathcal{K}})$ does to $(1 + \sqrt{\frac{K}{M}})^2$. Thus, substituting the convergence results to the inequalities, we have $\delta_K \approx 2\sqrt{\frac{K}{M}} + \frac{K}{M}$. But the following example shows that something may have gone wrong there. Let $\frac{K}{M} = 1/2$, then $\delta_K \approx 2\sqrt{\frac{1}{2}} + \frac{1}{2} = \sqrt{2} + \frac{1}{2} \approx 1.4 + 0.5 = 1.9$, a number greater than 1!!! Thus, a caution must be used for that result!!!
- Now using (63) or the inequality $1 - \delta_K \leq \lambda_{\min}(F_{\mathcal{K}}^T F_{\mathcal{K}})$

$$\|(F_{\mathcal{K}}^T F_{\mathcal{K}})^{-1}\| \leq \frac{1}{1 - \delta_K}$$

- This means

$$\begin{aligned}
|\langle v_1, f_j \rangle| &= \left| \left[F_{\mathcal{K}^c}^T F_{\mathcal{K}} \right]_{j\text{th row}} \left(F_{\mathcal{K}}^T F_{\mathcal{K}} \right)^{-1} \text{sgn}(x_{\mathcal{K}}) \right| \\
&\leq \left| \left[F_{\mathcal{K}^c}^T F_{\mathcal{K}} \right]_{j\text{th row}} \text{sgn}(x_{\mathcal{K}}) \right| \left\| \left(F_{\mathcal{K}}^T F_{\mathcal{K}} \right)^{-1} \right\| \\
&\leq \frac{1}{1 - \delta_{\mathcal{K}}} \left| \left[F_{\mathcal{K}^c}^T F_{\mathcal{K}} \right]_{j\text{th row}} \text{sgn}(x_{\mathcal{K}}) \right| \tag{64} \\
&\leq \frac{1}{1 - \delta_{\mathcal{K}}} |(\mu \mu \cdots \mu) \text{sgn}(x_{\mathcal{K}})| \\
&= \frac{\mu K}{1 - \delta_{\mathcal{K}}} \leq 1
\end{aligned}$$

- Thus, the sufficient condition is that $K \leq \frac{1 - \delta_{\mathcal{K}}}{\mu}$ or $\mu \leq \frac{1 - \delta_{\mathcal{K}}}{K}$. We have defined

$$\mu = \max_{i \neq j} |\langle f_i, f_j \rangle| \tag{65}$$

- It would be interesting to compare this with the result of Candes and Tao. Namely, which one is tighter, $\delta_{\mathcal{K}} + K\mu \leq 1$ vs. $\delta_{\mathcal{K}} + \theta_{\mathcal{K},\mathcal{K}} + \theta_{\mathcal{K},2\mathcal{K}} < 1$?
- It would be also interesting to compare this result with $K \leq \frac{1}{2} \left(1 + \frac{1}{\mu} \right)$ (39) or to that of Elad and Bruckstein $K \leq \frac{0.9142}{\mu}$.

Summarizing the results so far, we have Theorem 2.

Theorem 2. (Sufficient condition for the *L1 uniqueness*) Let $K \geq 1$. Let $\mu = \max_{i \neq j} |\langle f_i, f_j \rangle|$. Let F be a sensing matrix such that $\delta_K + \mu K < 1$. Then, the L1 solution is the L0 solution.

- The sufficient condition in Theorem 2 can be made by a better bound at (63). That is, we use

$$\left\| (F_K^T F_K)^{-1} \right\| \leq \frac{1}{\lambda_{\min}(F_K^T F_K)}. \quad (66)$$

- Then, the sufficient condition becomes $K \leq \frac{\lambda_{\min}(F_K^T F_K)}{\mu}$.

3. Uniqueness Proofs with Mutual Coherence μ

Candes and Tao's L0 and L1 uniqueness sufficient conditions are given in terms of the RI constants such as δ_{2K} and $\theta_{K,K}$. Uniqueness proofs based on these parameters are not convenient in the aspect that they are difficult to find and check for a given matrix. However, those based on mutual coherence μ is much easier to compute and check for a given matrix. It is of interest to find the sufficient conditions for solution uniqueness in terms of the mutual coherence μ as well as to the restricted isometry constants. We aim to end this section with an attempt to obtain a new set of sufficient conditions for the L0 and L1 unique solutions with respect to the mutual coherence μ so that the result can be applied to any $M \times N$ compressed sensing matrices.

A. Comparison of $\theta_{K,K} + \theta_{K,2K}$ and $K\mu$

It would be interesting to compare the two quantities, $\theta_{K,K} + \theta_{K,2K}$ and $K\mu$. First, this comparative study shall give us the insight as to which one is better as the sufficient condition for the L0 uniqueness. Second, the meaning of these parameters can be learned through this process.

First, let us show that $\theta_{K,K'} \leq \delta_{K+K'}$.

Note that $\delta_{K+K'}$ is the RIP constant for $K+K'$ -sparse signals. For example, every $K+K'$ columns should be linearly independent. In addition, there are two mutually exclusive support sets \mathcal{K} and \mathcal{K}' . Without loss of generality, let us assume unit energy K -sparse and K' -sparse signals v and v' , i.e., $\|v\| = \|v'\| = 1$. Then, the RIP can be given by the following inequalities:

$$1 - \delta_{K+K'} \leq \frac{\|Fv + Fv'\|_2^2}{\|v + v'\|_2^2} \leq 1 + \delta_{K+K'} \quad (67)$$

In other form, it can also be given by the following inequalities

$$1 - \delta_{K+K'} \leq \frac{\|Fv - Fv'\|_2^2}{2} \leq 1 + \delta_{K+K'}$$

which can be written, after multiplied by -1, as

$$-1 - \delta_{K+K'} \leq -\frac{\|Fv - Fv'\|_2^2}{2} \leq -1 + \delta_{K+K'} \quad (68)$$

Adding (67) and (68), we have

$$-\delta_{K+K'} \leq \frac{\|Fv + Fv'\|_2^2 - \|Fv - Fv'\|_2^2}{4} \leq \delta_{K+K'} \quad (69)$$

The term at the center is $|\langle Fv, Fv' \rangle|$ and we have shown that $|\langle Fv, Fv' \rangle| \leq \delta_{K+K'}$.

Similar development gives

$$\delta_{K+K'} \leq \max(\delta_K, \delta_{K'}) + \theta_{K,K'} \quad (70)$$

Thus, we have

$$\theta_{K,K'} \leq \delta_{K+K'} \leq \max(\delta_K, \delta_{K'}) + \theta_{K,K'} \quad (71)$$

Now let us compare $\theta_{K,K} + \theta_{K,2K}$ with $K\mu$.

Note that we have defined the constant $\theta_{K,K'}$, the *restricted orthogonality constant*, to be the smallest quantity such that

$$\theta_{K,K'} \geq \frac{|\langle Fv, Fv' \rangle|}{\|v\| \|v'\|} \quad (72)$$

for all K -sparse v and K' -sparse vector v' where the two supports sets are disjoint, $\mathcal{K} \cap \mathcal{K}' = \emptyset$.

Thus, it is the supremum of all possible values in the R.H.S.

Note that the maximum correlation can be written as

$$\begin{aligned} \mu &= \max_{i,j} |\langle f_i, f_j \rangle| \\ &= \max_{\text{all 1-sparse signals } v, v'} |\langle Fv, Fv' \rangle| \end{aligned} \quad (73)$$

Thus, we can say μ is a special case of $\theta_{1,1}$.

Now, can we compare $\theta_{1,1}$ and $\theta_{1,2}$.

$$\begin{aligned}
\theta_{1,2} &= \sup \left| \langle f_i, Fv \rangle \right| \\
&= \sup \left| f_i^T Fv \right| \\
&= \sup \left| \left(v_{i_1} f_i^T f_{i_1} + v_{i_2} f_i^T f_{i_2} \right) \right| \\
&= \sup \left(\left| v_{i_1} \right| \mu + \left| v_{i_2} \right| \mu \right) \\
&= \mu \sup \left(\left| v_{i_1} \right| + \left| v_{i_2} \right| \right) \\
&= \sqrt{2} \mu
\end{aligned} \tag{74}$$

Note that the first inequality is obvious by the definition of maximum correlation μ . The second inequality is due to the normalization, i.e., $v_{i_1}^2 + v_{i_2}^2 = 1$. That is, given a vector v with the constraint $v_{i_1}^2 + v_{i_2}^2 = 1$, $\sup_{v \in \mathbb{R}^2, v_{i_1}^2 + v_{i_2}^2 = 1} \left(\left| v_{i_1} \right| + \left| v_{i_2} \right| \right) = \sqrt{2}$.

How about $\theta_{2,2}$?

$$\begin{aligned}
\theta_{2,2} &= \sup \left| \langle Fv, Fv' \rangle \right| \\
&= \sup \left| \langle f_{i_1} v_{i_1} + f_{i_2} v_{i_2}, f_{i_3} v_{i_3} + f_{i_4} v_{i_4} \rangle \right| \\
&= \sup \left| \left(f_{i_1} v_{i_1} + f_{i_2} v_{i_2} \right)^T \left(f_{i_3} v_{i_3} + f_{i_4} v_{i_4} \right) \right| \\
&= \sup \left| \left(v_{i_1} f_{i_1}^T f_{i_3} v_{i_3} + v_{i_1} f_{i_1}^T f_{i_4} v_{i_4} + v_{i_2} f_{i_2}^T f_{i_3} v_{i_3} + v_{i_2} f_{i_2}^T f_{i_4} v_{i_4} \right) \right| \\
&= \sup \mu \left(\left| v_{i_1} \right| \left| v_{i_3} \right| + \left| v_{i_1} \right| \left| v_{i_4} \right| + \left| v_{i_2} \right| \left| v_{i_3} \right| + \left| v_{i_2} \right| \left| v_{i_4} \right| \right) \\
&= \mu \sup \left(\left| v_{i_1} \right| + \left| v_{i_2} \right| \right) \left(\left| v_{i_3} \right| + \left| v_{i_4} \right| \right) \\
&= 2\mu
\end{aligned} \tag{75}$$

Lemma 6. Let $\|v\|_2 = 1$. Then, $\sup \|v\|_1 = \sqrt{K}$ s.t. $\|v\|_0 = K$. In addition,

$$\theta_{K,K} = \mu \|v\|_1 \|v'\|_1 = K\mu \quad \text{and} \quad \theta_{K,2K} = \sqrt{K} \sqrt{2K} \mu = \sqrt{2} K \mu.$$

Use the result that $\theta_{K,K'} \leq \delta_{K+K'}$, and $\delta_K + \mu K < 1$. Then, we have $\theta_{\frac{K}{2}, \frac{K}{2}} + \mu K < 1$ which leads to $\frac{K}{2} \mu + \mu K < 1$. Finally we have $K < \frac{2}{3} \frac{1}{\mu}$. If $K < \frac{2}{3} \frac{1}{\mu}$, then the L1 minimization gives the unique sparsest solution. This statement, however, has a problem because we have used a lower bound, $\theta_{K,K'} \leq \delta_{K+K'}$, to replace δ_K , rather than using an upper bound. Now we aim to replace the RIP constant δ_K with the maximum absolute correlation μ . Holger Rauhut has obtained an upper bound.

B. Connections to Other Results

Proposition (H. Rauhut) Let $A \in \mathbb{C}^{M \times N}$ with unit norm columns, coherence μ , 1-coherence function $\mu_1(K) = \max_{I \subset [N]} \max_{\substack{\mathcal{K} \subset [N] \setminus \{I\} \\ |\mathcal{K}| \leq K}} \sum_{j \in \mathcal{K}} |\langle a_j, a_I \rangle| \leq K\mu$, and restricted isometry constant δ_K . Then,

- $\mu = \delta_2$,
- $\mu_1(K) = \max_{\mathcal{K} \subset [N], |\mathcal{K}| \leq K+1} \|A_{\mathcal{K}}^* A_{\mathcal{K}} - e\|_{1 \rightarrow 1}$ where $\|A\|_{p \rightarrow p} := \max_{\|x\|_p=1} \|Ax\|_p$ is the operator norm of a matrix from L_p norm into L_p norm,
- $\delta_K \leq \mu_1(K-1) \leq \mu(K-1)$.

Substituting the third result of Rauhut, i.e., $\delta_K \leq \mu(K-1)$, into the result of Theorem 2, i.e., $\delta_K + \mu K < 1$ (64), we have the following bound, $\delta_K + \mu K \leq \mu(K-1) + \mu K < 1$, which leads to

$$K \leq \frac{1}{2} \left(1 + \frac{1}{\mu} \right). \quad (76)$$

This is a new result obtained in this lecture note. We have obtained a new sufficient condition based on the maximum correlation for the unique L1 solution. Note that this sufficient condition is exactly what we have obtained with the dictionary made of two ortho normal basis in the previous section, i.e., $K \leq \frac{1}{2} \left(1 + \frac{1}{\mu} \right)$ (39). This is a surprising result that the condition holds also for a general $M \times N$ sensing matrices.

Theorem 3. (Sufficient condition for the L1 uniqueness) Let $\mu = \max_{i \neq j} |\langle f_i, f_j \rangle|$. Let F be any $M \times N$ sensing matrix with mutual coherence μ . Let the sparsity of the signal x satisfy $K \leq \frac{1}{2} \left(1 + \frac{1}{\mu} \right)$. Then, the L1 solution is the L0 solution.

Remarks:

- Is $\frac{1}{\mu}$ a sufficient condition for uniqueness on L0 solution for the general $M \times N$ cases as well? Donoho and Elad have obtained that the sufficient condition for L0 uniqueness is also $K \leq \frac{1}{2} \left(1 + \frac{1}{\mu} \right)$. Obviously, this can be improved.
- The result in Theorem 3 can be also compared to that of Elad and Bruckstein $K \leq \frac{0.9142}{\mu}$ where two orthogonal basis were used. Thus, the rate is 1/2, i.e., $N = 2M$.

Now, a research question is to find out if Elad and Bruckstein's results can be generalized as well to general $M \times N$ cases. This should be an interesting research direction.

A bad approach. One may think the following direction may work for improving the L0 uniqueness condition; but it turns out that it leads to a useless bound. We know that $0 < \delta_{2K} < 1$ is a sufficient condition that the L0 solution is unique due to Candes and Tao. Now let us suppose we choose one from the following relations, i.e., $\mu K = \theta_{K,K} \leq \delta_{2K} \leq (2K-1)\mu$. Note that the lower bound, $\mu K \leq \delta_{2K}$ and letting $\mu K \leq 1$, should not be used since a number smaller than δ_{2K} does not enforce $\delta_{2K} < 1$. A number larger than δ_{2K} that is also smaller than 1, can enforce $\delta_{2K} < 1$. The only relation that we can have is $\delta_{2K} \leq (2K-1)\mu$. But this is Theorem 3.

The other direction is to use a lower bound on the mutual coherence. The following lemma states it.

Lemma 7. If $K \leq \frac{1}{2} \frac{N}{(N-1)\mu^2 + 1}$, then $K \leq \frac{M}{2}$.

Proof. The mutual coherence is lower bounded by $\mu(M, N) \geq \sqrt{\frac{N-M}{M(N-1)}}$. Rearranging this

with respect to M , we obtain $M \geq \frac{N}{(N-1)\mu^2 + 1}$. We have the lower bound $\frac{N}{(N-1)\mu^2 + 1}$

be larger than $2K$. This will guarantee that $M \geq 2K$.

Q.E.D.

See Lemma 1 and Lemma 2 we have proved earlier in pg. 53 (L0 uniqueness conditions). We write them here more explicitly.

Lemma 8. Let the spark of a given matrix A be greater than $2K$. Then, $2K \leq M$ is necessary. But the converse does not hold in general.

Proof. Let the spark S of a given matrix A be greater than $2K$, $S > 2K$. From the Singleton bound, $S \leq M + 1$. Thus, $M \geq S - 1 \geq 2K$. The converse does not hold because the spark of an $M \times 2K$ matrix can have any spark, i.e., $S \in \{2, 3, \dots, M + 1\}$.

Q.E.D.

Remarks. Let the dimension of a given sensing matrix A be $M \times N$ with $M \geq 2K$. Let $y = Fx$ and x be K -sparse. We wish to be in a position that the L0 minimization solution subject to $y = Fx$ gives the unique and exact solution. The statement is true when the sensing matrix is made of Fourier transform matrix (with prime N) or of the Vandermonde frame. But it is not true in general. When the sensing matrix is constructed from i.i.d. Gaussian distribution, what is found in the literature is $M \geq \lambda K$ where $\lambda = O(\log(N/K))$, then, a random sensing matrix satisfies the RIP condition, and thus, the L0 solution is unique

and exact. See our discussion in Chapter III.4.

Theorem 4. (the L0 uniqueness theorem in terms of mutual coherence) Let an $M \times N$ matrix A have mutual coherence μ . Let $y = Ax$. Let λ be the oversampling factor. If the sparsity of signal x satisfies $K \leq \frac{1}{\lambda} \frac{N}{(N-1)\mu^2 + 1}$, then L0 solution is unique and exact with probability close to 1.

Discussion with M. Elad at SPARS'11: Elad mentioned that Candes once had aimed at improving and obtaining a result similar to that of Theorem 4 involving $\frac{1}{\mu^2}$. We need to find this source.

4. Ensembles of Random Sensing Matrices

Candes and Tao [8] studied three kinds of ensembles of random matrices, random Gaussian, random Bernoulli, and the Fourier transform matrices (rows selected at random from unitary Fourier transform matrix), to see when they satisfy the uniform uncertainty principle or the RIP. They have studied the eigenvalues of the Gram matrix, $F^T F$, and related them to the probability with which the RIP with a certain value of δ_{2K} is satisfied. Their focus has been to draw a condition on the number of measurements needed to make the random sensing matrix to be stable so that it can be used to find the K -sparse signal exactly under L1 recovery criteria with a large probability.

A. The $\log(N)$ factor for random ensembles

Namely, the main result states that if the number of measurements obeys

$$O(K \log \frac{N}{K}) \quad (77)$$

then, the L1 minimization reconstructs the unique and exact solution with probability close to 1. Similar results hold for the three kinds of random matrices. Since these results are given in a similar format, let us give the result on the Fourier case here.

Theorem 1. (Candes [33]) Let y be M Fourier coefficients with its frequencies selected uniformly at random. Let x be K -sparse. Then, the $M \times 1$ Fourier coefficient vector can be written as $y = Fx$ where the M rows of F are taken from the rows of the $N \times N$ Fourier transform matrix and the M row indices correspond to the M selected frequencies. N here needs not be prime. Let the number of measurements satisfy

$$M \geq cK \log N. \quad (78)$$

Then, minimizing L1 norm reconstructs x with overwhelming probability. In details, if the constant c is of the form $22(\delta+1)$, then, the probability of success exceeds $1 - O(N^{-\delta})$.

In [33], Candes summarized the up-till-then results on Gaussian measurements. The results was that if

$$M \geq cK \log(N / K) \quad (79)$$

then, the sensing matrix satisfies the RIP condition such that $\delta_{2K} + \delta_{3K} < 1$ or better, $\delta_{2K} + \theta_{K,2K} < 1$ with probability exceeding $1 - O(e^{-\gamma N})$ for some $\gamma > 0$, and thus minimizing L1 reconstructs the unique and exact solution x . For binary measurements, it was conjectured that the similar result would hold.

B. The $\log(N)$ factor, is it really needed for a random sensing matrix?

It was an interest of Park and Lee in [34] to find out how the logarithmic dependence on the block length N came to exist in determining the number of measurements M when random Gaussian matrices are used. They noted that the signal recovery can be divided into two parts, the support set recovery part which involves complexity of order $\binom{N}{K}$ and the other is the signal value recovery part given the support which involves the complexity of finding the K non-zero values via the least squares solution (5).

We know that the second part is simple once the support set is given. Thus, in [34], they aimed to remove the impact of the second part by considering binary alphabet for non-zero values of the K sparse signal, the simplest of all. They then moved on to investigate how many measurements M , for a given N , are needed for a Gaussian matrix to satisfy the Restricted Isometry Property (RIP) with high probability.

To our surprise, the result is still the same. That is, the number of measurements sufficient to satisfy the RIP with high probability is still $M = O(K \log(N/K))$. In one aspect, this result may imply that the size of the alphabet is not a determining factor to the number of measurements, but the support set recovery part is. In another aspect, the bound involved in the calculation of the sufficient condition, i.e. the union bound, is perhaps too loose to tell any thing precisely at all. To determine what to do next, it should be of interest to describe the derivation to the result, which is given as follows:

- The aim again, is to obtain a sufficient condition on the number of measurements when the non-zero elements of the K -sparse signal are taken from the binary set $\{0,1\}$.
- First, they obtained a result in which the energy of the measurement vector $\|y\|_2^2 = \|Ax\|_2^2$ is a chi-square random variable with M degrees of freedom, with mean K and variance $2K^2/M$. It should be noted that it is not a too difficult task to bound a tail event probability of this well known random variable.
- Second, they use the result in the first step to get an upper bound on the probability of a large deviation event $\left\{ \left| \|Ax\|_2^2 - \|x\|_2^2 \right| \geq \delta \|x\|_2^2 \right\}$, $\delta \in (0,1)$, for a specific binary K sparse signal x . They use the Chernoff bound to tail bound the event. The result is $2e^{-M\delta/2} (\delta+1)^{M/2}$, an upper bound on the probability of the two tails (both ends). Namely, the large deviation event occurs with probability less than $2e^{-M\delta/2} (\delta+1)^{M/2}$. An obvious check is to see if M increases, the probability gets smaller.
- Third, we move on to require the matrix A , the collection of $M \times 1$ columns, to satisfy the restricted isometry property with the RIP constant δ_K for all possible binary K sparse signals, and aim to upper bound the probability of violation. From the perspective of success, we should check if each and every set of K columns of matrix A is

independent. From the perspective of failure, we should check if any set of K columns violates the restricted isometry condition. We use the second approach and the union bound.

- Then, the upper bound on the probability of violation is obtained simply by multiplying the size of the collection of K -sparse support sets $\binom{N}{K}$ to the result of the second step, i.e., $2e^{-M\delta/2}(\delta+1)^{M/2}\binom{N}{K}$. Exponential expression of this result gives the sufficient condition, i.e., $M \geq O(K \log(N/K))$; thus, the bound is of the form $P_f \leq e^{-c_1(M - c_2 K \log(\frac{N}{K}))}$ for some constants c_1 and c_2 .
- In the following section, we provide a little more detail in this process.

Remark 1. We make note of the fact that the $\log(N)$ factor maybe is the result of union bounding. Union bounds are not tight, especially when we have M approaching K . Tight union bounding techniques such as the Gallager's random coding or tight bounding techniques can be useful to obtain tight results. Before deciding to pursuing this direction, though, it would be wise to study a lower bound, perhaps using the Fano's inequality, and check if the lower bound has the $\log(N)$ factor or not. See Chapter IV.4.

Remark 2. Suppose using a Vandermonde measurement frame instead of the Gaussian matrices. The use of Vandermonde matrices is one of the important factors in reducing the number of measurements. The reason is that they guarantee that any set of M or less column vectors of an $M \times N$ Vandermonde matrix is linearly independent by design. In other words, the probability of a large deviation event $\left\{ \left| \|\mathbf{A}x\|_2^2 - \|x\|_2^2 \right| \geq \delta \|x\|_2^2 \right\}$ is exactly zero for any x as long as $M \geq 2K$; and thus, the union bound is zero as well. Hence, the probability that the Vandermonde frame satisfies the RIP is exactly one if $M \geq 2K$. In other words, the probability that the Vandermonde frame does not satisfy the RIP is exactly zero if $M \geq 2K$.

C. The $\log(N)$ factor, derived for binary K -sparse signals
Let an $M \times N$ matrix \mathbf{A} be constructed from i.i.d. Gaussian $\mathcal{N}(0, \frac{1}{M})$. Let x be any binary signal whose support set is randomly drawn from S , the collection of all distinct subsets of $\{1, 2, \dots, N\}$ of size K . Thus, we aim to investigate the support recovery problem by considering binary K -sparse signals. Note that the matrix \mathbf{A} and the signal vector x are independently drawn.

The event that a sample matrix \mathbf{A} fails to satisfy the following condition for a given particular x whose support is a subset of S , and for a given $\delta \in (0, 1)$

$$(1 - \delta)\|x\|_2 \leq \|\mathbf{A}x\|_2 \leq (1 + \delta)\|x\|_2 \quad (80)$$

is defined as $Fail(\mathbf{A}, x)$ or the success event as $Success(\mathbf{A}, x)$. Note that the only random

part of this event is in matrix \mathbf{A} .

We have a result in [34] that the probability of failure is identical for each x in S , and is given by

$$\Pr(\text{Fail}(\mathbf{A}, x)) \leq 2 \exp\left(-\frac{M\delta}{2}\right) (\delta+1)^{\frac{M}{2}}. \quad (81)$$

Note that the observation vector $\mathbf{y} = \mathbf{A}x$, a summation of K randomly chosen columns of Gaussian \mathbf{A} , is a Gaussian random vector. Thus, the probability of such an event can be easily defined and its upper bound can be obtained. We note that this probability should depend only on the variance of the Gaussian, and thus M , and the cardinality of the support set K .

We are now interested in the event that a random matrix \mathbf{A} , a collection of N randomly chosen columns of Gaussian, satisfies (80) for each and every support set $\mathcal{K} \in S$. This event can be defined as $\text{Success}(\mathbf{A})$. In other words, we note that the matrix \mathbf{A} should satisfy the condition (80) for each and every $\mathcal{K} \in S$, i.e., $\text{Success}(\mathbf{A}) = \bigcap_{x \in S} \text{Success}(\mathbf{A}, x)$. The failure event $\text{Fail}(\mathbf{A})$ can be described as the union of each failure event, i.e., $\text{Fail}(\mathbf{A}) = \bigcup_{x \in S} \text{Fail}(\mathbf{A}, x)$. Hence, the probability of $\text{Fail}(\mathbf{A})$ is bounded by

$$\begin{aligned} P_f &:= \Pr\left(\bigcup_{x \in S} \text{Fail}(\mathbf{A}, x)\right) \\ &\leq \sum_{x \in S} \Pr(\text{Fail}(\mathbf{A}, x)) \quad , \\ &= \binom{N}{K} \Pr(\text{Fail}(\mathbf{A}, x_1)) \end{aligned} \quad (82)$$

where $\Pr(\text{Fail}(\mathbf{A}, x_1)) \leq 2 \exp\left(-\frac{M\delta}{2}\right) (\delta+1)^{\frac{M}{2}}$. Hence the probability $\text{Fail}(\mathbf{A})$ is bounded by

$$\begin{aligned} P_f &\leq 2(eN/K)^K \exp\left(-\frac{M\delta}{2}\right) (\delta+1)^{\frac{M}{2}} \\ &\leq 2 \exp\left(-\frac{M}{2}(\delta - (\delta+1)) + K[\log(eN/K)]\right) \end{aligned} \quad (83)$$

Since $\delta \in (0,1)$, $\delta - (\delta+1)$ must be positive, note that L.H.S. can converge to zero for the sufficiently large K and $M = O(K \log(N/K))$.

5. Stable Recovery Property

Signals are not exactly sparse, in practice. That is, the signal may have small portion of dominant transform coefficients and large portion of transform coefficients that are indeed close to zero. But in many cases, the small transform coefficients are not exactly zeros. Thus, signals are sparse in a loose sense. There exists a model mismatch in a sparse model. In addition, there exist measurement noise, always. When signals are measured, they are contaminated with noise.

Would the L1 optimization recovery provide a faithful performance under these practical settings? Would the measurement noise and the model mismatch completely ruin the nice properties of the Compressed Sensing? Compressed Sensing theorists have considered these questions and provided answers, and a collection of results have been obtained to address them. To state the results first, the news is on the positive side.

The L1 recovery algorithm provides stable recovery results.

This is to mean that the model mismatch and the observation errors do not amplify in the process of L1 minimization routines.

The readers who are interested in this subject are referred to Donoho, Elad, Temlyakov [29] and Candes, Romberg, and Tao [30].

6. The Chapter Problems

Problem 1 Show $\mu \geq \frac{1}{\sqrt{M}}$. Assume that the columns of the orthonormal matrices A and B have the unit energy.

Problem 2 Prove the uncertainty principle. Must use the notation of this note, i.e., $x = As_1$ and $x = Bs_2$ throughout the proof. (Hint: see Elad's paper) Provide a succinct (less than 10 statements) sketch of the proof in English. Note that the second part is more important than the first part.

Problem 3 Provide a succinct sketch of the proof for the L0 uniqueness. Less than 10 sentences again.

Problem 4 Do the same for the L1 uniqueness proof.

Problem 5 Do the L1 uniqueness proof using the proof-by-contradiction method.

Problem 6 (Review of Sampling Theorem) Here we would like to review the sampling theorem of the Shannon and Nyquist.

- a. First, show that the train of impulses in the time domain is again the train of impulses in the frequency domain under the Fourier transform:

$$\mathcal{F}\left\{\sum_{n=-\infty}^{\infty} \delta(t - nT_s)\right\} = \frac{1}{T_s} \sum_{k=-\infty}^{\infty} \delta\left(f - k\frac{1}{T_s}\right), \quad (84)$$

where $\mathcal{F}(x) = X(f) := \int_{-\infty}^{\infty} x(t) e^{-2\pi ft} dt$ denotes the Fourier transform coefficients $X(f)$ and $\delta(t)$ is the dirac delta function. (Hint: use the Fourier series).

- b. Second, use MATLAB to see this. Use a time domain vector $x = [1 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0 \ 1 \ 0 \ 0 \ 0]$. Taking `fft(x)` should give $\frac{1}{4}[1000100010001000]$.
- c. Third, find the F.T. of a signal $x(t) = \begin{cases} 1, & t \in [-T_p, T_p] \\ 0, & \text{o.w.} \end{cases}$. Show that the

energy is conserved. Sketch your results.

- d. Fourth, find the inverse F.T. of $X(f) = \begin{cases} 1, & f \in [-W, W] \\ 0, & \text{o.w.} \end{cases}$, where we have defined the inverse F.T. as $\mathcal{F}^{-1}(X(f)) = x(t) := \frac{1}{2\pi} \int_{-\infty}^{\infty} X(f) e^{2\pi ft} dt$. Show that the energy is conserved. Sketch your results.
- e. Fifth, use the developments so far to derive the Shannon's sampling theorem. See the Shannon's 1948 paper. (Following is the exact words captured from the paper)

If a function of time $f(t)$ is limited to the band from 0 to W cycles per second it is completely determined by giving its ordinates at a series of discrete points spaced $\frac{1}{2W}$ seconds apart in the manner indicated by the following result.⁵

Theorem 13: Let $f(t)$ contain no frequencies over W . Then

$$f(t) = \sum_{-\infty}^{\infty} X_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

where

$$X_n = f\left(\frac{n}{2W}\right).$$

Problem 7 (Let us see the effect of prime N)

- a. We can create an example of the impulse trains similar to Problem 6.b, with a prime block length N . (True/False with reasons)
- b. Write the general expressions for the discrete Fourier transform pair. Verify the expression you've got by showing (i) orthogonality, and (ii) Inverse transform of the transform of a signal gets you back the original signal.
- c. Make a 16 by 16 Discrete Fourier Transform matrix. Call this F . What is the RIP constant for this matrix F ? You can obtain it from Monte Carlo simulation (say, from a thousand random trials)
- d. Now take the first 8 rows of the matrix F and form a fat sensing matrix F_1 . Find the RIP constant of this matrix F_1 . Compare it with that of c .
- e. Make a 17 by 17 Discrete Fourier Transform matrix. Do c & d again, and compare the results. Now, what can you say about the effects of prime N ?

HW#3 starts from here.

Problem 8 (Prove Lemma 2.2 of [9]) Use the notation of this note. Otherwise there is no point. Succinctly summarize the proof in English (10 sentences). Note that the latter part is very important.

Problem 9 (The L0 uniqueness proof for prime N , Theorem 1.1 of [10]) Study the proof of the theorem, and (i) discuss the method of proof, among the three methods we discussed in class, (ii) discuss if there is any difference that we have done in the class, (iii) discuss the effect of prime N .

Problem 10 (The L1 uniqueness proof for prime N , [10]) Read Section II of [10] and discuss the similarity and difference of the materials there with the materials covered in class regarding the L1 uniqueness proof.

Problem 11 (Subdifferential and subgradient) Draw the subdifferential of the second function in Figure 14 in Chapter VII.1.

Problem 12 (Lagrange duality)

a. Consider the following problem

$$\begin{aligned} \min \quad & \frac{1}{2} x^T \begin{pmatrix} 2 & 1 \\ 1 & 2 & 1 \\ & 1 & 2 \end{pmatrix} x + \left(\frac{1}{2} \quad \frac{1}{2} \quad 1\right) x + 3 \\ \text{subject to} \quad & \begin{bmatrix} \frac{1}{2} & 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 & \frac{1}{3} \end{bmatrix} x = \begin{bmatrix} 2 \\ 3 \end{bmatrix} \end{aligned}$$

- ◆ Is the primal function convex?
- ◆ Obtain the Lagrange dual function.
- ◆ Find the K.K.K. conditions.
- ◆ Solve the optimization problem using the K.K.K. conditions.
- ◆ What can you say about the strong or weak duality in this case?

b. (Capacity of two parallel channels) Consider the capacity of two parallel AWGN channels given a fixed power budget P . Note that capacity of an AWGN channel is $\frac{1}{2} \log_2(1 + \frac{P_i}{N_i})$ where P_i is the transmit power spent for channel i , $i=1, 2$. We aim to distribute the given power P to the two channels so that the capacity of the two channels can be maximized given the noise powers of the two channels N_1 and N_2 respectively.

- ◆ Set up a primal problem
- ◆ Provide answers to the same set of questions given in a.
- ◆ Solve the problem for $P = 1$, $N_1 = 0.5$ and $N_2 = 1$. Namely determine the power levels and the capacity.

Problem 13 (*Open Problem*) Compare the sufficient condition bound obtained in the lecture with that of Candes and Tao.

Chapter IV. INFORMATION THEORETIC CONSIDERATION

1. K -sparse signal model

We would like to specify the input vector, K -sparse signal \mathbf{x} in the following manner:

- (Index Profile \mathbf{t}) It has K non-zero entries. We put the indices of the non-zero entries in to a vector $\mathbf{t} = (t_1, t_2, \dots, t_K)$ and call it an Index Profile. Each entry $t_k \in \{1, 2, \dots, N\} = \mathcal{N}$ denotes the index of a non-zero entry in \mathbf{x} . We let \mathcal{S}_t be referred to as the set of all possible Index Profiles. We note its size is

$$|\mathcal{S}_t| = \binom{N}{K}. \quad (85)$$

In this note, we assume each profile is equi-probable unless otherwise stated.

- (Value Profile \mathbf{s}) We put the values of the K non-zero entries into a vector $\mathbf{s} = (s_1, s_2, \dots, s_K)$, call it Value Profile. A value profile \mathbf{s} can be determined from a distribution. For example, we may use Gaussian, Bernoulli, or a hybrid distribution. We use a pdf $f_s(\mathbf{s})$ to denote a VP distribution. For the example of complex valued Gaussian multivariate random vector, the pdf is given by

$$f_s(\mathbf{s}) = \frac{1}{\pi^N |\mathbf{C}_s|} \exp \left[-\frac{1}{2} (\mathbf{s} - \bar{\mathbf{s}})^* \mathbf{C}_s^{-1} (\mathbf{s} - \bar{\mathbf{s}}) \right] \quad (86)$$

where $\bar{\mathbf{s}} := E\{\mathbf{s}\}$ is the mean vector of the Gaussian multivariate \mathbf{s} and $\mathbf{C}_s := E\{(\mathbf{s} - \bar{\mathbf{s}})(\mathbf{s} - \bar{\mathbf{s}})^*\}$ is the covariance matrix.

- (A hybrid distribution case) The preceding steps 1 and 2 may be good enough for a K -sparse signal in exact sense—exactly K non-zero elements. The following would be useful when we want it to include all those \mathbf{x} whose support set size is smaller than or equal to K . This is thus just a trivial extension. At any rate, we may use a hybrid distribution to cover such a case. In such a case, the IP set \mathcal{S}_t should include all possible IPs whose size is smaller than or equal to K . Then, the size of the IP set is equal to the number of points in a Hamming sphere of size K , i.e.,

$$|\mathcal{S}_t| = V_2(N, K) = \sum_{k=0}^K \binom{N}{k}. \quad (87)$$

The number of non-zero entries, k , is now a random variable with the following distribution

$$f_K(k) = \frac{\binom{N}{k}}{V_2(N,K)}. \quad (88)$$

Using the two distributions (86) and (88), we may obtain a hybrid distribution .

2. The Entropy of K -sparse signals

We are interested in asking the following questions.

- How much information in terms of bits can be represented by the K -sparse signal \mathbf{x} ? Namely, we ask how large the entropy of the K -sparse signal \mathbf{x} is.
- How much information in terms of bits can be represented by the signal vector $\mathbf{y}_{M \times 1}$? This question can be answered being divided into the two exclusive cases, the first case is when $M \geq 2K$ and the complement case. In this section, we let \mathbf{F} be an $M \times N$ Fourier transform matrix with prime N . This case simplifies our answer to this question. Note that since $\mathbf{y}_{M \times 1} = \mathbf{F}_{M \times N} \mathbf{x}$ is given, as long as the map is one-to-one correspondent, which is the case when $M \geq 2K$, the entropy of \mathbf{x} is the entropy of $\mathbf{y}_{M \times 1}$. It would be a challenging task if the same question can be thrown when the sensing matrix is Gaussian.

Lemma 1. Let \mathbf{F} be an $M \times N$ Fourier transform matrix with prime N where $M \geq 2K$. Let \mathbf{x} be a K sparse signal. Then, the entropy of \mathbf{y} given \mathbf{F} is $H(\mathbf{x})$, i.e., $H(\mathbf{y} | \mathbf{F}) = H(\mathbf{F}\mathbf{x} | \mathbf{F}) = H(\mathbf{x})$. If $M < 2K$, then $H(\mathbf{y} | \mathbf{F}) = H(\mathbf{F}\mathbf{x} | \mathbf{F}) \leq H(\mathbf{x})$.

The entropy of K -sparse signal \mathbf{x} is

$$\begin{aligned} H(\mathbf{x}) &= H(\mathbf{t} = (t_1, \dots, t_K), \mathbf{s} = (s_1, \dots, s_K)) \\ &= H(\mathbf{t}) + H(\mathbf{s} | \mathbf{t}) \\ &= H(\mathbf{t}) + H(\mathbf{s}) \end{aligned} \quad (89)$$

Assuming that the support set of size K is uniform randomly distributed, the entropy of $H(t_1, \dots, t_K)$ can be written as

$$H(t_1, \dots, t_K) = \log_2 \binom{N}{K}. \quad (90)$$

Using the Stirling's approximation for the factorial function, we can show

$$\log_2 \left(\frac{1}{N+1} \right) + NH \left(\frac{K}{N} \right) \leq \log_2 \binom{N}{K} \leq NH \left(\frac{K}{N} \right). \quad (91)$$

For a large N , thus, we note that $\log_2 \binom{N}{K} \cong NH \left(\frac{K}{N} \right)$.

But when K is small compared to N , the entropy function $H \left(\frac{K}{N} \right) = \frac{K}{N} \log_2 \frac{N}{K} + \left(\frac{N-K}{N} \right) \log_2 \left(\frac{N}{N-K} \right)$ can be approximated with the first term only, which is $H \left(\frac{K}{N} \right) \approx \frac{1}{N} K \log_2 \frac{N}{K}$.

This implies that for $K \ll N$

$$NH \left(\frac{K}{N} \right) \approx K \log_2 \left(\frac{N}{K} \right). \quad (92)$$

Remark 1. This result is quite interesting. In the Compressive Sensing literature, the R.H.S. of (92) appears as the number of random projection measurements sufficiently needed for a reliable recovery. It is also same as the number of bits sufficiently required to search for the support set of size K .

Now we note that if $|\Omega| \geq 2|T|$, or $M \geq 2K$, any compression map from \mathbf{x} to $\mathbf{y} = \mathbf{F}\mathbf{x}$ is one-to-one correspondent. Thus, the entropy of \mathbf{y} is also the same as that of the input \mathbf{x} .

Using the result of Theorem 1.3 of Candes-Romberg-Tao [10], we can prove the following Lemma.

Lemma 2. If $|T| \leq \frac{|\Omega|}{2}$, $H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) = 0$; otherwise if $|T| \leq |\Omega| < 2|T|$, $H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \leq H(\mathbf{t})$. If $|\Omega| < |T|$, $H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \leq H(\mathbf{x})$;

Proof: Let us do this for the first case. For the first case, the map $\mathbf{F}_{M \times N}$ is one-to-one correspondent from any K -sparse signal \mathbf{x} to $\mathbf{y}_{M \times 1}$. Thus, we have $H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) = 0$.

Now let's consider the second case:

$$\begin{aligned} H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) &= H(\mathbf{t}, \mathbf{s} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \\ &= H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) + \underbrace{H(\mathbf{s} | \mathbf{t}, \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})}_{=0} \\ &= H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \\ &\leq H(\mathbf{t}) \end{aligned} \tag{93}$$

The second line is due to conditional entropy. The third is due to the fact that the uncertainty on \mathbf{s} is zero once \mathbf{t} and $\mathbf{y}_{M \times 1}$ are given. One may use the LS estimator for the over determined problem formed from those columns, known from \mathbf{t} , of $\mathbf{F}_{M \times N}$. The uncertainty on \mathbf{x} given $\mathbf{y}_{M \times 1}$ is left only on the unknown vector \mathbf{t} .

On the third case, $H(\mathbf{s} | \mathbf{t}, \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \neq 0$.

Thus, we only have a trivial result, i.e., $H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \leq H(\mathbf{x})$.

Q.E.D.

It is interesting to note that in fact $H(\mathbf{s} | \mathbf{t}, \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) = 0$ as long as $|T| \leq |\Omega|$.

Another line of thought can be that using the knowledge that the signal is K -sparse, we can find the vector \mathbf{t} exhaustively. In such a case, we note that, there is about $H(\mathbf{t})$ much of uncertainty left on the input.

3. Mutual Information

Lemma 3: Let $M \geq K$. The mutual information $I(\mathbf{x}; \mathbf{y}_{M \times 1} | \mathbf{F}_{M \times N})$ between \mathbf{x} and $\mathbf{y}_{M \times 1}$ given the measurement matrix is $H(\mathbf{s}) + H(\mathbf{t}) - H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})$.

Proof:

$$\begin{aligned}
 I(\mathbf{x}; \mathbf{y}_{M \times 1} | \mathbf{F}_{M \times N}) &= H(\mathbf{x} | \mathbf{F}_{M \times N}) - H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \\
 &= H(\mathbf{t} | \mathbf{F}_{M \times N}) + H(\mathbf{s} | \mathbf{t}, \mathbf{F}_{M \times N}) - H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \\
 &= H(\mathbf{s} | \mathbf{t}, \mathbf{F}_{M \times N}) + H(\mathbf{t} | \mathbf{F}_{M \times N}) - H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) \\
 &\leq H(\mathbf{s}) + H(\mathbf{t}) - H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})
 \end{aligned} \tag{94}$$

In the second line, we have $H(\mathbf{x} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) = H(\mathbf{s}, \mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})$ which is equal to $H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) + H(\mathbf{s} | \mathbf{y}_{M \times 1}, \mathbf{t}, \mathbf{F}_{M \times N})$; and the second term is zero (no uncertainty on \mathbf{s} given the three). Then, we have $H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})$. Proceeding with the inequality, we use “conditioning reduces the entropy” on $H(\mathbf{s} | \mathbf{t}, \mathbf{F}_{M \times N}) \leq H(\mathbf{s})$. The equality is met if the vector \mathbf{s} is independent from the joint information of \mathbf{t} and $\mathbf{F}_{M \times N}$.
Q.E.D.

Corollary 4. Let $M \geq 2K$. $I(\mathbf{x}; \mathbf{y}_{M \times 1} | \mathbf{F}_{M \times N}) = H(\mathbf{x})$.

This holds because given $\mathbf{y}_{M \times 1}$ and $\mathbf{F}_{M \times N}$, there is no uncertainty on the position vector \mathbf{t} . That is, $H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) = 0$.

Remark 1: Corollary 4, in fact, says that the sparse signals can be measured and perfectly recoverable with “incomplete” measurements, as long as one has M independent “incomplete” measurements. In fact, Corollary 4 shows that the “incomplete” measurement is not indeed incomplete since the “incomplete” measurement $\mathbf{y}_{M \times 1}$ in fact contains all the information that the input vector \mathbf{x} contains within it. This is of course dependent upon the structure of the sensing matrix $\mathbf{F}_{M \times N}$. The meaning of Corollary 4 should be clear. In the compressive sensing literature, the sparse-solution can be found through exhaustive search.

Discussion on lossy vs. lossless compression.

In information theory, we compare entropy of the input to the output. We say a compression is made when a source whose apparent rate is higher than the entropy of the source can be represented at a rate closer to the entropy.

We say that the compression is lossless if the original signal can be recovered with 100% accuracy from the compressed signals. We say that the compression is lossy if the original signal can be recovered only with a certain amount of distortion.

For the lossless compression, the encoding scheme includes the Huffman codes, Lempel-Ziv codes, Arithmetic codes, and Run-Length codes. These codes can be used to encode a source whose encoding rate is close to the entropy rate of the source. As long as the rate is greater

than the entropy rate, we can find a codebook that achieves zero errors.

Shannon has shown that there exists a block code whose coding rate is arbitrarily close to the entropy rate of the source when the block length is sufficiently large. This is achieved when we allow the probability of error to be arbitrarily small but non zero.

When we want the coding rate to be smaller than the entropy $R < H(x)$, then we should look for a rate-distortion code which reproduces the source with a certain amount of distortion.

Corollary 5. The mutual information $I(\mathbf{x}; \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})$ between the K -sparse signal \mathbf{x} and the received signal $\mathbf{y}_{M \times 1}$ and $\mathbf{F}_{M \times N}$ is $I(\mathbf{x}; \mathbf{y}_{M \times 1} | \mathbf{F}_{M \times N})$.

Proof: Using the chain rule, we have $I(\mathbf{x}; \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N}) = I(\mathbf{x}; \mathbf{F}_{M \times N}) + I(\mathbf{x}; \mathbf{y}_{M \times 1} | \mathbf{F}_{M \times N})$. But the first term is zero because the two are mutually independent.
Q.E.D.

At this point, it is of interest to find how big the non-zero term $H(\mathbf{t} | \mathbf{F}_{M \times N}) - H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})$ is.

First, we note that the first term is simply $H(\mathbf{t})$. We cannot reduce uncertainty on the combination of non-zero positions in the unknown vector \mathbf{x} by knowing the measurement matrix $\mathbf{F}_{M \times N}$.

Second, we should ask the question if the uncertainty on the combination $H(\mathbf{t} | \mathbf{y}_{M \times 1}, \mathbf{F}_{M \times N})$ can be reduced by knowing $\mathbf{y}_{M \times 1}$ in addition to $\mathbf{F}_{M \times N}$.

4. Lower Bound on Probability of Sensing Error via Fano's Inequality

Fano's inequality has been used to prove the converse of the channel capacity theorem. That is, a sequence of channel codes has error probability which vanishes as the length of the code grows to infinity must have its code rate smaller than the information theoretic channel capacity. It may lead to a lower bound on the sensing bound. Let us see this possibility in this section.

Let us note that the input is \mathbf{t} , the support set. From the support \mathbf{t} , we generate \mathbf{x} , the signal. Then, we observe the measurement \mathbf{y} . Now, suppose any decision device, and its outcome $\hat{\mathbf{t}}$ which is made from the measurement \mathbf{y} . Thus, we have a Markov chain $\mathbf{t} \rightarrow \mathbf{x} \rightarrow \mathbf{y} \rightarrow \hat{\mathbf{t}}$. Suppose the sensing matrix is known to the decision device.

Theorem 6. (Fano's inequality)

$$H_b(P_e) + P_e \log(|\mathbf{t}|-1) \geq H(\mathbf{t} | \mathbf{y}_{M \times 1}) \quad (95)$$

where $P_e := \Pr(\hat{\mathbf{t}} \neq \mathbf{t})$ means the support set detection error probability, $|\mathbf{t}|$ means the size of the support set, $H(\mathbf{t} | \mathbf{y}_{M \times 1})$ is a conditional entropy, and $H_b(\cdot)$ is the binary entropy function.

The technique for proof is available in Chapter 2 of Cover and Thomas.

Now, let us apply the inequality to lower bound the probability of support set recovery error. From the Fano's inequality, and making use of a bound $H_b(P_e) \leq 1$, the decision error probability can be lower bounded as follows:

$$P_e \geq \frac{H(\mathbf{t} | \mathbf{y}_{M \times 1}) - 1}{\log|\mathbf{t}|-1}. \quad (96)$$

Now suppose that $M \geq 2K$, then the map is one-to-one correspondent from \mathbf{t} to $\mathbf{y}_{M \times 1}$; thus, $H(\mathbf{t} | \mathbf{y}_{M \times 1}) = 0$ which leads to a trivial bound.

This problem is simple because we are currently dealing with the Fourier sensing matrices with prime N , the reason for $H(\mathbf{t} | \mathbf{y}_{M \times 1}) = 0$ for $M \geq 2K$.

At any rate, this lower bound result is consistent with the result summarized in Remark 2 at Chapter III.4.B.

Remark 1. (Fano inequality result is consistent with the upper bound result in Chapter III.4.B) In Chapter III.4.B, see Remark 2 there, we have discussed how the probability that the Vandermonde frame does not satisfy the RIP is exactly zero if $M \geq 2K$. That was because any collections of the columns of the $M \times 2K$ Vandermonde matrix are all linearly independent, the spark is larger than $2K$, thus unique L_0 solution is obtainable for any K -sparse input vector x . The $M \times N$ partial Fourier matrix with prime N has the same property. Thus, as in the case of the Vandermonde frame, the upper bound on the probability of support set recovery error is exactly zero. This is consistent with the lower bound made by the Fano's inequality.

Remark 2. (Results on Gaussian matrices) It would be perhaps more interesting when we consider Gaussian ensembles for sensing matrices. Tang and Nehorai considered the Gaussian ensembles, used the union bound approaches with Chernoff upper bound on pairwise error probabilities, used a derivative of the Fano's inequality to obtain the lower bound on the probability of support set recovery error. They reported that they have obtained a necessary condition showing that the $\log\left(\frac{N}{K}\right)$ term cannot be removed, see the first paragraph on page 1385 in [35].

Chapter V. SPARSE RECOVERY ALGORITHMS

1. Linear Programming

Now, let us consider finding the solution of the L1 minimization problem. Following Chen-Donoho-Saunders [17], we will call it a Basis Pursuit problem:

$$(BP) \quad \min_{x \in \mathbb{R}^N} \|x\|_1 \quad \text{s.t.} \quad y = Ax. \quad (97)$$

where the matrix A is of size $M \times N$. They say that BP is an optimization principle, not an algorithm.

Recasting the BP problem into an LP problem can be done in the following way:

$$\begin{aligned} \min_{(x,u)} \sum_i u_i &= 1^T u + 0^T x = \begin{bmatrix} 0^T & 1^T \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \\ \text{s.t.} \quad x - u &= \begin{bmatrix} e & -e \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \leq 0, \\ -x - u &= -\begin{bmatrix} e & e \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \leq 0, \\ -u &= \begin{bmatrix} 0 & -e \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \leq 0, \\ [A \quad 0] &\begin{pmatrix} x \\ u \end{pmatrix} - b = 0 \end{aligned} \quad (98)$$

where we define e as the identity matrix. Note that both the inequality constraint and the equality constraint functions are affine as well. Note that the third constraint, $u \geq 0$, is redundant given the first two.

Another approach taken by Huo and Donoho is as follows:

$$\begin{aligned} \min_x 1^T u \\ \text{s.t.} \quad y = Bu \end{aligned} \quad (99)$$

where $u := \begin{bmatrix} x_+^T & x_-^T \end{bmatrix}^T$, $x_+ := \max\{x, 0\}$, $x_- := \max\{-x, 0\}$, and $B := [A \ ; \ -A]$.

We solve this by an interior point method.

There are a number of MATLAB program routines that can be utilized to solve this problem as well. They include the ones good for large scale problems. See the MATLAB Help menu, and look for “linear programming.”

We note that (98) takes after the form of the standard linear programming problem:

$$\begin{aligned}
 & \min c_0^T x \\
 (LP) \quad & \text{s.t. } f_i(x) := c_i^T x + d_i \leq 0, \quad i = 1, 2, \dots, m \\
 & Ax - b = 0
 \end{aligned} \tag{100}$$

where $c_i \in \mathbb{R}^N$, $b \in \mathbb{R}^M$, A an $M \times N$ matrix and $d_i \in \mathbb{R}$ are given as constants. Note that each function, including the objective, inequality and equality constraint functions, is affine. Thus, the problem is convex and a linear program.

In the L_1 -magic package, Candes and Romberg have outlined a number of L_1 recovery algorithms. They provide several L_1 recovery MATLAB programs to solve a number of different problems on which they have published papers. They are built based on two generic linear programming methods which have their underpinnings in the Convex Optimization by Boyd and Vendenberghe [12]. We will provide these two generic linear programming methods:

- The first one is based on the Lagrange duality theory. This is the interior point method exposed below.
- The second one is based on the log-barrier method.

2. Solving Linear Program via Lagrange Dual Interior Point Method

We first obtain the Karush-Kuhn-Tucker conditions for the linear program. We then aim to find the optimal point (x, λ, v) that satisfies the K.K.T. conditions. We will take the Lagrange dual approach to find the optimal point. We anticipate that the strong duality holds since the problem is convex. To search for the optimal point, an interior point method will be used.

Please note that the L1 magic package by Candes and Romberg has a routine that does implement this algorithm. We also provide the MATLAB code and the manual of our own (with corrections of typos in the L1 magic package) in HW set #4 solution manual of this book.

At the optimal point x^* , there exist Lagrange multiplier vectors $\lambda^* \geq 0$, $\lambda \in \mathbb{R}^m$ and $v^* \in \mathbb{R}^M$.

The K.K.T. conditions are

$$\begin{aligned} c_0 + \sum_{i=0}^m \lambda_i^* c_i + A^T v^* &= 0 \\ \lambda_i^* f_i(x^*) &= 0, \quad i = 1, \dots, m \\ Ax^* - b &= 0, \\ f_i(x^*) &\leq 0, \quad i = 1, \dots, m \end{aligned} \tag{101}$$

This cannot be solved as a linear equation since it has the inequality constraints and the second set of equations is not linear (it's a product of variables). Thus, we first like to use an interior condition to get rid of the inequality condition. We will use the Newton method such that in a number of iterations, we would like to have the solution of the Newton's method approach to the optimal points x^* , $\lambda^* \geq 0$, and v^* . Let us denote k as the iteration index, and λ^k , x^k and v^k as the primal and dual variables at the k -th iteration.

That is, while trying to approach to the optimal solution, we would like to enforce

- $\lambda_i > 0$, $i = 1, \dots, m$ and
- $f_i(x^*) < 0$, $i = 1, \dots, m$

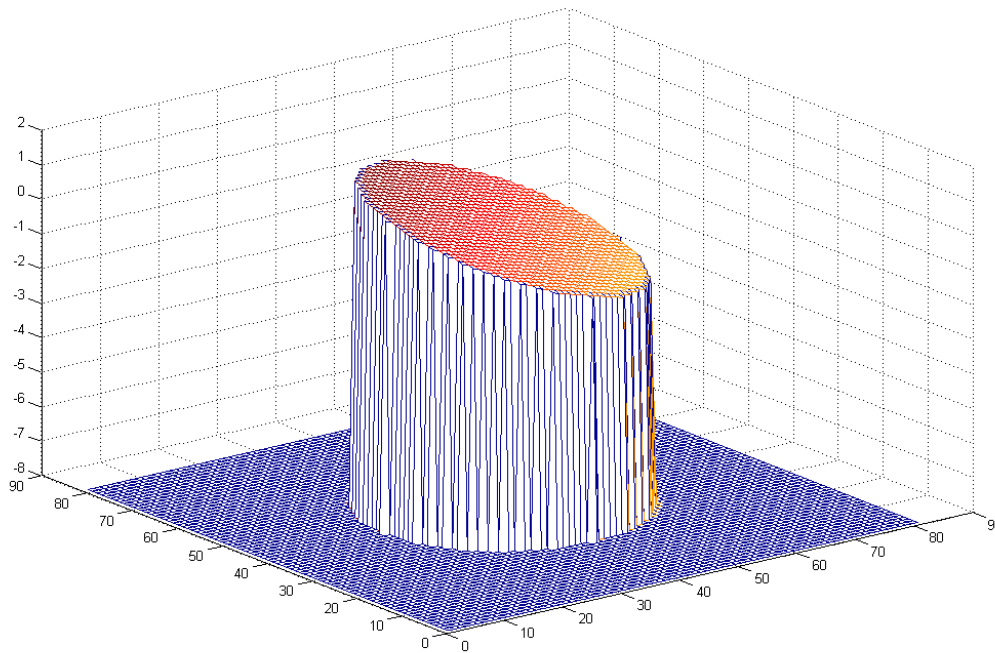


Figure 4: An example of linear objective function $\min_x [1 \ -1]x$ s.t. $x^T x - 1 \leq 0$. Note that the optimal point is achieved at the perimeter of the feasible set.

The objective of approaching the solution from the interior can be achieved by employing an ever getting smaller slack variable $\frac{1}{\tau^k}$ for $\tau > 0$ which shrinks as the iteration proceeds, i.e.,

$$\lambda_i^k f_i(x^k) + \frac{1}{\tau^k} = 0, \quad i = 1, \dots, m \quad (102)$$

Note that as the iteration index proceeds the slack variable tends to zero, and thus each equation approaches $\lambda_i^* f_i(x^*) = 0$, $i = 1, \dots, m$. We will have to choose λ^k , x^k carefully at each iteration so that they remain within the interior of the primal and dual feasible sets. We want the solution at the boundary to be approached from the interior.

Note that (102) replaces both $\lambda_i^* f_i(x^*) = 0$, $i = 1, \dots, m$ and $f_i(x^*) \leq 0$, $i = 1, \dots, m$ in (101).

Now we have the following three equations

$$\begin{aligned}
r_g(x, \lambda, v) &:= c_0 + \sum_{i=0}^m \lambda_i c_i + A^T v \\
r_s(x, \lambda, v) &:= - \underbrace{\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix}}_{\Lambda} \underbrace{\begin{pmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{pmatrix}}_f - \frac{1}{\tau} \mathbf{1}, \quad i=1, \dots, m \\
r_p(x, \lambda, v) &:= Ax - b
\end{aligned} \tag{103}$$

We combine all three as a single residual function $r(x, \lambda, v) = \begin{pmatrix} r_g \\ r_s \\ r_p \end{pmatrix}$. Refreshing our objective, we aim to find optimal point (x, λ, v) such that $r(x, \lambda, v) = 0$. Since the equation is non linear, we use the Newton's method to find the roots of the equation.

(Newton's method) We use the first order approximation (linear approximation) to $r(x, \lambda, v)$. That is, at a particular point (x, λ, v) and the residual $r(x, \lambda, v)$ at that point, we find the Jacobian. Linear approximation of the residual $r(x, \lambda, v)$ using the Taylor expansion around the point (x, λ, v) gives

$$r(x + \Delta x, \lambda + \Delta \lambda, v + \Delta v) \approx r(x, \lambda, v) + J(x, \lambda, v) \begin{pmatrix} \Delta x \\ \Delta \lambda \\ \Delta v \end{pmatrix} \tag{104}$$

We aim to find the step so that the residual at the next point is zero, i.e., $r(x + \Delta x, \lambda + \Delta \lambda, v + \Delta v) = 0$. Then, we have

$$J(x, \lambda, v) \begin{pmatrix} \Delta x \\ \Delta \lambda \\ \Delta v \end{pmatrix} = -r(x, \lambda, v) \tag{105}$$

Rewriting it with the details, we have

$$\begin{pmatrix} 0 & C^T & A^T \\ -\Lambda C & -F & 0 \\ A & 0 & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta \lambda \\ \Delta v \end{pmatrix} = - \begin{pmatrix} c_0 + \sum_{i=0}^m \lambda_i c_i + A^T v \\ -\Lambda f - \frac{1}{\tau} \mathbf{1} \\ Ax - b \end{pmatrix} \tag{106}$$

where C is an $m \times N$ matrix with c_i^T as its rows, $i=1, \dots, m$, and F is an $m \times m$ diagonal with $F_{ii} = f_i(x)$, $i=1, \dots, m$.

We note the second equation can be rewritten as

$$\Delta\lambda = -F^{-1}\Lambda C\Delta x - F^{-1}\Lambda f - F^{-1}\frac{1}{\tau}\mathbf{1}_m \quad (107)$$

And the other two equations can be written as

$$\begin{aligned} C^T\Delta\lambda + A^T\Delta v &= -\left(c_0 + \sum_{i=0}^m \lambda_i c_i + A^T v\right) \\ A\Delta x &= -(Ax - b) \end{aligned} \quad (108)$$

Substituting (107) into the first equation of (108),

$$\begin{aligned} C^T\left(-F^{-1}\Lambda C\Delta x - F^{-1}\Lambda f - F^{-1}\frac{1}{\tau}\mathbf{1}_m\right) + A^T\Delta v &= -\left(c_0 + \sum_{i=0}^m \lambda_i c_i + A^T v\right) \\ -C^T F^{-1}\Lambda C\Delta x - C^T F^{-1}\Lambda f - C^T F^{-1}\frac{1}{\tau}\mathbf{1}_m + A^T\Delta v &= -\left(c_0 + \sum_{i=0}^m \lambda_i c_i + A^T v\right) \end{aligned}$$

and finally have

$$-C^T F^{-1}\Lambda C\Delta x + A^T\Delta v = -\left(c_0 + \sum_{i=0}^m \lambda_i c_i + A^T v\right) + C^T F^{-1}\Lambda f + C^T F^{-1}\frac{1}{\tau}\mathbf{1}_m. \quad (109)$$

Note that

$$\begin{aligned} C^T F^{-1}\Lambda f &= C^T \Lambda F^{-1} f = C^T \Lambda \mathbf{1}_m \\ &= \begin{pmatrix} - & c_1^T & - \\ & \vdots & \\ - & c_m^T & - \end{pmatrix}^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix} \mathbf{1}_m \\ &= \begin{pmatrix} \lambda_1 c_1 & \cdots & \lambda_m c_m \end{pmatrix} \mathbf{1}_m \\ &= \sum_{i=1}^m \lambda_i c_i \end{aligned}$$

and $C^T F^{-1}\frac{1}{\tau}\mathbf{1}_m = \frac{1}{\tau}C^T f^{-1}$ where $f^{-1} := \left(\frac{1}{f_1(x)} \cdots \frac{1}{f_m(x)}\right)^T$.

Then, the rest can be written as

$$\begin{pmatrix} -C^T F^{-1}\Lambda C & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta v \end{pmatrix} = \begin{pmatrix} -c_0 + \frac{1}{\tau}C^T f^{-1} - A^T v \\ b - Ax \end{pmatrix} \quad (110)$$

Thus, we can solve (110) to obtain $(\Delta x, \Delta v)$ first, and then we get $\Delta\lambda$ from (107).

The step $(\Delta x, \Delta \lambda, \Delta v)$ obtained from (110) and (107) provides a direction to move in the multi-dimensional space.

Next, we determine how big a move is needed in the found direction. This step size, denoted as $s, 0 < s \leq 1$, needs to be carefully chosen at each move to ensure that the next step must remain as an interior point. Namely, it would be nice if one can choose the largest possible step size while $x + s \Delta x$ and $\lambda + s \Delta \lambda$ stays interior to the feasible sets, for all $i = 1, \dots, m$, i.e.,

$$f_i(x + s \Delta x) < 0 \quad \text{and} \quad \lambda + s \Delta \lambda > 0.$$

3. Solving Second Order Cone Programs with Log-Barrier Method

This section aims to describe the log barrier method which is useful to solve second order cone programs (SOCPs). SOCPs arise in Compressed Sensing problems which include measurement noise. Namely, the following problem belongs to the category of SOCPs:

$$(P1) \quad x^* := \arg \min \|x\|_1 \quad \text{subject to} \quad \|Ax - b\|_2 \leq \varepsilon, \quad (111)$$

where $b = Ax_0 + e$, x_0 is the sparse signal, and e is used to denote the noise term. There is a stability result that x^* is very close to x_0 . More precisely, if x_0 is sufficiently sparse and the error is bounded, $\|e\|_2 \leq \varepsilon$, then, $\|x^* - x_0\|_1 \leq c \cdot \varepsilon$ for some constant c .

The algorithm and a manual are provided in the L1 magic package by Candes and Romberg. We also provide the MATLAB code and the manual of our own (with corrections of some typos noted in the L1 magic package) in HW set #4 solution manual of this book.

Note that the main problem (P1) in this section can be recasted as a problem of a second-order cone problem (SOCP) such that

$$x^* := \arg \min_{(x,u)} \sum_i u_i$$

$$\text{subject to} \begin{cases} x_i - u_i \leq 0, & \text{for each } i \\ -x_i - u_i \leq 0, & \text{for each } i \\ \frac{1}{2} (\|Ax - b\|_2^2 - \varepsilon^2) \leq 0 \end{cases} \quad (112)$$

The log-barrier method, as the name indicates, for is characterized by placing a barrier function to the objective function. This barrier is created by adding a barrier function, or a penalty function, to the objective function. The job of the barrier is to make an optimization procedure, an iterative algorithm implemented with a Newton's method, to remain inside the feasible set while searching for the optimal solution. The new objective function formed with the addition of the barrier to the original objective function should remain close to the original function inside the feasible set, and approach infinity at the boundary of the feasible set. Once the search is started from inside the feasible set, then, the search will remain inside the feasible set.

The following figure should be helpful.

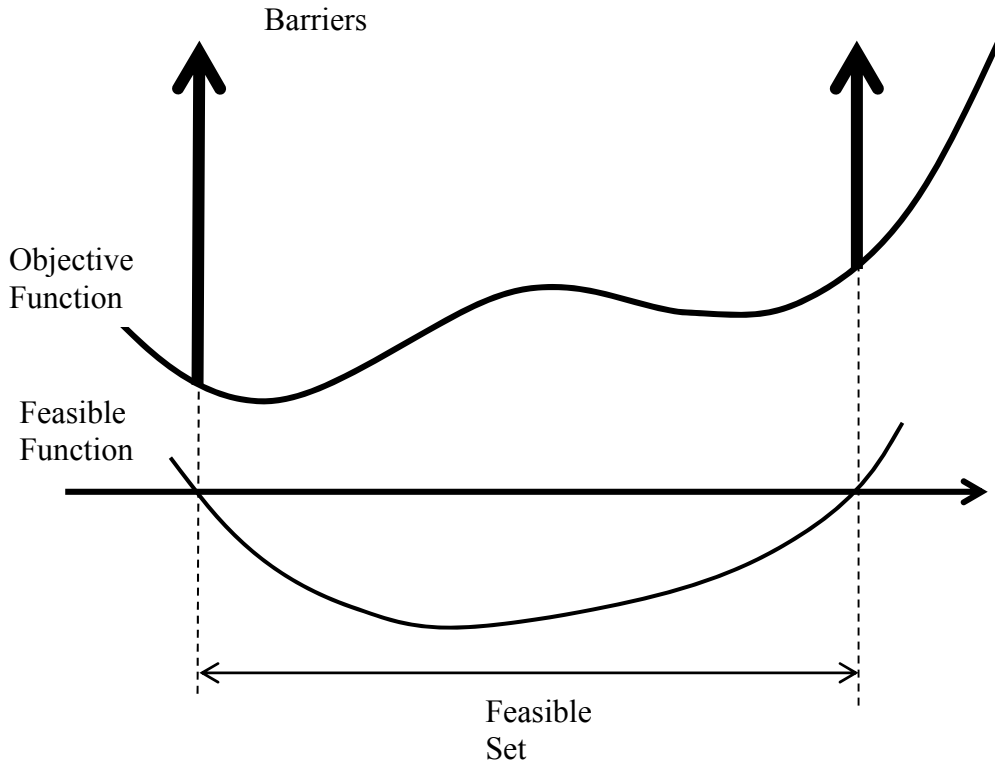


Figure 5: The Concept of Placing Barriers At the Boundary of the Feasible Set

A. A log barrier function

Now, it should be of interest to find a suitable barrier function. Note that our selection should take into account the Newton's method. We consider the log function only in this section. The log function is continuous and differentiable, and thus should be suitable for Newton's method.

To explain the choice of the barrier function, let us consider the following optimization problem:

$$\min_z f_o(z), \quad \text{subject to} \begin{cases} f_i(z) \leq 0 & \text{for some } i \in I \\ f_e(z) = 0 \end{cases} \quad (113)$$

where $f_i(z)$ for some $i \in I$ are the inequality constraint functions and $f_e(z)$ is the equality constraint. Using the log barrier method, we cast the problem in the following form:

$$\min_z f_o(z) + \frac{1}{\tau_k} \sum_{i=1}^m -\log(-f_i(z)), \quad \text{subject to } f_e(z) = 0. \quad (114)$$

Note here that we have moved up all the inequality constraints to the objective part, except the equality constraint. We negate each inequality constraint function $f_i(z)$, i.e., $-f_i(z)$.

Then, inside the feasible set, $-f_i(z)$ is positive and when the argument z approaches the boundary of feasible set, $-f_i(z)$ is getting close to zero. Taking the log, i.e., $\log(-f_i(z))$, then, the result gets smaller in magnitude inside the feasible set but it approaches negative infinity near the boundary. We then negate the result again to form the barrier function, i.e., $-\log(-f_i(z))$. At this point, the barrier function is close to positive infinite near the boundary of the feasible set. In order to subdue the additive influence of the log-barrier function, we scale the log barrier function by $\frac{1}{\tau_k}$, i.e., $-\frac{1}{\tau_k}\log(-f_i(z))$. We increase the value of the denominator τ^k in each iteration where k is the iteration index. Then, each barrier function will behave like a big wall near the boundary of the feasible set.

The duality gap: It can be shown that $|f_o(z^*) - f_o(z)| \leq \frac{m}{\tau_k}$ where m is used to denote the number of inequality constraints in (113). The measure on the R.H.S. $\frac{m}{\tau_k}$ can thus be used as the duality gap. This shows that the solution z^* is within a distance of $\frac{m}{\tau_k}$ from the optimal solution.

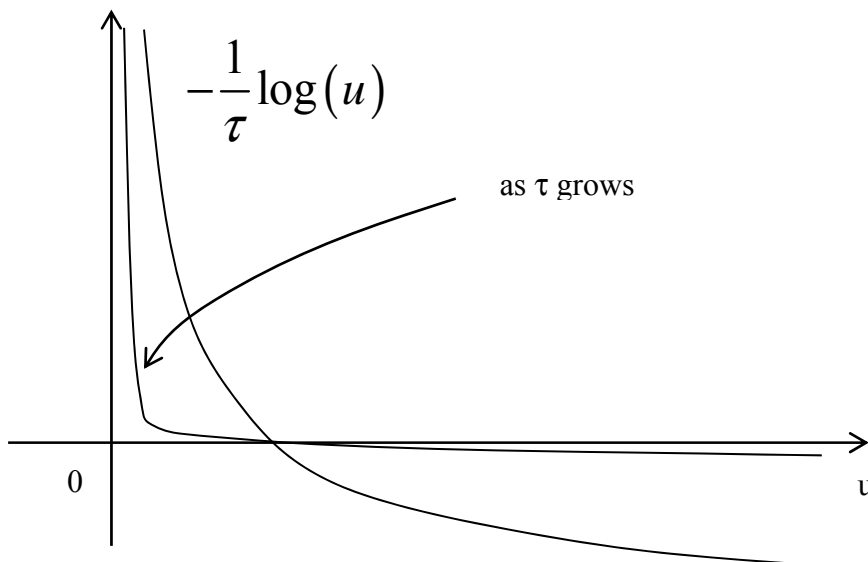


Figure 6: A log-barrier function

B. The interior log barrier method for solving SOCPs

We now aim to apply the log-barrier method to the problem given in (112). To derive an algorithm which is good for a general set of problems, we also assume there is an equality constraint of the form, $A_1x = b_1$. This way, we can handle recovery problem with equality constraint as well. Then, we have

$$\min_{(x,u)} \underbrace{\sum_i u_i + \frac{1}{\tau^k} \left[\sum_i -\log(-x_i + u_i) + \sum_i -\log(x_i + u_i) - \log\left(-\frac{1}{2}(\|Ax - b\|_2^2 - \varepsilon^2)\right) \right]}_{=: f(z)}$$

$$s.t. \quad A_1 x - b_1 = 0$$

(115)

where we have defined the variable $z := (x; u)$.

The quadratic approximation of the objective function is

$$f(z + \Delta z) \approx f(z) + g_z^T \Delta z + \frac{1}{2} \Delta z^T H_z \Delta z := q(z + \Delta z)$$

The gradient g_z and the Hessian H_z are what we need to find for this approximation. They are

$$g_z = \begin{bmatrix} \frac{1}{\tau^k} \left(-\frac{1}{f_{u_1}} + \frac{1}{f_{u_2}} - \frac{2A^T(Ax-b)}{((Ax-b)^T(Ax-b) - \varepsilon^2)} \right) \\ 1^T + \frac{1}{\tau^k} \left(\frac{1}{f_{u_1}} + \frac{1}{f_{u_2}} \right) \end{bmatrix} \in \mathfrak{R}^{2N}$$

$$H_z = \begin{bmatrix} \frac{1}{\tau^k} \left(\frac{1}{f_{u_1}^2} + \frac{1}{f_{u_2}^2} + \frac{A^T(Ax-b)(Ax-b)^T A - A^T A f_\varepsilon}{f_\varepsilon^2} \right) & \frac{1}{\tau^k} \left(-\frac{1}{f_{u_1}^2} + \frac{1}{f_{u_2}^2} \right) \\ \frac{1}{\tau^k} \left(-\frac{1}{f_{u_1}^2} + \frac{1}{f_{u_2}^2} \right) & \frac{1}{\tau^k} \left(\frac{1}{f_{u_1}^2} + \frac{1}{f_{u_2}^2} \right) \end{bmatrix} \in \mathfrak{R}^{2N \times 2N}$$

where $f_{u_1} = x - u$, $f_{u_2} = -x - u$, and $f_\varepsilon = \frac{1}{2}(\|Ax - b\|_2^2 - \varepsilon^2)$. In addition, we have used

$$\frac{\partial}{\partial z} \left(\frac{2A^T(Ax-b)}{(Ax-b)^T(Ax-b) - \varepsilon^2} \right) = \frac{A^T(Ax-b)(Ax-b)^T A - A^T A f_\varepsilon}{f_\varepsilon^2}.$$

a. Construct the matrix vector equation with the unknown $(\Delta z, \nu)$

Now, given that x is feasible, i.e., $A_1 x = b_1$, we aim to find the step Δz which is the solution of the following optimization problem:

$$\min_{\Delta z} q(\Delta z) \quad s.t. \quad [A_1 \quad 0] \Delta z = 0$$

The Lagrangian is

$$L(\Delta z, \nu) = q(\Delta z) + ([A_1 \quad 0] \Delta z)^T \nu$$

From the KKT conditions, i.e., setting the gradient of the Lagrangian function equals to zero, we can obtain

$$\frac{d\left(q(\Delta z)+\left([A_1 \ 0]\Delta z\right)^T v\right)}{d\Delta z}=g_z+H_z\Delta z+\left[A_1 \ 0\right]^T v$$

$$\frac{d\left(q(z+\Delta z)+\left([A_1 \ 0]\Delta z\right)^T v\right)}{dv}=\left[A_1 \ 0\right]\Delta z$$

Finally, we get

$$\begin{bmatrix} H_z & [A_1 \ 0]^T \\ [A_1 \ 0] & 0 \end{bmatrix} \begin{bmatrix} \Delta z \\ v \end{bmatrix} = \begin{bmatrix} -g_z \\ 0 \end{bmatrix}$$

We solve this equation and obtain the step direction Δz . From this we can obtain the next solution z^{k+1}

$$z^{k+1}=z^k+q_{step}\Delta x.$$

The step size q_{step} is obtained heuristically, which is explained below a little bit. Iteration will continue until the duality gap is smaller than a certain prescribed value, i.e., $\frac{m}{\tau_k} \leq \eta$. Here $m=2N+1$, the number of all inequality constraints.

Namely, the algorithm should go as follows:

1. Input: Set a feasible value z^1 , a tolerance η , a parameter μ , the barrier constant τ , and $k=1$.
2. Compute the gradient $g_z(z^k)=:g_z^k$ and the Hessian $H_z(z^k)=:H_z^k$ at z^k .
3. Solve $\begin{bmatrix} H_z & [A_1 \ 0]^T \\ [A_1 \ 0] & 0 \end{bmatrix} \begin{bmatrix} \Delta z \\ v \end{bmatrix} = \begin{bmatrix} -g_z \\ 0 \end{bmatrix}$ and obtain Δz .
4. Obtain $z^{k+1}=z^k+q_{step}\Delta x$ while q_{step} is selected semi-heuristically.
5. Check the stopping criterion. If yes, stop and return z^k , else set $\tau^{k+1}=\mu\tau^k$, $k=k+1$, and go to Step 2.

- Determination of initial guess

- We can use the same initial values for $x^{k=0}, u^{k=0}$ as in the case of L_1 norm with equality constraint such that
- $x^{k=0}=A_1^T b$
- $u^{k=0}=0.95|x^{k=0}|+0.1 \times \max\left[|x^{k=0}|\right]$

- $v^{k=0}$: From the KKT condition, we know that
$$\begin{bmatrix} H_z & [A_1 \ 0]^T \\ [A_1 \ 0] & 0 \end{bmatrix} \begin{bmatrix} \Delta z \\ v \end{bmatrix} = \begin{bmatrix} -g_z \\ 0 \end{bmatrix}$$
 such that $v^{k=0} = -A_1 g_z^{k=0}$

- Step size selection

- From the KKT condition, we obtain Δz^k , which is the Newton step direction. Then, the step size q is chosen so that it is the maximum step size satisfying all the following requirements

- 1) All feasibility constraints are satisfied, i.e.,

$$f_{u_1,i}(x_i + q_{step} \Delta x_i, u_i + q_{step} \Delta u_i) < 0 \quad \text{for } i = 1, \dots, N$$

$$f_{u_2,i}(x_i + q_{step} \Delta x_i, u_i + q_{step} \Delta u_i) < 0 \quad \text{for } i = 1, \dots, N$$

$$f_\varepsilon(x_i + q_{step} \Delta x_i) < 0$$

- 2) The function had decreased sufficiently

$$f(z^{k+1}) < f(z^k) + \alpha q_{step} g_z^k \Delta z$$

- The requirement basically states that the decrease must be within a certain percentage of that predicted by the linear model.
- Usually, we set $\alpha = 0.01$.
- If the conditions 1) and 2) are not satisfied, we reduce the step-size $q_{step} = \beta q_{step}$ and check it again, where $\beta = 0.5$ is chosen in Candes and Romberg's package.

4. Homotopy Algorithms

A. Motivation

In this section, we aim to study the paper by Donoho and Tsaig [18] where the so-called Homotopy algorithm was proposed to solve the Basis Pursuit problem given in (97), the L1 minimization solution, i.e., $\min \|x\|_1$ subject to $y = Ax$. The motivation for studying this algorithm includes:

- It is a fast algorithm perhaps suitable for large scale L1 minimization problems. As Donoho and Tsaig comments (see the caption in Fig. 1 of [18] and the relevant discussion in the body text) “Homotopy *probably* solves L1 minimization problem.”
- It has the K iterative-step solution property. The K -sparse solution can be obtained in K iterative steps when sufficient sparsity is present. This is interesting.
- It is related to other algorithms such as Least Angle Regression (LARs), LASSO, Orthogonal Matching Pursuit (OMP). OMP is probably the fastest; but it is superior to the BP. There are cases where the OMP cannot find the sparse solution while L1 can.
- The interior point methods, and many other conventional L1 norm minimization routines, start at a dense solution and approach the sparse solution as iteration goes. Meanwhile, the Homotopy and the OMP build a sparse solution through the iterative steps by including or removing a sparse set of elements. If the solution is sparse, it is easier to accept now, the Homotopy and the OMP must be faster and simpler than the interior point methods. The complexity of Homotopy is $O(K^3+KMN)$.

The Homotopy method was originally proposed by Osborne, Presnell, and Turlach [21][22] for solving noisy over-determined L1 penalized least squares problem. Donoho and Tsaig [18] used it to obtain the sparse solution of underdetermined problem using the L1 norm minimization, i.e., the Basis Pursuit problem given in (97).

B. The Homotopy Problem

In the Homotopy problem, one aims to solve the following unconstrained L1 penalized least squares problem:

$$\text{(Homotopy)} \quad \min_x \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1 \quad (116)$$

The lambda is a utility variable, i.e., $\lambda \in [0, \infty)$. Given a fixed λ , we note, there is a corresponding solution, say x_λ . If the utility variable is varied, the solution traces a path.

There exists a set of solution paths $\{x_\lambda : \lambda \in [0, \infty)\}$. The desired L1 solution, we hope for, shall be approached as we change the utility variable. In this case, it is changed from a large value to smaller ones. See the illustration below.

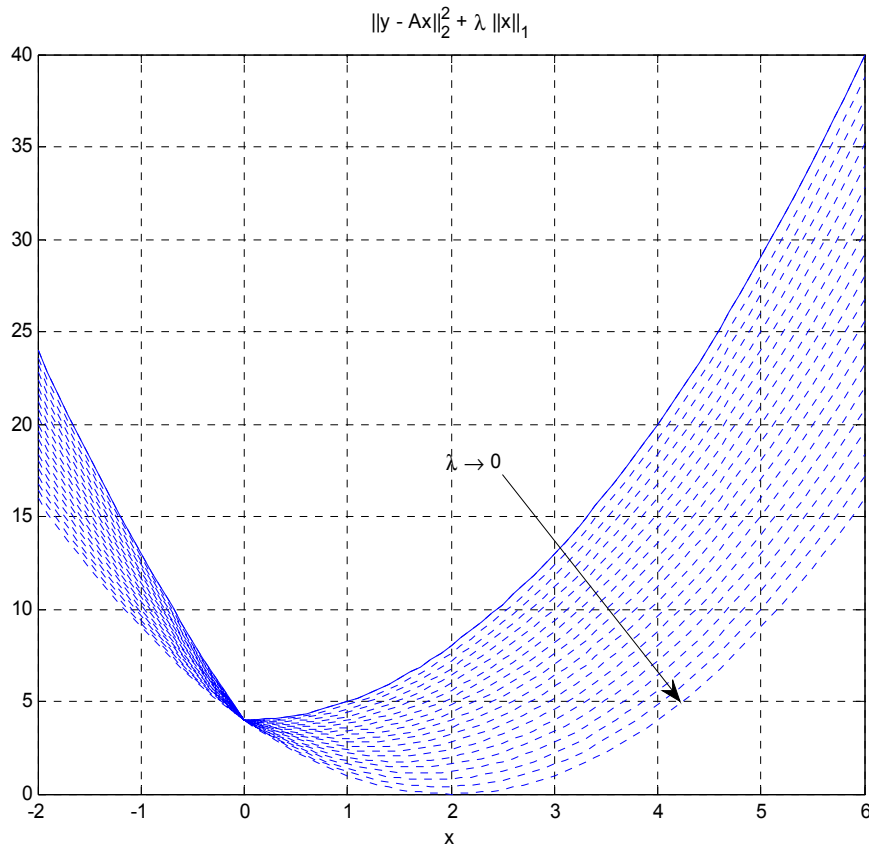


Figure 7: Variation of the penalized objective function as lambda is changing from 4 to 0.

Why? Why not from a smaller value to larger values of lambda?

We note that a large λ can be used to put more emphasis on the L1 norm while for a smaller λ the more emphasis on the L2 norm of the difference. Thus, for a large λ the solution x_λ shall be close to the zero vector. Smallness of x dominates the minimization result whether or not such an x is a feasible solution. As $\lambda \rightarrow 0$, we further note, the emphasis is put more on the feasibility of the solution. At this point, it shall be nice to recall the nature of finding the solution of the L1 norm minimization. We make the L1 ball to grow until it hits the feasibility space for the first time. The first feasible solution at which the L1 ball touches the feasibility space is the L1 norm minimized solution. Thus, it now makes a little more sense why we

shrink the value of lambda started off from a large value. Thus, as λ shrink from a large value to a smaller one, x_λ converges to the L1 norm minimization solution, the solution of the Basis Pursuit problem (97). It looks like starting from a solution closely satisfying $y = Ax$.

LASSO: One can consider the following optimization problem, the so-called LASSO problem, that Tibshirani [23] has proposed:

$$\text{(LASSO)} \quad \min_x \|y - Ax\|_2^2 \quad \text{s.t.} \quad \|x\|_1 \leq q \quad (117)$$

where $q \in [0, \infty)$ and x_q a solution in (117). The set of solutions $\{x_q : q \in [0, \infty)\}$ identifies a solution path $x_q = 0$ as the utility variable q is increased from $q = 0$. As $q \rightarrow \infty$, x_q converges to the solution of the Basis Pursuit problem (97). It seems that the two problems are equivalent, LASSO and Homotopy, as there exists a reparametrization $q(\lambda)$ defined by $q(\lambda) = \|x_\lambda\|_1$ so that the solution path of the Homotopy(λ) and that of LASSO($q(\lambda)$) coincide.

Osborne *et. al* [21] found that the solution path follows a polygonal path. Based on these observation, the Homotopy algorithm is developed, which follows the solution path by jumping from vertex to vertex of this polygonal path. Perhaps, the meaning of this will become clearer after we have introduced the algorithm. For the moment, we can define the meaning of the active set. The active set is defined as the index set of non-zero elements of x at a particular iterative step of the algorithm. It starts at $x_\lambda = 0$ for a large λ with an empty active set. At each vertex, the active set is updated through the addition or removal of “active” elements. An active element here means the element which is believed to be an non-zero element of the sparse vector x using the clues gathered up to that particular step in the course of the algorithm. In a sequence of steps, the algorithm’s active set is guided to become the support set of x as λ is made to go to zero in a controlled manner.

The following figure, Figure 8, taken from Donoho and Tsaig (Fig. 1 of [18]) illustrates the position of Homotopy in the map of relations from L1 minimization, to Least Angle

Regression (LARs), and to the OMP algorithms:

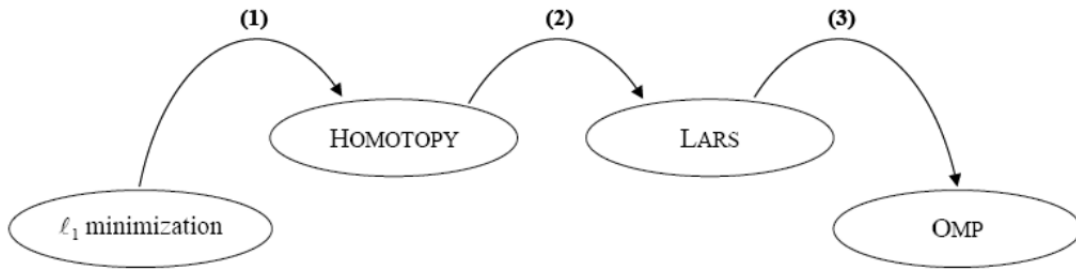


Figure 8: Relation map of algorithms. From Donoho and Tsaig [18].

- (1) Homotopy probably solves L1 minimization problems.
- (2) LARS is obtained from Homotopy by removing the sign constraint check. (Not removing an element from the active set.)
- (3) OMP and LARS are similar in structure, OMP solves a least squares problem at each iteration, whereas LARS solves a linearly penalized least-squares problem.

Theorem (The K -step Property of Homotopy). Let $\mu := \max_{i \neq j} \{|\langle a_i, a_j \rangle|\}$ where a_i s are the columns of the sensing matrix A . Let the sparsity of the vector x satisfy the following inequality

$$K \leq (\mu^{-1} + 1)/2. \quad (118)$$

Then, the Homotopy algorithm runs K steps and stops, delivering the correct solution.

Remarks: It is worthwhile to note that the sufficient condition (118) is identical to the sufficient condition that the solution of the BP problem is identical to the unique L0 norm minimization solution. In addition, Tropp [19][20] found that when (118) holds, even the OMP algorithm recovers the unique L0 solution. The OMP is a very simple greedy algorithm, which will be discussed shortly.

In fact, when the degree of restrictiveness of the sufficient condition (118) is considered, namely it is so restrictive that even the simplest OMP algorithm finds the correct unique solution. The theorem of Homotopy algorithm is not at all surprising because it is more advanced and complex than the OMP algorithm.

C. Two constraints from the subdifferential

Let $f_\lambda(\mathbf{x})$ denote the objective function of the Homotopy problem, i.e.,

$$f_\lambda(x) := \frac{1}{2} \|y - Ax\|_2^2 + \lambda \|x\|_1. \quad (119)$$

We will solve this problem and obtain x_λ as λ is started at a large value and varied while converging to zero. At each time λ is varied, the selection should be carefully made. The selection criterion is derived from the classical convex analysis.

From the classical convex analysis, a necessary condition for x_λ to be a minimizer of $f_\lambda(x)$ is that the zero vector is an element of the subdifferential of f_λ at x_λ , i.e., $0 \in \partial_{x_\lambda} f_\lambda(x_\lambda)$, where

$$\partial_{x_\lambda} f_\lambda(x_\lambda) = -A^T(y - Ax_\lambda) + \lambda \partial_{x_\lambda} \|x_\lambda\|_1 \quad (120)$$

and

$$\partial_{x_\lambda} \|x_\lambda\|_1 = \left\{ u \in \mathbb{R}^N \mid \begin{array}{l} u_i = \text{sgn}(x_{\lambda,i}), \quad x_{\lambda}(i) \neq 0 \\ u_i \in [-1, 1], \quad x_{\lambda}(i) = 0 \end{array} \right\}.$$

Let $\mathcal{K} = \{i : x_{\lambda}(i) \neq 0\}$ denote the support set of x_λ . Let us refer to $c = A^T(y - Ax_\lambda)$ as the vector of **residual correlations**.

Then, the condition $0 \in \partial_{x_\lambda} f_\lambda(x_\lambda)$ can be written equivalently as two conditions. The first one is the **sign agreement**,

$$c(i) = \lambda \cdot \text{sgn}(x_{\lambda}(i)), \quad \forall i \in \mathcal{K}. \quad (121)$$

In words, on the support of \mathcal{K} , the residual correlation must all have magnitude equal to λ , and the sign of the corresponding element of x_λ should match the sign of the residual correlation.

The second is the **upper-bound on residual correlation**,

$$|c(i)| \leq \lambda, \quad \forall i \in \{0, 1, \dots, N-1\} \setminus \mathcal{K} \quad (122)$$

It says that off the support of \mathcal{K} the residual correlation must have magnitude less than or equal to λ .

The Homotopy algorithm carefully traces out the optimal path x_λ that maintains the conditions (121) and (122) for all $\lambda \geq 0$. The key to its successful operation, as Donoho and Tsaig puts it, is that the path x_λ is a piecewise linear path, along the vertices of a polygon.

D. The Homotopy Algorithm

Homotopy is an iterative algorithm. Similar to the Newton's method, seen in the previous

sections, it finds the direction and the step-size. Then, the update of solution is made by moving the current solution estimate by the amount of step-size in the found direction. One notable aspect is that throughout its operation, the algorithm aims to maintain the active set I

$$I = \{j : |c_l(j)| = \|c_l\|_\infty = \lambda\} \quad (123)$$

which satisfies the conditions (121) and (122). As the iteration goes, the active set shall converge to the sparse support set \mathcal{K} . We would like to have the residual correlation be smaller each time.

The Homotopy Algorithm

1. **Initialize:** Given A and y ($= Ax$), set $x_0 = 0, I = \emptyset, c_0 = A^T y, d_0 = 0, A_I = \emptyset$, and $\lambda = \|c_0\|_\infty$. Let the iteration index be $l = 0, 1, 2, \dots$

2. **Update direction:**

Obtain an updated direction vector d_l for the active set by solving

$$A_I^T A_I d_l(I) = \text{sgn}(c_l(I)) \quad (124)$$

Let $d_l(I^c) = 0$.

(This direction update d_l ensures that all the magnitudes of residual correlations on the active set decline equally.)

3. **Determine the step-size s_l :**

There are two cases of constraint violation (for the two constraints).

Case 1: A nonactive element of c_l would increase in magnitude beyond λ , violating the upper bound in (122). This first occurs when

$$s_l^+ = \min_{i \in I^c} \left\{ \frac{\lambda - c_l(i)}{1 - a_i^T A_I d_l(I)}, \frac{\lambda + c_l(i)}{1 + a_i^T A_I d_l(I)} \right\} \quad (125)$$

where the minimum is taken only over positive arguments.

Call the minimizing index i^+ .

Case 2: An active coordinate crosses zero, violating the sign agreement in (121). This first occurs when

$$s_l^- = \min_{i \in I} \left\{ \frac{-x_l(i)}{d_l(i)} \right\} \quad (126)$$

where the minimum is taken only over positive arguments.

Call the minimizing index i^- .

The step-size is determined as the minimum of the above two:

$$s_l = \min\{s_l^+, s_l^-\} \quad (127)$$

4. **Update the active set, the solution estimate, the residual correlation, and λ .**

$$\begin{cases} I = I \cup \{i^+\} & \text{if } s_l = s_l^+ \\ I = I - \{i^-\} & \text{otherwise.} \end{cases}$$

$$x_{l+1} = x_l + s_l d_l$$

$$c_{l+1} = A^T (y - Ax_{l+1})$$

$$\lambda = \|c_{l+1}\|_\infty$$

5. **If residual correlation is zero, then stop; otherwise let $l = l + 1$ and go to Step 2 :**

This algorithm terminates when $\|c_l\|_\infty = 0$, which indicates the solution has been reached.

E. Proof of the Theorem

There are introductory assumptions for the proof of the Theorem.

1. Let $y = Ax$ with $\|x\|_0 = K$ and x has its nonzeros in the first K positions.
2. If two or more coordinates are candidates to enter the active set, assume the algorithm inserts them one at a time, on separate stages.
3. Let x_l denote the Homotopy solution at the l th step, $r_l = y - Ax_l$ the residual at the step, and $c_l = A^T r_l$ the corresponding residual correlation vector.

Definition (Correct Term Selection Property). The Homotopy algorithm has the *Correct Term Selection Property* at a given problem $y = Ax$, if at each iteration, the algorithm selects a new term to enter the active set from the support of x .

If the Homotopy has the the *Correct Term Selection Property*, at the termination, the support set of the solution is guaranteed to be a *subset* of the support set of x .

Definition (Sign Agreement Property). Homotopy algorithm has the *Sign Agreement Property* if at every step l for all $j \in I$ $\text{sgn}(x_l(j)) = \text{sgn}(c_l(j))$.

That is, the Homotopy has the sign agreement property if, at every step of the algorithm, the residual correlations in the active set agree in sign with the corresponding solution coefficients. This ensures that the algorithm never removes elements from the active set.

Lemma 1. *The Homotopy algorithm has the K -step solution property if and only if it has the correct term selection property and the sign agreement property.*

Proof: Converse proof first. After K steps, correct term selection property implies that the active set is a subset of the support of x , i.e., $I \subseteq \{1, 2, \dots, K\}$. The sign agreement property ensures that no variable leaves the active set. Thus, after K steps, $I = \{1, 2, \dots, K\}$, the Homotopy algorithm recovers the correct sparsity pattern. To show that the algorithm terminates at the K step, the step-size s_K is chosen so that, for some $j \in I^c$,

$$|c_K(j) - s_K a_j^T A_I d_K(I)| = \lambda_K - s_K$$

with $\lambda_K = \|c_K(I)\|_\infty$.

In addition, for the K th update, we have $A_I d_K(I) = A_I (A_I^T A_I)^{-1} A_I^T r_K = r_K$ since r_K is contained in the column space of A (because $r_K = y - Ax_K$ and $y = Ax$ is noiseless.)

Hence, $c_K(I^c) = A_I^T A_I d_K(I)$ and $s_K = \lambda_K$ is chosen to satisfy the sign agreement constraint. Therefore, the solution at step K has $Ax_K = y$. Since y has a unique representation in terms of the columns of A_I , we conclude that $x_K = x$.

Now let us do the forward part. It is obvious, since violation of either the correct term

selection property or the sign agreement property would result in a number of steps greater than k or an incorrect solution. \square

Lemma 2. (Correct Term Selection) Let the sufficient condition of the Theorem holds, i.e., $K \leq (\mu^{-1} + 1)/2$. Let the residual at the l th step be written as a linear combination of the first K columns in A , i.e., $r_l = \sum_{j=1}^k w_l(j) a_j$. Then the next step of the Homotopy algorithm selects an index from among the first K .

Proof. We will show that at the l th step

$$\max_{1 \leq i \leq K} |\langle r_l, a_i \rangle| > \max_{i > K} |\langle r_l, a_i \rangle| \quad (128)$$

and so at the end of the l th iteration, the active set is a subset of $\{1, 2, \dots, K\}$.

Let $G := A^T A$ denote the Gram matrix of A . Let $\hat{i} = \arg \max_{1 \leq i \leq K} w_l(i)$. The left-hand side of (128) is bounded below by

$$\begin{aligned} \max_{1 \leq i \leq K} |\langle r_l, a_i \rangle| &\geq |\langle r_l, a_{\hat{i}} \rangle| \\ &= \left| \sum_{j=1}^K w_l(j) G_{\hat{i}, j} \right| \\ &\geq |w_l(\hat{i})| - \sum_{j \neq \hat{i}} |G_{\hat{i}, j}| |w_l(j)| \\ &\geq |w_l(\hat{i})| - \mu \sum_{j \neq \hat{i}} |w_l(j)| \\ &\geq |w_l(\hat{i})| - \mu(K-1) |w_l(\hat{i})| \end{aligned} \quad (129)$$

since $\|a_j\|_2^2 = 1$ for all j and $G_{i,j} \leq \mu := \max_{i \neq j} |\langle a_i, a_j \rangle|$ for $j \neq \hat{i}$. On the right-hand side of (128), for $i > K$, we have

$$\begin{aligned} |\langle r_l, a_i \rangle| &\leq \sum_{j=1}^K |w_l(j)| |G_{ij}| \\ &\leq \mu \sum_{j=1}^K |w_l(j)| \\ &\leq K\mu |w_l(\hat{i})|. \end{aligned} \quad (130)$$

Then, we note that (128) holds if the lower bound (129) is greater than or equal to the upper bound (130), i.e.,

$$1 - (K-1)\mu > K\mu. \quad (131)$$

□

Recall that in the Homotopy algorithm the sign agreement condition (121) is used to remove an index i from the active set at stage l if the sign agreement condition is violated.

The following lemma shows that, **when x is sufficiently sparse**, the sign disagreement never happens. Namely, **at each stage of the algorithm**, the residual correlations in the active set **agree in sign** with the direction of change of the corresponding terms in the solution. In other words, the solution moves in the right direction at each step. In particular, it implies that throughout the Homotopy solution path, the sign agreement property is maintained. Again this happens when **when x is sufficiently sparse**, i.e., $K \leq (\mu^{-1} + 1)/2$.

Lemma 3. (Sign Agreement) Suppose that $y = Ax$, where x has only K nonzeros, with K satisfying $K \leq (\mu^{-1} + 1)/2$. For $l \in \{1, 2, \dots, K\}$, let $c_l = A^T r_l$, and the active set I be defined as in (123). Then, the updated direction d_l defined by (12) satisfies

$$\text{sgn}(d(I)) = \text{sgn}(c_l(I)). \quad (132)$$

Proof: Let $\lambda_l = \|c_l\|_\infty$. We will show that $|\mathbf{d}_l(I) - \text{sgn}(c_l(I))| < 1$ for $i \in I$, which means that (132) holds. From (12) and (9), we have $\lambda_l (A_l^T A_l - e_d) d_l(I) = -\lambda_l d_l(I) + c_l(I)$ where e_d is the identity matrix. This yields

$$\begin{aligned} & \|\lambda_l d_l(I) - c_l(I)\|_\infty \\ & \leq \|A_l^T A_l - e_d\|_{(\infty, \infty)} \cdot \|\lambda_l d_l(I)\|_\infty \\ & \leq \frac{1-\mu}{2} \|\lambda_l d_l(I)\|_\infty \\ & \leq \frac{1-\mu}{2} (\|c_l(I)\|_\infty + \|\lambda_l d_l(I) - c_l(I)\|_\infty) \\ & = \frac{1-\mu}{2} (\lambda_l + \|\lambda_l d_l(I) - c_l(I)\|_\infty) \end{aligned} \quad (133)$$

where $\|\cdot\|_{(\infty, \infty)}$ denotes the induced L^∞ operator norm.

Rearranging terms, we get

$$\|\lambda_l d_l(I) - c_l(I)\|_\infty \leq \lambda_l \cdot \frac{1-\mu}{1+\mu} < \lambda_l$$

and so

$$\|\mathbf{d}_l(I) - \text{sgn}(c_l(I))\|_\infty < 1.$$

Thus (21) follows. □

5. Bayesian Compressive Sensing on the Graph Model

There are a number of approaches in the compressive sensing literature relevant to the title of this section. A few of the authors are Dror Baron and Baraniuk, Donoho, *et. al* (Donoho, A. Maleki, and A. Montanari), and others.

Our approach in this section will be different from these other approaches, independently done.

Our main focus in this section has been to

- Derivation of a Belief Propagation (Message Passing) algorithm which determines the posterior probability on the signal values given the observation. In this regard, our algorithm derivation is similar to that of Gallager for his probabilistic decoding method of low density parity check code.
- Derivation of the state distribution and the support set recovery algorithm. Once the support set is recovered, the determination of the signal values will be done via solving the over-determined Least Squares (LS) solution. This process enables early breaking out of the costly iterative process.
- Demonstration that the derived algorithm can be applied to real or complex valued signals, quantized values and elements of the signals and the sensing matrix, as well as to them over Galois field $GF(q)$. Our preliminary results show that $GF(q)$ results well follow the prediction made by the Gilbert Varshamov bounds discussed in Chapter VI.2.

Major research problems still remain include:

- Determination of the limit performance of this Bayesian approach, in comparison with the sufficient conditions we have obtained in the previous sections.
- Connections to the RIP conditions and the sufficient conditions to L1 unique solution. Here the matrix is designed to be sparse, rather than the dense matrices often assumed in the majority of compressive sensing literature. There exists an intriguing relation between the density and the RIP constant.
- Determination of the density of the sensing matrices (the number of non-zero elements). There is a trade-off relation between the density and the performance. Increasing the density, the minimum distance of the code grows, but the iterative algorithm covered in this section shall not work very well. This is what we have

learned from our coding theory class. There the values have been restricted to 1s and 0s. Here the alphabet is much larger, the set of real numbers. In the decoding process, however, the alphabet is often quantized and discretized densities have been used for iterations. Then, the question is to investigate the performance trade-off relation between the quantization effect (noise) and the overall complexity. The other approach is to use a high density sensing matrix and use a Gaussian approximation. When there are many connections in the graph, a dense graph, a Gaussian approximation may work out well. The trade-off here will be that since high density, the matrix will have a large minimum distance, the spark, but the reconstruction perhaps is just an approximation, rather than an exact recovery of the solution. This is interesting research point which may have deep impact to the coding theoretic practice of iterative message passing algorithms.

A. Iterative Compressive Sensing Algorithm

There are two states: $S_t = 0$ or $S_t = 1$ for $t = 0, 1, \dots, N-1$. The state S_t is taken from the 0-1 Bernoulli random variable with parameter $p = \frac{K}{N}$ where K is the approximate sparsity. When the state is zero, the sample x_t is taken from the zero state distribution, say $f_0(x) \sim \mathcal{N}(0, \sigma_0^2)$; or when the state is one, it is taken from the first state distribution, i.e., $f_1(x) \sim \mathcal{N}(0, \sigma_1^2)$. Roughly, therefore, there are K first states out of N positions.

Now the problem is to determine the realization of the input vector $x = (x_0 \ x_1 \ \dots \ x_{N-1})$ given the observation of the syndrome of the following parity check relation

$$\begin{aligned}
 \text{P1:} \quad y := \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{M-1} \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 & -100 & -10 \\ -11 & -1 & 0 & 00 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & & \\ 0 & 0 & 1 & -100 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} & \quad (134) \\
 & =: Fx
 \end{aligned}$$

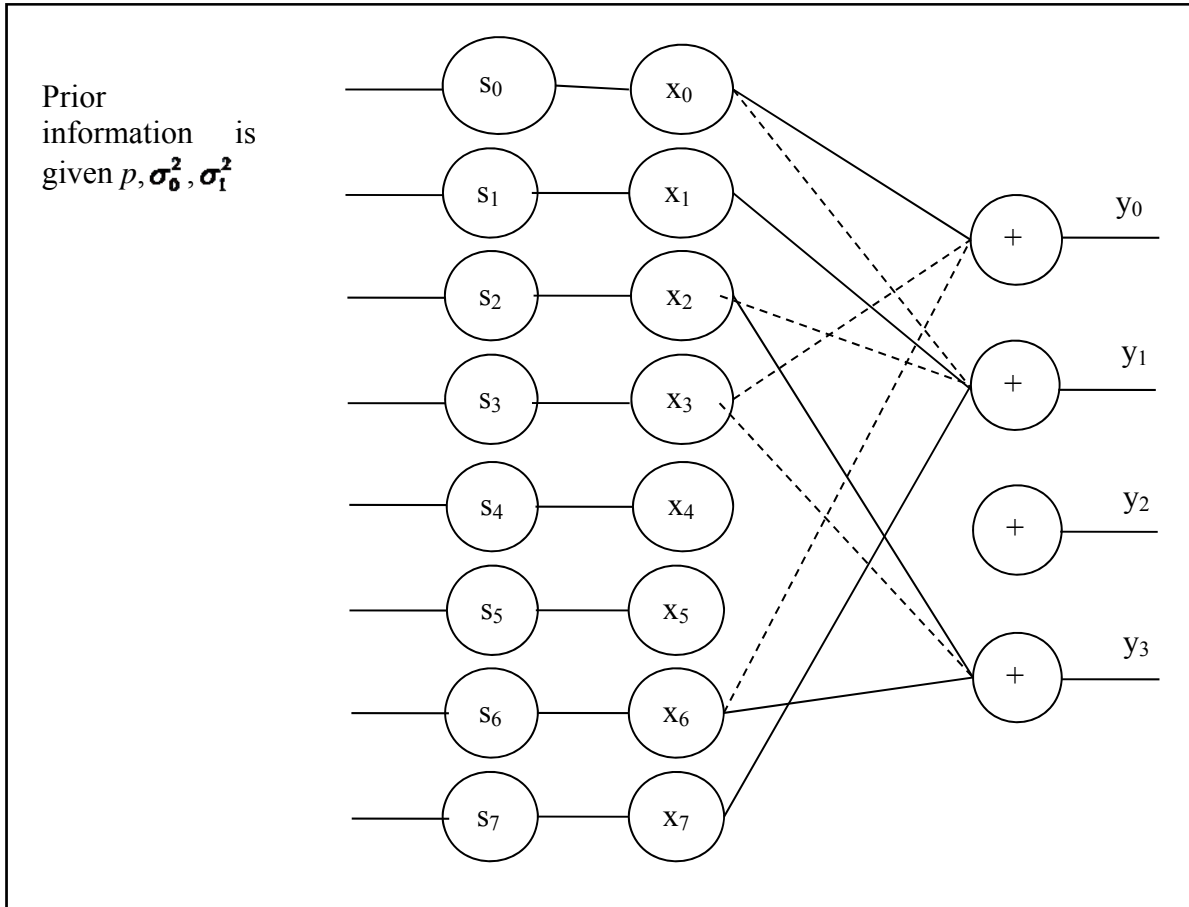


Figure 9: The Graph

We now aim to design the iterative density passing algorithm.

The purpose of the message passing algorithm is to obtain the distributions of the signal values $P(x_0 = \tau_0 | \mathbf{y}, C)$ and the those of the states $\Pr\{S_i = 1 | r, C\}$, use them to determine the support set \mathcal{K} , the non-zero element states, by thresholding, say $\Pr\{S_i = 1 | r, C\} \geq \frac{1}{2}$. Once all the non-zero states have been determined, the unknown values on each of the states can be determined by solving the over-determined Least Squares Estimate(LSE) of the unknown K -sparse vector, i.e., $x_{\mathcal{K}} = (A_{\mathcal{K}}^T A_{\mathcal{K}})^{-1} A_{\mathcal{K}}^T y$. When this estimation is good enough, i.e. $\|r - Ax_{\mathcal{K}}\|_2 \leq \delta$, a certain threshold, one can stop the iteration. A suitable threshold δ can be determined from a probabilistic analysis. Obviously, the larger the threshold, the algorithm can stop earlier but the greater the inclination to errors.

B. The Distribution of the Signal Value

The decoding theorem is given here, which is given as an example to determine the first value $x_{i=0}$ without loss of generality. Note that the first signal node x_0 is connected to the first observation y_0 as well as to the second observation y_1 . The first observation is then connected to the signal nodes x_3 and x_6 ; the second one to x_1 and x_7 .

Theorem 1: The a posteriori probability that the first value, $x_0 = \tau_0$, takes a certain point in

an alphabet \mathcal{X} , given the observation \mathbf{y} and enforcing the checks (checks should be satisfied), is given by

$$\begin{aligned}
 P(x_0 = \tau_0 | \mathbf{y}, C) &= \frac{P(x_0 = \tau_0 | \mathbf{y})}{P(C | \mathbf{y})} \left[\sum_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}} P(C_0 | x_0 = \tau_0, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{y}) P(\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0} | \mathbf{y}) \right] \\
 &\quad \times \left[\sum_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} P(C_1 | x_0 = \tau_0, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}, \mathbf{y}) P(\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1} | \mathbf{y}) \right]
 \end{aligned}
 \tag{135}$$

We apply the same procedure and obtain a similar AP result for each τ_0 in the alphabet. We repeat this procedure for each element of the input vector x . Note that one can make a Maximum A Posteriori decision at any time given the posterior distribution for each element.

Proof: Let us consider the following derivation to determine the distribution of the first value, x_0 . Note that without loss of generality, we can repeat the same procedure for the rest of the variables, x_i s.

$$\begin{aligned}
 P(x_0 = \tau_0 | \mathbf{y}, C) &= \sum_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}} \sum_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} P(x_0 = \tau_0, \underbrace{x_3 = \tau_3, x_6 = \tau_6}_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}}, \underbrace{x_1 = \tau_1, x_2 = \tau_2, x_7 = \tau_7}_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} | \mathbf{y}, C) \\
 &= \sum_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}} \sum_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} \frac{P(x_0 = \tau_0, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}, \mathbf{y}, C)}{P(\mathbf{y}, C)} \\
 &= \sum_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}} \sum_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} \frac{P(C | x_0 = \tau_0, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}, \mathbf{y})}{P(\mathbf{y}, C)} \times \\
 &\quad P(x_0 = \tau_0, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1} | \mathbf{y}) P(\mathbf{y}) \\
 &= \sum_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}} \sum_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} \frac{P(C_0 | x_0 = \tau_0, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{y}) P(C_1 | x_0 = \tau_0, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}, \mathbf{y})}{P(C | \mathbf{y})} \times \\
 &\quad P(x_0 = \tau_0, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1} | \mathbf{y}) P(\mathbf{y}) \\
 &= \frac{P(x_0 = \tau_0 | \mathbf{y})}{P(C | \mathbf{y})} \left[\sum_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}} P(C_0 | x_0 = \tau_0, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{y}) P(\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0} | \mathbf{y}) \right] \times \\
 &\quad \left[\sum_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} P(C_1 | x_0 = \tau_0, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}, \mathbf{y}) P(\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1} | \mathbf{y}) \right]
 \end{aligned}
 \tag{136}$$

Q.E.D.

A single round of calculation of the posterior distribution $\{P(x_t = \tau_t | \mathbf{y}, C) : \tau_t \in \mathcal{X}\}$ for each and every variable, $t = 0, 1, \dots, N-1$, constitute a single iteration. We repeat the iteration again and again. In a single iteration, all N different posterior distributions are updated once. Why do we need iterations? Why can't we be just satisfied with a single iteration? This has to do with our choice on the density of the sensing matrix. We have chosen sparse sensing matrix. Thus, the graph is sparsely connected. This is because that way the iterative algorithm works well. In one iteration, only local information is gathered. Through many iterations, the whole information available from the entire observation \mathbf{y} and from the checking relations presented in the sparse graph, can be gathered. Enough number of iterations should be repeated before an attempt to make a Maximum A Posteriori (MAP) decision on the value of each signal node. One may choose to apply the MAP rule on the posterior distribution obtained at the end of a prescribed number of iterations. That is, we determine the value of each variable via the following rule

$$\hat{x}_t := \arg \max_{\tau_t} P(x_t = \tau_t | \mathbf{y}, C) \quad (137)$$

for each t , $t = 0, 1, \dots, N-1$.

Note that $\left[\sum_{\mathbf{x}_{0,0}=\tau_{0,0}} P(C_0 | x_0 = \tau_0, \mathbf{x}_{0,0} = \tau_{0,0}, \mathbf{y}) P(\mathbf{x}_{0,0} = \tau_{0,0} | \mathbf{y}) \right]$ is obtained from the convolution of probability density functions for random variables x_3 and x_6 . Similarly, $\left[\sum_{\mathbf{x}_{0,1}=\tau_{0,1}} P(C_1 | x_0 = \tau_0, \mathbf{x}_{0,1} = \tau_{0,1}, \mathbf{y}) P(\mathbf{x}_{0,1} = \tau_{0,1} | \mathbf{y}) \right]$ for all x_0 is obtained from the convolution of random variables x_1, x_2 and x_7 . The convolution operations can be computed in the frequency domain using FFT and IFFT.

For clarity, we use the following examples.

Example) From the system of equations $\mathbf{y} = F\mathbf{x}$, we note that the first equation is $x_0 = y_0 + x_3 + x_6$; and the second equation is $x_0 = y_1 - x_1 + x_2 - x_7$.

$$\mathbf{y} := \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{M-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & -100 & -10 \\ -11 & -1 & 0 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & -100 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix}$$

Here y_0 is an observed value; thus, a determined and given value. We know the distributions of the random variables x_3, x_6 , and those of x_1, x_2 and x_7 . We aim to find the distribution of x_0 that satisfies the two equations simultaneously. For the first equation, we note that, the distribution of x_0 is determined by the convolution of the distribution of x_3 and that of x_6 . For the second equation, it is determined by the convolution of the distributions of x_1, x_2 and x_7 : One may obtain the convolution of x_1, x_2 first; and the output of this convolution is convolved again with the distribution of x_7 .

C. The Distribution of the Signal Value in the Presence of Noise

We now consider a little bit different problem. We choose $\sigma_0^2=0$ in P1. Thus, the vector x is exactly sparse, which means that when the state is zero, the value of the variable node is zero with probability 1. In addition, the observation is made under the additive white Gaussian noise with zero mean and variance of $\frac{N_o}{2}$ where N_o is the single sided power spectral density of the noise.

Now the problem is to determine the realization of the input vector $\mathbf{x} = (x_0 x_1 \cdots x_{N-1})$ given the noisy observation of the syndrome of the following parity check relation

$$\begin{aligned} \mathbf{r} := \begin{pmatrix} r_0 \\ r_1 \\ \vdots \\ r_{M-1} \end{pmatrix} &= \begin{pmatrix} 1 & 0 & 0 & -100 & -10 \\ -11 & -1 & 0 & 00 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & & \\ 0 & 0 & 1 & -100 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ \vdots \\ x_{N-1} \end{pmatrix} + \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_{M-1} \end{pmatrix} \\ &=: \mathbf{F}\mathbf{x} + \mathbf{w} \\ &=: \mathbf{y} + \mathbf{w} \end{aligned} \quad (138)$$

Theorem 2. The decoding theorem is given by the following equation. It is given as an example to determine the first value $x_{i=0}$ without loss of generality. It is the result of the Maximum A Posteriori criterion which is used to determine the first value x_0 :

$$\begin{aligned} P(x_0 = \tau_0 | \mathbf{r}, C) &= \sum_{\mathbf{w}} P(x_0 = \tau_0 | \mathbf{w}, \mathbf{r}, C) p(\mathbf{w}) \\ &= \sum_{\mathbf{w}} P(x_0 = \tau_0 | \mathbf{y} = \mathbf{r} - \mathbf{w}, C) p(\mathbf{w}) \end{aligned} \quad (139)$$

The proof is obvious and omitted. We can use Theorem 1 to determine $P(x_0 = \tau_0 | \mathbf{y} = \mathbf{r} - \mathbf{w}, C)$. Thus, the result can be obtained by averaging it over the distribution of the noise.

D. Distribution of the Binary State Value

Since there are only two states, it might be more robust to determine the state at each node, instead of the exact value x_i . Once the states are found, the over-determined problem can be solved, and the unknown sparse vector \mathbf{x} can be found.

Given the distribution $P(x_0 = \tau_0 | \mathbf{r}, C)$, there are number of ways we can obtain the LLRs for the state likelihoods.

1. Use the Kulback-Leibler Distance to the distribution of the zero state probability, i.e., $D(P(x_0 | \mathbf{r}, C) || f_0(x))$ and to that of the non-zero state probability, i.e.,

$D(P(x_0 | \mathbf{r}, C) \| f_1(x))$. Compare the result and determine the state value.

2. Find the probability of the state value given the prior distribution and the posterior $P(x_0 | \mathbf{r}, C)$. We are given from the problem set up the conditional probability distribution $p\{x_0 = \tau | S_0 = 0\} = f_0(\tau)$ and $p\{x_0 = \tau | S_0 = 1\} = f_1(\tau)$, and the prior distribution on the state values, namely, the state S_t is the 0-1 Bernoulli($p = \frac{\kappa}{N}$), i.e. $\Pr\{S_0 = 1\} = p$. Next suppose that observation r is made, and the posterior distribution $p(x_0 = \tau | r, C)$ has been obtained by enforcing the parity-checking relations to r . Now, we ask

$$\begin{aligned} \gamma_0 &:= \Pr\{S_0 = 1 | r, C\} = \int p\{S_0 = 1, x_0 = \tau | r, C\} d\tau \\ &= \int \Pr\{S_0 = 1 | x_0 = \tau, r, C\} p\{x_0 = \tau | r, C\} d\tau \end{aligned} \quad (140)$$

We note the term $\Pr\{S_0 = 1 | x_0 = \tau, r, C\}$. It is equal to $\Pr\{S_0 = 1 | x_0 = \tau\}$ once the value x_0 is given, r and C do not provide any additional information to the state value.

Now, let us work on the conditional probability $\Pr\{S_0 = 1 | x_0 = \tau\}$

$$\begin{aligned} \Pr\{S_0 = 1 | x_0 = \tau\} &= \frac{\Pr\{S_0 = 1, x_0 = \tau + d\tau\}}{\Pr\{x_0 = \tau + d\tau\}} \\ &= \frac{p(x_0 = \tau | S_0 = 1) \Pr(S_0 = 1)}{p(x_0 = \tau)} \\ &= \frac{\frac{\kappa}{N} f_1(x_0 = \tau)}{\frac{\kappa}{N} f_1(x_0 = \tau) + (1 - \frac{\kappa}{N}) f_0(x_0 = \tau)} \end{aligned} \quad (141)$$

Finally, one can design the decision maker, i.e.,

$$\text{Binary State Decision: } S_0 = \begin{cases} 1, & \text{if } \gamma_0 \geq 0.5 \\ 0, & \text{o.w.} \end{cases} \quad (142)$$

Theorem 3. (The Posterior Binary State Distribution) Let S_t is the 0-1 Bernoulli($p = \frac{\kappa}{N}$). Let the conditional probability distribution on the signal values at different states be $p\{x_0 = \tau | S_0 = 0\} = f_0(\tau)$ and $p\{x_0 = \tau | S_0 = 1\} = f_1(\tau)$ respectively. Let $p\{x_t = \tau | r, C\}$ be obtained from the message passing algorithm given by Theorem 2. Then, the probability of state S_t , $t = 0, 1, \dots, N-1$, given r and applying the check relation C , is obtained as

$$\begin{aligned} \gamma_t &:= \Pr\{S_t = 1 | r, C\} \\ &= \int \frac{\frac{\kappa}{N} f_1(x_0 = \tau)}{\frac{\kappa}{N} f_1(x_0 = \tau) + (1 - \frac{\kappa}{N}) f_0(x_0 = \tau)} p\{x_t = \tau | r, C\} d\tau \end{aligned} \quad (143)$$

for $t = 0, 1, 2, \dots, N-1$.

Support Set Recovery with Log Likelihood Ratios

We note that $\Pr\{S_t = 1 | r, C\}$ may not be a probability measure given that $\Pr\{S_t = 1 | r, C\} + \Pr\{S_t = 0 | r, C\}$ is not equal to zero. This may be due to the approximation we have made in (140) for $\Pr\{S_0 = 1 | x_0 = \tau, r, C\}$, with $\Pr\{S_0 = 1 | x_0 = \tau\}$ (141).

To resolve this issue, one may use normalization. The other approach is to use the log ratio. We may opt to use the log ratio because the algorithm may be simplified with the log-ratio approach. Let us see.

$$\begin{aligned} \text{LR}(S_0) &:= \log \frac{\Pr\{S_0 = 1 | r, C\}}{\Pr\{S_0 = 0 | r, C\}} \\ &= \log \frac{\int \frac{\frac{K}{N} f_1(x_0 = \tau)}{\frac{K}{N} f_1(x_0 = \tau) + (1 - \frac{K}{N}) f_0(x_0 = \tau)} p\{x_0 = \tau | r, C\} d\tau}{\int \frac{(1 - \frac{K}{N}) f_0(x_0 = \tau)}{\frac{K}{N} f_1(x_0 = \tau) + (1 - \frac{K}{N}) f_0(x_0 = \tau)} p\{x_0 = \tau | r, C\} d\tau} \\ &= \log \frac{\int \frac{K}{N} \tilde{f}_1(x_0 = \tau) p\{x_0 = \tau | r, C\} d\tau}{\int (1 - \frac{K}{N}) \tilde{f}_0(x_0 = \tau) p\{x_0 = \tau | r, C\} d\tau} \end{aligned}$$

Note that it is in the form of log-sum-product. Express the integral operation as the Riemann summation. Then, we have

$$\begin{aligned} &\log \frac{\int \frac{K}{N} \tilde{f}_1(x_0 = \tau) p\{x_0 = \tau | r, C\} d\tau}{\int (1 - \frac{K}{N}) \tilde{f}_0(x_0 = \tau) p\{x_0 = \tau | r, C\} d\tau} \\ &= \log \left(\frac{K}{N - K} \right) + \log \left(\sum_{\tau} \exp(lp_{1,\tau}) \right) - \log \left(\sum_{\tau} \exp(lp_{0,\tau}) \right) \end{aligned}$$

where

$$\begin{aligned} lp_{1,\tau} &= \log \left(\tilde{f}_1(x_0 = \tau) p\{x_0 = \tau | r, C\} \right) \\ lp_{0,\tau} &= \log \left(\tilde{f}_0(x_0 = \tau) p\{x_0 = \tau | r, C\} \right) \end{aligned}$$

Note that the argument inside the logarithm is the product of two probability measures; thus, log of the product should be negative.

Now, let us discuss one of the log products

$$\begin{aligned} lp_{1,\tau} &= \log \left(\tilde{f}_1(x_0 = \tau) p\{x_0 = \tau | r, C\} \right) \\ &= \log \left(\tilde{f}_1(x_0 = \tau) \right) + \log \left(p\{x_0 = \tau | r, C\} \right) \end{aligned}$$

Giving attention to the second term, we have

$$\begin{aligned} \log(p\{x_0 = \tau | r, C\}) &= \log\left(\frac{P(x_0 = \tau | \mathbf{y})}{P(C | \mathbf{y})}\right) \\ &+ \log\left[\sum_{\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}} P(C_0 | x_0 = \tau, \mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0}, \mathbf{y}) P(\mathbf{x}_{0,0} = \boldsymbol{\tau}_{0,0} | \mathbf{y})\right] \\ &+ \log\left[\sum_{\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}} P(C_1 | x_0 = \tau, \mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1}, \mathbf{y}) P(\mathbf{x}_{0,1} = \boldsymbol{\tau}_{0,1} | \mathbf{y})\right] \end{aligned}$$

Now, we note that the second and the third term are in the form of log-sum-product again. Let us find a way to simplify them.

First, let us use the FFT technique to get rid of the cumbersome convolution operation. Recall that convolution in the time domain (the pdfs in this problem) is equivalent to the multiplication in the frequency domain (the product of characteristic function).

We may apply the FFT technique to express the convolution. For example, the first term is the convolution of two random variables x_3 and x_6 , i.e., $x_3 + x_6 = y_0 - \tau$. Let us denote the FFT and IFFT operators as F and IF . Then, the pdf of the convoluted random variable is obtained by

$$\begin{aligned} pdf_{3,6}(t) &= pdf_3 \star pdf_6 \\ &= IF\left(F_3(pdf_3) \times F_6(pdf_6)\right)(t) \end{aligned} \tag{144}$$

Then, substitute $t = y_0 - \tau$ and we have

$$\log(p\{x_0 = \tau | r, C\}) = \log\left(\frac{P(x_0 = \tau | \mathbf{y})}{P(C | \mathbf{y})}\right) + \log[pdf_{3,6}(y_0 - \tau)] + \log[pdf_{1,2,7}(y_1 - \tau)].$$

Now, substitute the result to the previous expression:

$$\begin{aligned} lp_{1,\tau} &= \log(\tilde{f}_1(x_0 = \tau) p\{x_t = \tau | r, C\}) \\ &= \log(\tilde{f}_1(x_0 = \tau)) + \log\left(\frac{P(x_0 = \tau | \mathbf{y})}{P(C | \mathbf{y})}\right) + \log[pdf_{3,6}(y_0 - \tau)] + \log[pdf_{1,2,7}(y_1 - \tau)] \end{aligned}$$

Similarly we have

$$\begin{aligned}
lp_{0,\tau} &= \log(\tilde{f}_0(x_0 = \tau) p\{x_0 = \tau | r, C\}) \\
&= \log(\tilde{f}_0(x_0 = \tau)) + \log\left(\frac{P(x_0 = \tau | \mathbf{y})}{P(C | \mathbf{y})}\right) + \log[pdf_{3,6}(y_0 - \tau)] + \log[pdf_{1,2,7}(y_1 - \tau)]
\end{aligned}$$

Combining the results so far, we have

$$\begin{aligned}
\log \frac{\Pr\{S_0 = 1 | r, C\}}{\Pr\{S_0 = 0 | r, C\}} &= \log\left(\frac{K}{N-K}\right) + \log\left(\sum_{\tau} \exp(lp_{1,\tau})\right) - \log\left(\sum_{\tau} \exp(lp_{0,\tau})\right) \\
&= \log\left(\frac{K}{N-K}\right) + \underbrace{\max_{\tau} \{lp_{1,\tau}\}}_{=: \max_1} + \log\left(1 + \sum_{\tau \neq \max \text{ index}} \exp[-(\max_1 - lp_{1,\tau})]\right) \\
&\quad - \underbrace{\max_{\tau} \{lp_{0,\tau}\}}_{=: \max_0} - \log\left(1 + \sum_{\tau \neq \max \text{ index}} \exp[-(\max_0 - lp_{0,\tau})]\right)
\end{aligned}$$

Now, it becomes very simple to do the operation.

E. Useful Mathematical Identities for Log-Sum-Products

Product of real numbers

$$\prod_i a_i = \prod_i \operatorname{sgn}(a_i) \exp\left(\sum_i \log(|a_i|)\right)$$

An example is $ab = \operatorname{sgn}(a)\operatorname{sgn}(b)\exp[\log|a| + \log|b|]$.

A logarithmic identity

We may use the following identity on the log sum of more than two numbers for positive numbers A , B , and C ,

$$\begin{aligned} \log(A+B+C) &= \log(\exp(\log A) + \exp(\log B) + \exp(\log C)) \\ &= \log(e^a + e^b + e^c) \\ &= \log(e^a(1 + e^{-(a-b)} + e^{-(a-c)})) \\ &= a + \log(1 + e^{-(a-b)} + e^{-(a-c)}). \end{aligned} \tag{145}$$

where it is supposed that $a = \max(a, b, c)$.

F. The Message Passing Algorithm

The message passing algorithm is given as the following:

1. Initialization: Set $P(x_t = \tau_t | \mathbf{y}) = \frac{\kappa}{N} f_1(x_t = \tau_t) + (1 - \frac{\kappa}{N}) f_0(x_t = \tau_t)$ for all t . Determine a threshold δ for stopping criterion.
2. Run message passing routine: Do the convolution (or the FFT/IFFT) routine for each t , obtaining $P(x_t = \tau_t | \mathbf{y}, C)$ for all t .
3. Run the active set recovery routine. An index t will be decided to be added to the active if the log ratio, $LR(S_t)$, for $t = 0, 1, 2, \dots, N-1$, is greater than zero, i.e.,

$$I = \{t : LR(S_t) > 0.0\}$$

4. Check if I is \mathcal{K} : Run $x_t = (A_t^T A_t)^{-1} A_t^T y$. When this value is good enough, i.e. $\|r - Ax_t\|_2 \leq \delta$ the threshold, the iteration can be put to stop. Otherwise, return to step 2 and repeat.

G. Donoho's message passing algorithm for compressed sensing

Donoho et al. discuss a new iterative algorithm which builds upon a standard linear programming (LP) based iterative algorithm. According to the authors, this new algorithm brings dramatically faster reconstruction than the LP-based approach. The idea for improvement seems to be partly borrowed from the iterative message passing algorithm on bipartite code graphs popularly used in the coding theory community (They refer to Richardson and Urbanke's *Modern Coding Theory*, Cambridge University Press, 2008). One key difference could be the addition of a corrective term to the iterative algorithm. Research in this direction shall be very interesting.

Reference

1. D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," vol.106, no.45, *Proceedings of National Academy of Science*, Nov. 10, 2009.

6. The Expander Graph Approach

In previous sections, we have demonstrated that the solution to the under-determined linear system of equations could be found using the L1 optimization approach, which has been known in the literature as “basis pursuit”. While L1 optimization can be done in polynomial-time (often $O(N^3)$), this may still be infeasible in applications where N is quite large. In [1], motivated by the existence of bipartite expander graphs [6], a new scheme for compressive sensing with deterministic performance guarantees based on expander graphs was proposed. The recovery complexity of this algorithm is $O(K)$, where K is the number of non-zero entries in the signal vector. Research in this direction shall be of interest.

7. Section References

- [1] Sina Jafarpour , Weiyu Xu , Babak Hassibi, and Robert Calderbank, “Efficient and robust compressed sensing using optimized expander graphs” *IEEE Transactions on Information Theory*, 55 (9). pp.2009.
- [2] W.Xu and B.Hassibi, “Efficient compressive sensing with deterministic guarantees using expander graphs,” *Proceedings of IEEE Information Theory Workshop, Lake Tahoe, 2007.*
- [3] W.Xu and B.Hassibi, “Further results on performance analysis for co-mpressive sensing using expander graphs,” *Conference Record of the Forty-First Asilomar Conference on Signal, Systems and Computers. ACSSC 2007.4-7 Nov.2007* Pages(s):621-625,2008
- [4] E. Candes, J. Romberg and T. Tao, “Robust Uncertainty Principles: Exact Signal Reconstruction from Highly Incomplete Frequency Information”, *IEEE Transactions on Information Theory* , 52(2) pp.489-509, Feb.2006.
- [5] E. Candes and T. Tao, “Near Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?”, *IEEE Trans.on Information Theory*, 52(12) pp. 5406 - 5425, Dec. 2006.
- [6] Michael Sipser and Daniel A. Spielman “Expander codes”, *IEEE Transactions on Information Theory* , vol.42, NO.6, Nov, 1996.
- [7] L.A. Bassalygo and M.S. Pinsker, “Complexity of an optimum nonblocking switching network without reconnections”, *Problems in Information Transmission*, vol 9 no 1, pp. 84-87, 1973.
- [8] Fan Chung and Linyuan Lu, “Concentration inequalities and martingale inequalities: A survey”, *Internet Mathematics* Vol.3, No.1:79-127.
- [9] E. Candes and T. Tao, “Decoding by Linear Programming”, *IEEE Trans. on Information Theory*, 51(12), pp. 4203 - 4215, Dec. 2005.

8. Chapter Problems

Design a small scale $y = Ax$ problem. The dimension of y should be at least 100 by 1. The dimension of A should be 100 by 200. The unknown x should be a K -sparse vector. Let $K = 10$. A is randomly generated from the Gaussian distribution. Make sure that the columns of A have energy 1 by normalizing each column.

6. Consider the linear program given in class. Namely, it is

$$\begin{aligned} \min_{(x,u)} \sum_i u_i &= 1^T u + 0^T x = \begin{bmatrix} 0^T & 1^T \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \\ \text{s.t. } x - u &= \begin{bmatrix} e & -e \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \preceq 0, \\ -x - u &= -\begin{bmatrix} e & e \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \preceq 0, \\ -u &= \begin{bmatrix} 0 & -e \end{bmatrix} \begin{pmatrix} x \\ u \end{pmatrix} \preceq 0, \\ [A \ 0] \begin{pmatrix} x \\ u \end{pmatrix} - b &= 0 \end{aligned}$$

- Use the MATLAB's standard routine to obtain the solution of linear program solution. Keep it for reference.
- Note that the constraint includes the non-negativeness of u . The programs in L_1 -magic package by Candes and Tao do not include this. Seeing the effect of including this term is one of our objectives.
- Use the Lagrange-dual interior point method to do the L_1 minimization. First obtain the KKT conditions. Write them down.
- Next obtain the Jacobian matrix discussed in class. Specify each and every component matrices. Specify their dimensions.
- Construct the two equations, one for $\Delta\lambda$ and the other for $(\Delta x, \Delta v)$.
- Construct the Netwon's step routine. Verify its correct operation by doing a small problem of your own with known solution. Explain how you have verified its correct operation. Include the verification results.
- Implement the step-size selection routine in MATLAB. Complete the

MATLAB program implementation by including the stopping condition.

- h. Verify the operation of your algorithm by solving the small scale problem.
 - i. Provide the evidence of the convergence of the solution estimate. Show the duality gap narrows as iteration goes on. Show that the mean square error between the solution estimate and the true solution narrows as the iteration goes on.
7. Repeat the first problem with the Log-Barrier method.
- a. Obtain the Hessian by specifying the component matrices. Specify their dimensions.
 - b. Construct the matrix vector equation with the unknown $(\Delta x, v)$.
 - c. Implement the Netwoen method routine in MATLAB.
 - d. As done in the first problem, provide the evidence that your program is working properly by making comparison with the first examples.
8. Repeat the first problem with the Homotopy algorithm. Do the full verification as to show that your algorithm is working properly.
9. Compare the recovery performance of the four algorithms: the MATLAB's built-in routine, the primal-dual interior point method, and the log-barrier method. To compare, do a Monte carlo simulation in which the sensing matrices are randomly generated and the sparse-signals are randomly generated. Fix the matrix size as 100 x 200, try to see the performance as you increase the sparsity $K = 10, 20, 30, 40, 50$.
10. Use the Bayesian method and verify your routine. Include into the comparison.
11. Use the log-barrier algorithm for SOCPs and solve TV_1 Problem. Do the Shepp-Logan phantom example. For this, refer to page 9 of the l1-magic program manual provided by Candes and Romberg.

Chapter VI. RECENT RESEARCH RESULTS

1. Deterministic Matrix Design

This part will be added soon.

2. Coding Theoretic Approach

This section is adopted from [8] (Section Reference). Compressed sensing has provided a new signal acquisition framework with which one can take samples of a given signal of interest while compressing it simultaneously. This compressed sample taking is done via linear projection of the given signal against a prescribed set of kernels, i.e., one linearly projected sample per kernel. In its standard form, this compressed sensing operation is developed over the field of real numbers. In this presentation, we are interested in the development of compressed sensing over the finite fields. Fundamental limits on sensing measurement requirements as we vary the size of the finite field will be discussed. When compressed sensing is put to work in digital systems, the signals and the kernels should be represented in a finite precision manner anyhow; thus, the study of compressed sensing over finite fields should be of interest for implementation point of view. We aim to present our understanding on compressive sensing via Gilbert Varshamov (GV) bounds, new results on average *spark* calculation results, and proposal of the a posteriori (AP) signal recovery algorithm, and provide discussion on how they are related with each other.

We make note of existence of a few prior studies relevant to the content of this section. Draper and Malekpour [2] have studied compressed sensing over finite fields and obtained fundamental bounds on sensing requirements using the error exponent analysis techniques of the channel coding theory. Ardestanizadeh, Cheraghchi, and Shokrollahi [6] have studied the question how much bit precision on the compressive measurements will be needed for good recovery of sparse signals of a finite size alphabet, say q . They assumed the use of Vandermonde frames [5] and obtained that the precision requirement is $O(K \log_2 q + K^2 \log \frac{N}{K})$. Zhang and Pfister [4] discussed the connection between compressed sensing and error correction codes, and proposed the use of low density parity check matrices over $\text{GF}(q)$ and a verification based iterative decoding schemes.

A. Compressed Sensing via Syndrome Decoding

In this section, we aim to draw analogy between parity checking in coding theory and the under-determined equation in compressed sensing by recasting the basic compressed sensing equation

$$y = Fx \tag{146}$$

as a coding theoretic parity-checking equation. Treat y as an $M \times 1$ syndrome vector, F as an $M \times N$ parity check matrix, $M < N$, and x as an $N \times 1$ K -sparse error vector. Note that this model is valid for real, complex, and finite fields $\text{GF}(q)$. Finite fields can be useful for implementing the CS system in digital forms, with a finite precision representation, say $\log_2(q)$ bit precision, done to the coefficients of the elements of the sensing matrices and signals.

We assume that $K \leq t$ where t means the number of errors a given code defined by an F can correct. Let $U = N - M$. The rate R of the code is U/N . We can then find the $N \times U$ generator matrix G from F using the relationship that $FG = 0$ (e.g. using Gaussian elimination on F) where 0 denotes the $M \times U$ all zero matrix. Let \mathcal{C} be the codebook—collection of all codewords. Each $N \times 1$ codeword c can be generated by multiplying an arbitrary $U \times 1$ message vector m to the generator matrix, i.e., $c = Gm$. We assume c is sent over a noisy channel where the noisy channel introduces an additive random error pattern x to c , and the output of the channel is $z = c + x$.

In this setting, parity checking on z shall return the zero syndrome, i.e., $y = Fz = F(c + x) = Fx$, unless there is zero errors or the error pattern x is a codeword, i.e. $x \in \mathcal{C}$; otherwise it will give a non-zero syndrome vector. The code is linear and hence it contains the all-zero codeword. The error correction capability of this code \mathcal{C} can be parameterized by its minimum distance d_{\min} . The minimum distance d_{\min} is the minimum Hamming weight (the number of non-zero coefficients) of any codeword, since the code is linear, i.e.,

$$d_{\min} \triangleq \min_{c \neq 0, c \in \mathcal{C}} w_H(c). \quad (147)$$

But a codeword is a word that satisfies the parity check equation, i.e., $Fx = 0$. From this observation, we may also write that d_{\min} is also the smallest number d that there exists a set of d columns of a matrix F that are linearly dependent; this definition is the same as that of the spark in compressed sensing. This discussion will continue further in Section B. From the coding theory, we note that, a code defined by its parity check matrix F with d_{\min} can correct *all t or smaller error patterns*, and t is given by

$$t = \left\lfloor \frac{d_{\min} - 1}{2} \right\rfloor \approx \frac{d_{\min}}{2}. \quad (148)$$

Our discussion up to (148) implies that all K -sparse error vectors x can be uniquely determined from the syndrome equation $y = Fx$ as long as $K \leq t$. Notice that this is a deterministic guarantee, rather than probabilistic, on the recovery of the sparse vectors. Such a code or a matrix F with d_{\min} can be constructed. Namely, we can construct an F so that any collection of less than or equal to $d_{\min} - 1$ columns of F is linearly independent. This means that $d_{\min} - 1$ can be as large as the rank of F which is further upper bounded by M since $M < N$. Hence, we have the Singleton bound,

$$d_{\min} - 1 \leq M. \quad (149)$$

Those codes that achieve the Singleton bound with equality are called *maximum distance*

separable codes. They include the *repetition* codes and *Reed Solomon* (RS) codes. Real- or complex-valued RS code like sensing matrix F with d_{min} can also be found. The examples are given in [5],[7]. From (149), we can obtain $\frac{M}{2N} \geq \frac{d_{min}}{2N} + \frac{1}{2N}$, by dividing both sides by $2N$. By defining the compression ratio $\rho_{comp} \triangleq \frac{M}{N}$ and the error correction ratio (ECR) $\rho_t \triangleq \frac{t}{N}$, we have

$$\rho_{comp} \geq 2\rho_t. \quad (150)$$

We call (150) the *CS Singleton bound*. Any x whose sparsity ratio $\rho_{sp} \triangleq \frac{K}{N}$ is smaller than or equal to ECR (i.e., $K \leq t$), can be uniquely determined from syndrome y . Fig. 1 shows the CS Singleton bound.

On the other hand, the Gilbert-Varshamov bound tells us the existence of a t error correcting linear block code. The rate R of such a code is given by,

$$R(\delta) \geq 1 - H_q(\delta) \quad (151)$$

where $\delta \triangleq \frac{d_{min}}{N}$, $R = \frac{N-M}{N}$ and $H_q(\delta)$ is the q -ary entropy function. It is the lower bound on the rate required to have the relative minimum distance δ . Eq. (151) can then be written as,

$$\rho_{comp} \leq H_q(2\rho_{sp}) \quad (152)$$

for $\rho_{sp} \in [0, 0.5]$.

It is interesting to note that $H_q(2\rho_{sp})$ approaches the line with slope 2 as q increases. The required code rate can be as large as what this lower bound predicts for a long block length. It is then an upper bound on the redundancy. The number of check equations required for a sensing matrix to have the relative minimum distance is at most what this bound can tell us. One needs at most this much redundancy to be able to find a sensing matrix with the relative minimum distance δ . It can be shown that an ensemble of parity check (PC) codes, say $\mathcal{C}(N, d_s, d_c, q)$ block codes of length N , check degree d_c , signal element degree d_s , and $\text{GF}(q)$, closely approach the GV bound from *above* as the degrees are increased. Thus, GV bounds in fact work as a benchmark, instead of upper bounds. The check degree and the signal element degree indicate the number of non-zero entries in any row of a sensing matrix and the number of non-zero entries in any column respectively. We focus on the cases here that the degrees are fixed for each row and column. Thus, for a compressed sensing system with a large field $\text{GF}(q)$, the sufficient condition is going to approach

$$\rho_{comp} \gtrsim 2\rho_{sp}. \quad (153)$$

This means that if $\rho_{comp} \gtrsim 2\rho_{sp}$, a good sensing matrix exists and thus one can be found. As the dimension of the system approaches infinity, a randomly selected code out of an ensemble will behave as good as what these bounds can predict, with probability getting close to 1.

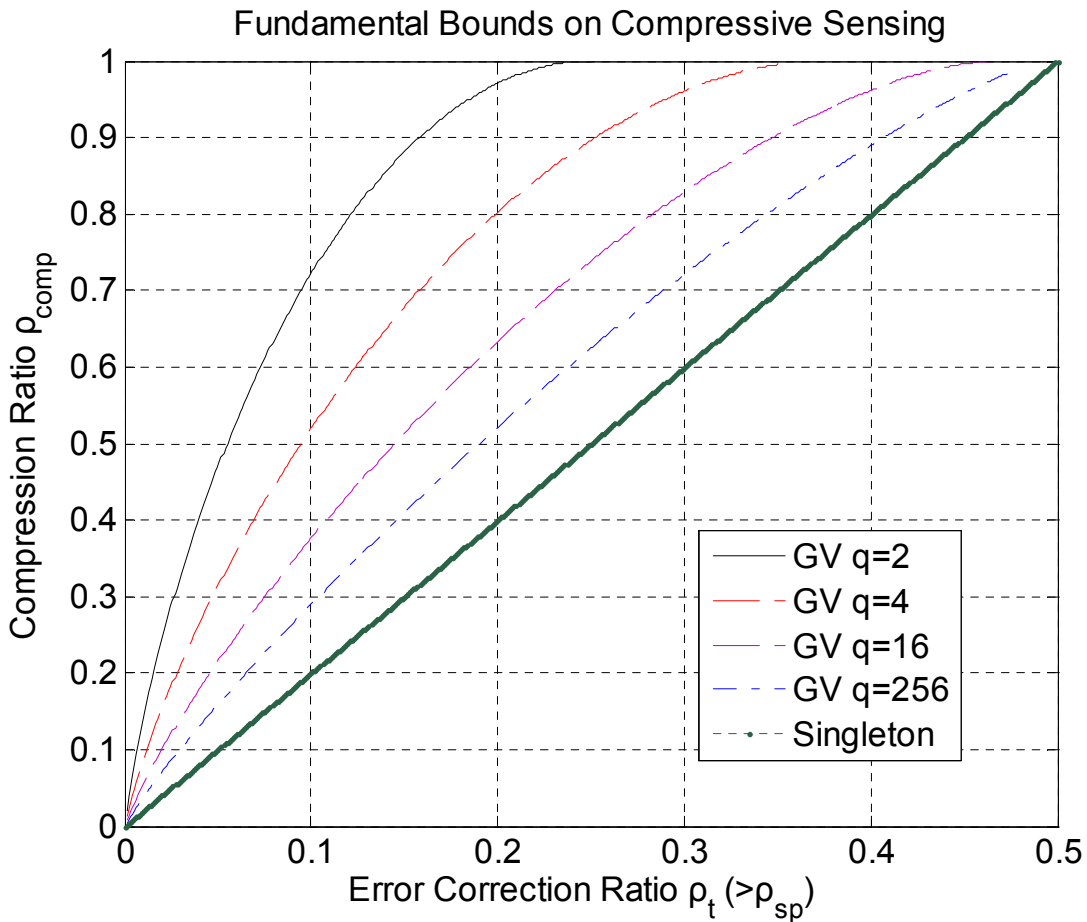


Figure 10 : Gilbert-Varshamov Compressed Sensing Bounds for Sensing Matrices over $GF(q)$ and The Singleton Bound

B. The Sparks of Sensing Matrices over $GF(q)$

In this section, we aim to find the ensemble average *sparks* of an $GF(q)$ LDPC codes. The *spark* of an $M \times N$ matrix is the smallest number S such that there exists a set of S columns of the matrix that are linearly dependent. One should note that spark for a sensing matrix and d_{min} for a parity check matrix are exactly the same. From the Singleton bound, then, $S-1 \leq M$. Finding the SPARK of a sensing matrix is of paramount importance in compressed sensing because it can provide a limit on how sparse a signal has to be for guaranteed unique recovery. For example, if the spark of a certain $M \times N$ sensing matrix F is given to be S , then any signal x with the sparsity K can be uniquely determined from the combinatorial L_0 minimization routine, as long as $K \leq \frac{S}{2}$. That is, $K \leq \frac{S}{2}$ is the sufficient condition for the L_0 norm minimization solution, subject to $y = Fx$ constraint, to return the unique solution. Otherwise, say $K = \frac{S}{2} + 1$ for example, the solution is not unique, which can be easily proved. The L_0 minimization is known as an NP-hard problem since it is combinatorial. The sufficient condition, thus, provides a meaningful benchmark on the required sparsity.

Finding the spark of a matrix is thus desired, but it requires a combinatorial search and hence is an NP-hard problem by itself. In this section, we find the average spark of an ensemble of sensing matrices. For a system with a large block length N , the average spark of an ensemble of sensing matrices is very close to the spark of an individual sensing matrix randomly selected out of the ensemble. That is, it can be shown that the spark of an

individual sensing matrix concentrates around the ensemble averaged spark.

Theorem 1 : The average spark of an ensemble of $\text{GF}(q)$ random sensing matrices, i.e., $\mathcal{C}(N, d_s, d_c, q)$, is given by

$$\text{SPARK}(N, d_s, d_c, q) = \min \left\{ d \in \{2, \dots, N\} \mid d(1 - d_v) \ln(q-1) + (d_s) \ln \frac{N_d}{\binom{N}{d}} \geq 0 \right\} \quad (154)$$

where the variable N_d is a function of the *check* degree d_c and it is given by

$$N_d = \text{Coeff}_d \left(p(x)^{\frac{N}{d_c}} \right) \quad (155)$$

where $p(x) = \sum_{i=0}^{d_c} p_i x^i$, $p_i = \binom{d_c}{i}$, for even i , $p_i = 0$ for odd i , $0 \leq i \leq d_c$, $\text{Coeff}_d(\cdot)$ denotes the coefficient of the term x^d in the expansion of the argument polynomial, and we assume $\frac{N}{d_c}$ is integer.

C. Signal Detection Algorithms

In this section, we aim to discuss how to detect the sparse signal x measured from a sensing matrix F selected randomly out of an ensemble $\mathcal{C}(N, d_s, d_c, q)$ [3].

The sparse signal values can be obtained by solving the following problem:

$$\tilde{x}_t := \arg \max_{\tau_t \in \text{GF}(q)} P(x_t = \tau_t \mid y, C) \quad \text{s.t. } y = Fx \quad (156)$$

where $t = 0, 1, \dots, N-1$ and the symbol “ C ” in the conditioning argument compartment means satisfaction of all the M “check” relations. The function $P(x_t = \tau_t \mid y, C)$ is the posterior distribution, given the observation, and after enforcing the check relations. This posterior distribution is updated for each element of the signal x .

Theorem 2: The a posteriori probability (AP) that the first value, $x_0 = \tau_0 \in \text{GF}(q)$, given the observation y and enforcing the checks (checks should be satisfied), is given by

$$P(x_t = \tau \mid y, C) = \frac{P(x_t = \tau \mid y)}{P(C \mid y)} \times \prod_{p=1}^{d_c} \left[\sum_{x_{t,p}} P(C_{i_p} \mid x_t = \tau_0, x_{t,p}, y) P(x_{0,0} \mid y) \right] \quad (157)$$

We apply the same procedure and obtain the AP result for each element of x . A single round of calculation of the posterior distribution $\{P(x_t = \tau_t \mid y, C) : \tau_t \in \mathcal{X}\}$ for each element, $t = 0, 1, \dots, N-1$, constitutes a single iteration. In a single iteration, therefore, all N different posterior distributions are updated once. We repeat this iteration multiple times. Why do we

need iterations? Why can't we be satisfied with a single iteration? This has to do with our choice on the density, controlled by the two degrees, of the sensing matrix. Choosing a sparse sensing matrix (small degrees) would be desired because when the matrix is sparse, the iterative algorithm works well from the experience we had on the low density parity check codes. In a single iteration, only local information is gathered because of sparse connections. Through iterations, it is hoped that and thus the algorithm is only sub-optimal, the entire information from observation y available via enforcing checking relations prescribed in the sparse matrix can be gathered. An enough number of iterations should be repeated before convergence can be seen on the value of each signal element.

It can be shown that the *check* posterior results, i.e., $\sum_{x_{t,p}} P(C_{i_p} | x_t = \tau_0, x_{t,p}, y) P(x_{t,p} | y)$ in (157), can be obtained from a series of convolution operations of the probability distribution functions of the signal variables connected to the pertinent check C_{i_p} . For example, suppose x_t is connected via its first check to say x_3 and x_6 , then it is the convolution of the two distributions, one for the signal element x_3 and the other for x_6 . The convolution operations can be conveniently done in the frequency domain using FFT and IFFT.

In [3], a couple of ideas on iterations based on identifying the support set detection are also included. One of them is aiming to obtain the posterior distribution of the state S_t of t -th signal element. A state value S_t is binary, 1 for the non-zero value of x_t , and 0 for the zero values. Then, the state posterior either $\Pr\{S_t = 1 | y, C\}$ and $\Pr\{S_t = 0 | y, C\}$ can be updated in each iteration. The log ratio of the posterior probabilities on the state is maintained in each iteration. For the state posterior calculation, the prior information that the signal is sparse is utilized. When the log ratio is greater than 0, then the pertinent state is more like to be 1; otherwise it is zero. At the end of each iteration, we can determine the non-zero states by thresholding the log ratios, collect the indices of non-zero states, and form an estimate of the support set. Once a support set estimate of size K is given, one can then attempt to solve the over-determined problem (by collecting only those columns of matrix F and those elements of x corresponding to the non-zero indices) and find a solution \tilde{x} . If this one is found to satisfy the observation, i.e. $y = F \tilde{x}$, i.e., it is declared to be the solution; then the iteration can stop.

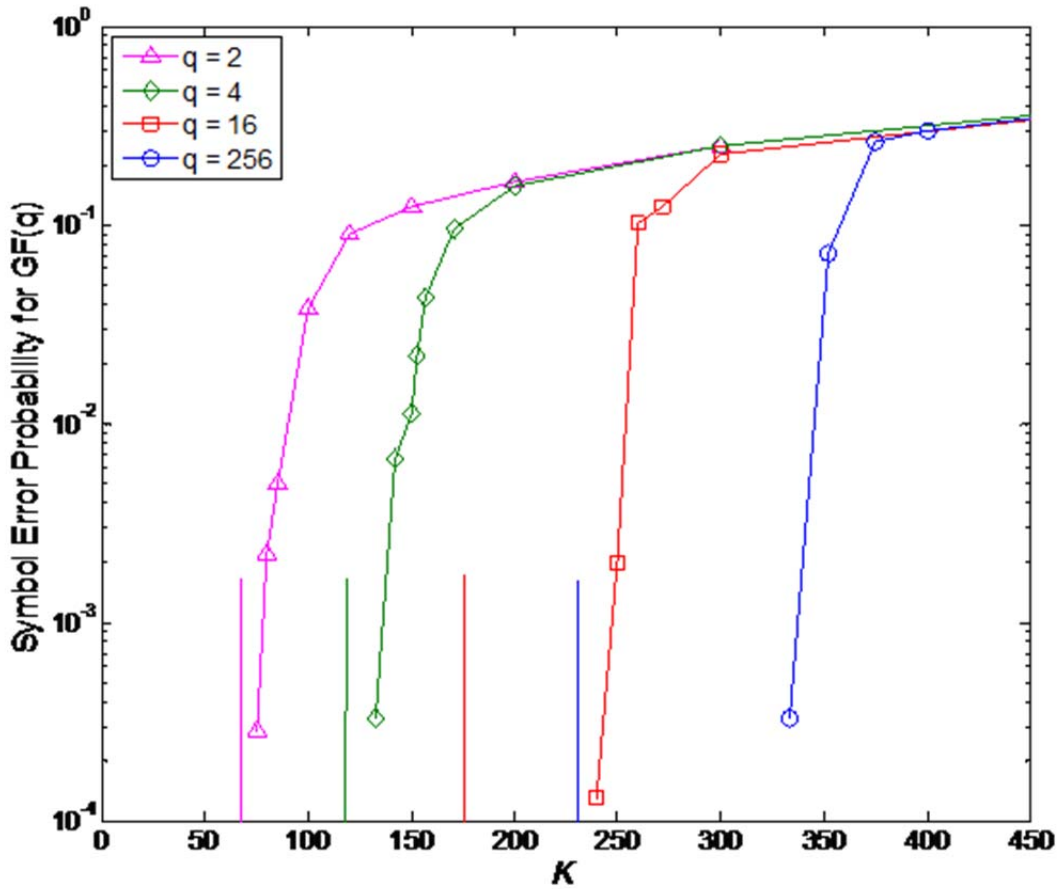


Figure 11 : Simulation results of a $C(N=1200, d_s = 3, d_c = 6)$ code with different field sizes.

D. Simulation Results

Figure 11 shows the Monte Carlo simulation results of our MAP algorithms. The block length is $N = 1200$. The number of observation is $M = 600$; the maximum of number iterations is 20. For each sensing matrix, selected randomly out of $C(N = 1200, d_s = 3, d_c = 6, q)$ ensemble, large enough signal vectors with sparsity K are simulated with an aim to obtain at least 1000 errors for each simulation point. In addition, the same procedure is repeated over for 50 matrix selections, and thus a little bit of averaging is also done for matrix selections within a particular ensemble. Also indicated in Figure 11 are the sparsities obtained from the Gilbert-Varshamov bounds for $q=2, 4, 16$ and 256 and for $(3, 6)$ code. They are indicated as the lines in Figure 11. In addition, Table I shows the sparsity of various rate 1/2 matrices.

Table 1: The Spark S and Relative Spark S/N (inside the parenthesis) obtained from Theorem 1 for $(N=1200, d_v, d_c)$ Ensembles and $GF(q)$. The rate M/N is 1/2.

(d_v, d_c) code	$q=2$	$q=4$	$q=16$	$q=256$
(3,6)	32 (0.027)	64 (0.54)	108 (0.09)	121(0.10)
(4,8)	78 (0.065)	146(0.12)	235 (0.20)	284(0.24)
(5,10)	102 (0.085)	187(0.16)	294 (0.25)	370(0.30)

E. Section Summary

We note that all three measures, the sparsity obtained from GV bounds, the ensemble average sparsity obtained from Theorem 1, and the simulation results of the iterative recovery algorithm, agree to the observation that as the field size q increases, a given sensing matrix of rate $1/2$ can have large spark and thus can be used to detect the signals with a larger sparsity K . Simulation results show that the iterative algorithm can far surpass the predictions made by the average sparks as well as by the GV bounds, which is very interesting, and calls for further study. We also note that as the field size is increased, the compressed sensing bound is $M \geq 2K$ for unique recovery under the Singleton bound. A sensing matrix that satisfies this can be found easily from the random construction, and the iterative recovery algorithm introduced here can be used to even surpass it. We note that the sparsity limits obtained from simulation are much larger. Namely, they are 70, 130, 240, and 325 obtained from simulation. The Singleton bound at $N=1200$ gives a spark of 600 for rate $1/2$ code.

F. Section References

- [1] D. Baron, S. Sarvotham, R. G. Baraniuk, "Bayesian Sensing Via Belief Propagation", *IEEE Sig. Proc.*, vol. 58, no. 1, pp. 269-280, Jan. 2010.
- [2] Stark C. Draper and Sheida Malekpour, "Compressed Sensing over Finite Fields," Proc. of *IEEE ISIT*, Seoul, Korea, 2009.
- [3] Heung-No Lee, *Introduction to Compressed Sensing*, Lecture Note, 2011 Spring Semester, GIST, Korea.
- [4] F. Zhang and H. D. Pfister, "Compressed Sensing and Linear Codes over Real Numbers," arXiv:0806.3243. Accepted for publication in *IEEE Trans. Info. Theory*.
- [5] M. Akcakaya and V. Tarokh, "A frame construction and a universal distortion bound for sparse representations," *IEEE Trans. on Signal Proc.*, vol. 56, pp. 2443-2450, 2008.
- [6] E. Ardestanizadeh, M. Cheraghchi, and A. Shokrollahi, "Bit Precision Analysis for Compressed Sensing," Prof. of *IEEE ISIT*, Seoul, Korea, June 28-July 3, 2009.
- [7] M. Vetterli, P. Marziliano, T. Blu, "Sampling signals with finite rate of innovation," *IEEE Trans. Signal Proc.*, vol.50, no. 6, pp. 1417-1428, June, 2002.
- [8] Heung-No Lee, JinTaek Sung, Suje Lee, "Compressed Sensing over $GF(q)$: Sensing Bounds and Recovery Algorithms," 24th Signal Processing Conference of IEEE Korean Signal Processing Society, vol. 24, no. 1, Sept, 2011.

3. Representation via Correlation vs. via Sparseness

We note that there are multiple ways to represent correlated data. With the sparse representation, we have one more. To name a few, they are Markov chain, statistical correlation via joint distribution, and sparse representation in a certain basis. It would be interesting research direction to see how these different ways of representing correlated signals are related with each other.

Relation between multiple elements x_i s in a vector can be represented as

- Probabilistic model: A signal x has inter-signal correlation. This can be modeled as the joint distribution, given by

$$p(x) = p(x_1, x_2, \dots, x_N)$$

where x_i s are the samples of the signal.

- Knowing a certain group of signal samples x_Σ , $\Sigma \subset \{1, 2, \dots, N\}$, can tell the samples of the other group of samples, x_M , $M \subset \{1, 2, \dots, N\} \setminus \Sigma$, via the conditional probabilities

$$p(x_M | x_\Sigma) = \frac{p(x_M, x_\Sigma)}{p(x_\Sigma)} \quad (158)$$

- We assume $H(x_M | x_\Sigma) < H(x_M, x_\Sigma)$ for correlated data.
- The number of checks required for signal reproduction will be less, for the correlated data.
- Sparse model: a signal x , say an image or a speech signal, is represented as $x = \Psi s$. It is possible to reduce the dimensionality of x via a sparsifying basis Ψ . Note that the energy of x and s are the same. But the number of non-zero samples are different.
- For sparse signals, the number of checks required for signal reproduction is smaller.

Chapter VII. REVIEW OF MATHEMATICAL RESULTS

1. Directional Derivatives, Subgradients, and Subdifferentials

In optimization problems, an evaluation of the objective function at a certain point is compared with that at a neighbor point. Thus, the notion of derivative can serve as a useful tool. But what should we do when the function we deal with is not differentiable at a certain point? Directional derivatives and subgradients are useful for such cases. As long as the function is convex, these tools may serve as an alternate tool for analysis. Please note that the textbook [11] is good for further reading on the materials presented in this subsection.

- (Hyperplane) Hyperplane is a set of vectors x satisfying $a'x = b$ ($a \neq 0$), i.e., $\{x \mid a'x = b\}$

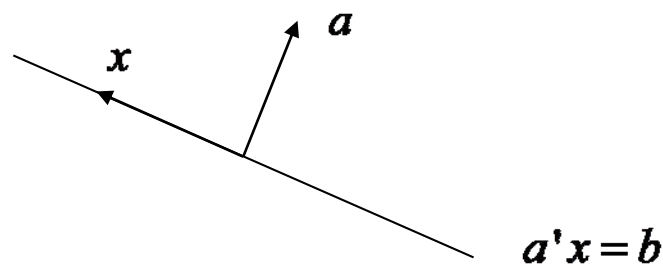


Figure 12: Hyperplane

- (Convex function) Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex U function.
 - The 2nd derivative is non negative for all $x \in \mathbb{R}^n$, $f''(x) \geq 0$, or
 - A convex function satisfies the Jensen's inequality (the average of function is larger than or equal to the function of average.)
- (Subgradient of f) : We say that a vector $d \in \mathbb{R}^n$ is a *subgradient* of f at a point $x \in \mathbb{R}^n$ if

$$f(z) \geq f(x) + (z - x)'d, \quad \forall z \in \mathbb{R}^n. \quad (159)$$

- $(\partial f(x), \text{ the subdifferential of } f)$: The *subdifferential* of f is referred to as the set of all subgradients.
- (Geometrical Illustration of Subdifferential) Let us rewrite (159)

$$f(z) - z'd \geq f(x) - x'd \tag{160}$$

each side of which can be rewritten as $n + 1$ dimensional inner product, i.e.,

$$(-d, 1) \begin{bmatrix} z \\ f(z) \end{bmatrix} \geq (-d, 1) \begin{bmatrix} x \\ f(x) \end{bmatrix} \tag{161}$$

Interpreting (161) geometrically, the set of all z is the supporting hyperplane to the *epigraph* of f at $(x, f(x))$:

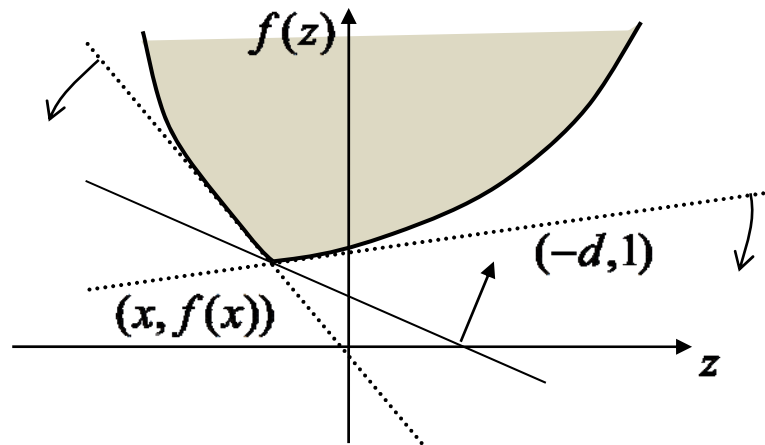


Figure 13: Geometrical illustration of a subgradient of a convex function f . Note that the space is $n + 1$ dimensional.

- (*Directional derivative*) : Subgradient and directional derivative are closely related.

Let $z = x + \alpha y$ for $\forall \alpha > 0, \forall y \in \mathbb{R}^n$. Then, the subgradient can be written as

$$\frac{f(x + \alpha y) - f(x)}{\alpha} \geq y'd$$

We say f is directionally differentiable at x in the direction of y if the limit of the L.H.S. exists as $\alpha \downarrow 0$. (cf, We say f is directionally differentiable at x if differentiable in all directions.)

- (Examples) Express the sub-differential of f as a function of x .

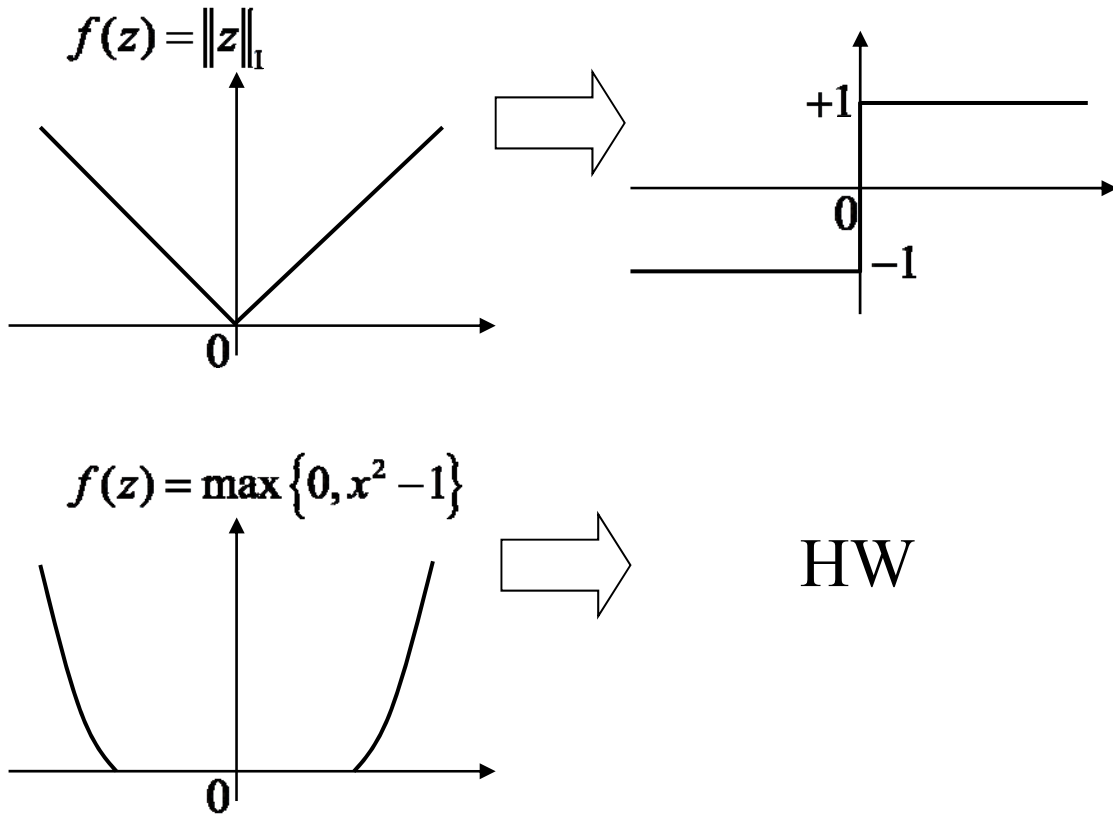


Figure 14: Illustration of subgradients

- As we can see from the examples, a subgradient of a function can be used as a linear approximation to the function. Namely, a subgradient provides an underestimate for a convex function, while an overestimate is provided for a concave function.
- Some properties of the subdifferential $\partial f(x)$ are given as follows:
 - Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex U function.
 - The subdifferential is *non-empty, convex, and compact (closed and bounded)* for all $x \in \mathbb{R}^n$.
 - The function f is differentiable at x with gradient $\nabla f(x)$ iff $\nabla f(x)$ is the unique subgradient at x .
 - Let $f_j : \mathbb{R}^n \rightarrow \mathbb{R}$, $j = 1, 2, \dots, m$, be convex functions and let $f = \sum_{j=1}^m f_j$. Then,

$$\partial f(x) = \sum_{j=1}^m \partial f_j(x) \quad (162)$$

- (ε -Subgradients) There is a notion of approximate subgradient. Let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ be a convex U function. We say that a vector $d \in \mathbb{R}^n$ is an ε subgradient of f at a point $x \in \mathbb{R}^n$ if

$$f(z) \geq f(x) + (z-x)' d - \varepsilon, \quad \forall z \in \mathbb{R}^n. \quad (163)$$

Geometrically, it can be interpreted in the following picture. Note that the change is tangential lines, and thus the change in the subdifferential.

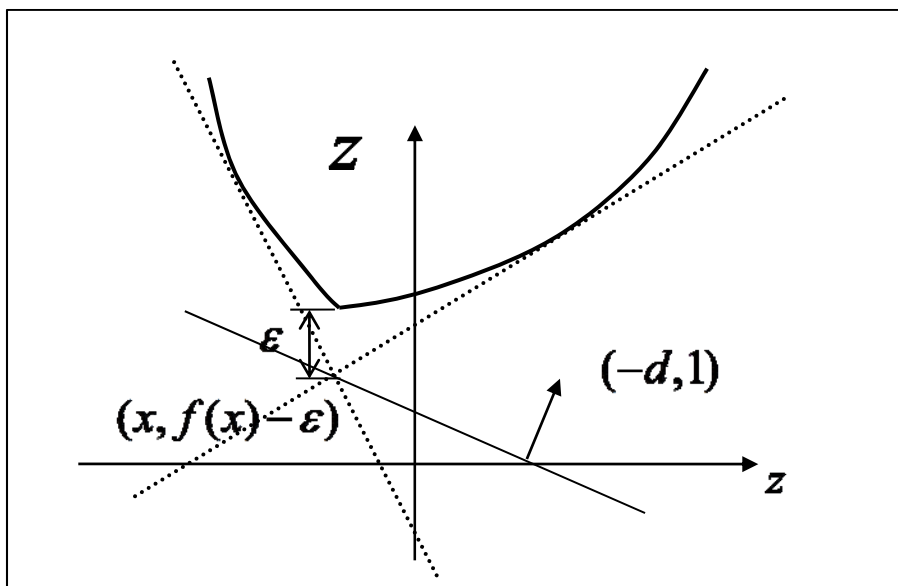


Figure 15: Geometrical illustration of a subgradient of a convex function f . Note that the space is $n + 1$ dimensional.

2. Duality and Convex Optimization

The word dual, an adjective, means twofold, double, or of pertaining two. Wikipedia defines the mathematical meaning of duality in the following way:

In mathematics, a duality translates concepts, theorems, or mathematical structures into other concepts, theorems, or structures, in a one-to-one fashion, often (but not always) by means of an involution operation: If the dual of A is B, then the dual of B is A.

The following are several charts from Convex Optimization [12]. They nicely summarize the Lagrange duality theory.

The standard optimization problem is

$$\begin{aligned} \min f_o(x) \\ \text{s.t. } f_i(x) \leq 0, \quad i = 1, \dots, m \\ h_j(x) = 0, \quad i = 1, \dots, p \end{aligned} \quad (164)$$

where the argument of the functions, $x \in \mathbb{R}^n$, is the variable over which the minimization is defined. The **domain** D of the functions is \mathbb{R}^n unless otherwise stated. More specifically, it is $D = \bigcap_{i=1}^m \text{dom } f_i \cap \bigcap_{i=1}^p \text{dom } h_i$. This set up is general: the objective function $f_o(x)$ does not need to be convex; any of the functions in the constraints need not be convex or concave either.

(Feasible set) A point x is called *feasible* if it satisfies the both sets of inequality and equality constraints. The collection of all feasible solutions is called the *feasible set* of the problem.

(Convex Problem) The optimization problem becomes a convex optimization problem if the **objective function $f_o(x)$ is convex and the feasible set is convex.**

The Lagrange dual theory is a powerful one which works not only for a convex optimization problem but also for the standard optimization problem. It gives us an option for working with the Lagrangian dual optimization problem, which will be precisely defined soon, instead of the original problem in (164) (called the primal problem) which may be difficult to deal with since it is a constrained problem. It is often convenient to deal with the Lagrangian dual optimization problem because the Lagrangian dual function is always concave over the dual variable space. In addition, it has the lower bound property such that the result of the dual optimization is always smaller than or equal to the result of the primal problem. Furthermore, the equality to the optimal result is guaranteed if the primal problem is convex. This motivates the study of Lagrange dual theory.

The **Lagrangian function** L is defined as the following:

$$L(x, \lambda, v) := f_o(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x) \quad (165)$$

It is obtained by adding the weighted sums of the constraint functions to the objective function.

- The lambdas $\{\lambda_i, i = 1, \dots, m\}$ are the Lagrange multipliers for the inequality constraints.
- The v 's $\{v_i, i = 1, \dots, p\}$ are the Lagrange multipliers for the equality constraints.

We note the following facts on the Lagrangian function:

- L is an affine function of the Lagrange variables (λ, v) for a fixed x . That is, we can express the Lagrangian function in the following form, the sum of a scalar plus the inner product, i.e.,

$$\begin{aligned} L(x, \lambda, v) &= f_o(x) + \begin{pmatrix} f_1(x) & \cdots & f_m(x) & h_1(x) & \cdots & h_p(x) \end{pmatrix} \begin{pmatrix} \lambda \\ v \end{pmatrix} \\ &= b(x) + A(x) \begin{pmatrix} \lambda \\ v \end{pmatrix} \end{aligned} \quad (166)$$

- The second term is negative if $\lambda \geq 0$ since each term $\lambda_i f_i(x) \leq 0$ is non positive for a feasible x .
- The third term is zero as long since each term $v_i h_i(x) = 0$ for a feasible x .

The **Lagrange dual function** $g(\lambda, v)$ is defined as the infimum of the Lagrange function $L(x, \lambda, v)$ over all x for a fixed set of Lagrange multipliers (λ, v) :

$$g(\lambda, v) := \inf_{x \in D} L(x, \lambda, v). \quad (167)$$

From this definition, we can make the following important observations:

1. $g(\lambda, v)$ is concave function of $(\lambda \in \mathbb{R}^m, v \in \mathbb{R}^p)$.
2. $g(\lambda, v)$ for $\lambda \geq 0$ is a lower-bound to p^* , the result of the primal optimization.

Since these results are central to the Lagrange dual theory, let us prove them here.

Proof of concavity. The direction is to use the Jensen's inequality ($g(\mathbb{E}\lambda) \geq \mathbb{E}g(\lambda)$).

Without loss of generality, we omit the part involving v and show the concavity of $g(\lambda)$.

Let $\theta \in [0, 1]$. We have

$$\begin{aligned}
g(\theta\lambda_1 + (1-\theta)\lambda_2) &= \inf_x \{b(x) + A(x)(\theta\lambda_1 + (1-\theta)\lambda_2)\} \\
&= \inf_x \{\theta b(x) + (1-\theta)b(x) + A(x)(\theta\lambda_1 + (1-\theta)\lambda_2)\} \\
&= \inf_x \{\theta[b(x) + A(x)\lambda_1] + (1-\theta)[b(x) + A(x)\lambda_2]\} \\
&\geq \inf_x \{\theta[b(x) + A(x)\lambda_1]\} + \inf_x \{(1-\theta)[b(x) + A(x)\lambda_2]\} \\
&= \theta \inf_x \{[b(x) + A(x)\lambda_1]\} + (1-\theta) \inf_x \{[b(x) + A(x)\lambda_2]\} \\
&= \theta g(\lambda_1) + (1-\theta)g(\lambda_2)
\end{aligned} \tag{168}$$

The one inequality in (168) is valid for all $\lambda \in \mathbb{R}^m$. To show this, let us consider two points x_1 and x_2 without loss of generality. There are only two cases we shall consider. The first case is when $a_1(x_1) > a_1(x_2)$ and $a_2(x_1) > a_2(x_2)$. Then, the L.H.S. is

$$\inf_{x_1, x_2} \{\theta a_1(x_1) + (1-\theta)a_2(x_1), \theta a_1(x_2) + (1-\theta)a_2(x_2)\} = \theta a_1(x_2) + (1-\theta)a_2(x_2),$$

while the R.H.S. is

$$\inf_x \{\theta a_1(x_1), \theta a_1(x_2)\} + \inf_x \{(1-\theta)a_2(x_1), (1-\theta)a_2(x_2)\} = \theta a_1(x_2) + (1-\theta)a_2(x_2).$$

Thus, we have the equality in this case.

The second case is when $a_1(x_1) > a_1(x_2)$ but $a_2(x_1) < a_2(x_2)$. In addition, we let $\theta(a_1(x_1) - a_1(x_2)) < (1-\theta)(a_2(x_2) - a_2(x_1))$. Then, the L.H.S. is

$$\inf_{x_1, x_2} \{\theta a_1(x_1) + (1-\theta)a_2(x_1), \theta a_1(x_2) + (1-\theta)a_2(x_2)\} = \theta a_1(x_1) + (1-\theta)a_2(x_1),$$

while the R.H.S. is

$$\inf_x \{\theta a_1(x_1), \theta a_1(x_2)\} + \inf_x \{(1-\theta)a_2(x_1), (1-\theta)a_2(x_2)\} = \theta a_1(x_2) + (1-\theta)a_2(x_1).$$

Now let us compare the two results. Since $a_1(x_1) > a_1(x_2)$, the L.H.S. is strictly larger than the R.H.S. Thus, we have the inequality in this case. Furthermore, let $\theta(a_1(x_1) - a_1(x_2)) > (1-\theta)(a_2(x_2) - a_2(x_1))$. Then, the L.H.S. is

$$\inf_{x_1, x_2} \{\theta a_1(x_1) + (1-\theta)a_2(x_1), \theta a_1(x_2) + (1-\theta)a_2(x_2)\} = \theta a_1(x_2) + (1-\theta)a_2(x_2),$$

while the R.H.S. stays the same

$$\inf_x \{\theta a_1(x_1), \theta a_1(x_2)\} + \inf_x \{(1-\theta)a_2(x_1), (1-\theta)a_2(x_2)\} = \theta a_1(x_2) + (1-\theta)a_2(x_1)$$

Now let us compare the two results. Since we have assumed $a_2(x_1) < a_2(x_2)$, the L.H.S is again strictly larger than the R.H.S. Thus, we have the inequality in this case.
Q.E.D.

Proof on the lower-bounding part. See the chart below.

From the definition, we have

$$\begin{aligned} g(\lambda, v) &:= \inf_{x \in D} L(x, \lambda, v) \\ &= \inf_{x \in D} \left\{ f_o(x) + \sum_{i=1}^m \lambda_i f_i(x) + \sum_{i=1}^p v_i h_i(x) \right\} \end{aligned}$$

Let \tilde{x} is a feasible point and $\lambda \succeq 0$. Then, we have the following relations

$$f_o(\tilde{x}) \geq L(\tilde{x}, \lambda, v) \geq \inf_{x \in D} L(x, \lambda, v) = g(\lambda, v). \quad (169)$$

Now minimizing the objective fuction over all feasible points, we have the result

$$p^* := f_o(x^*) \geq g(\lambda, v). \quad (170)$$

Q.E.D.

The following is a graphical illustration that the Lagrange dual function is concave.

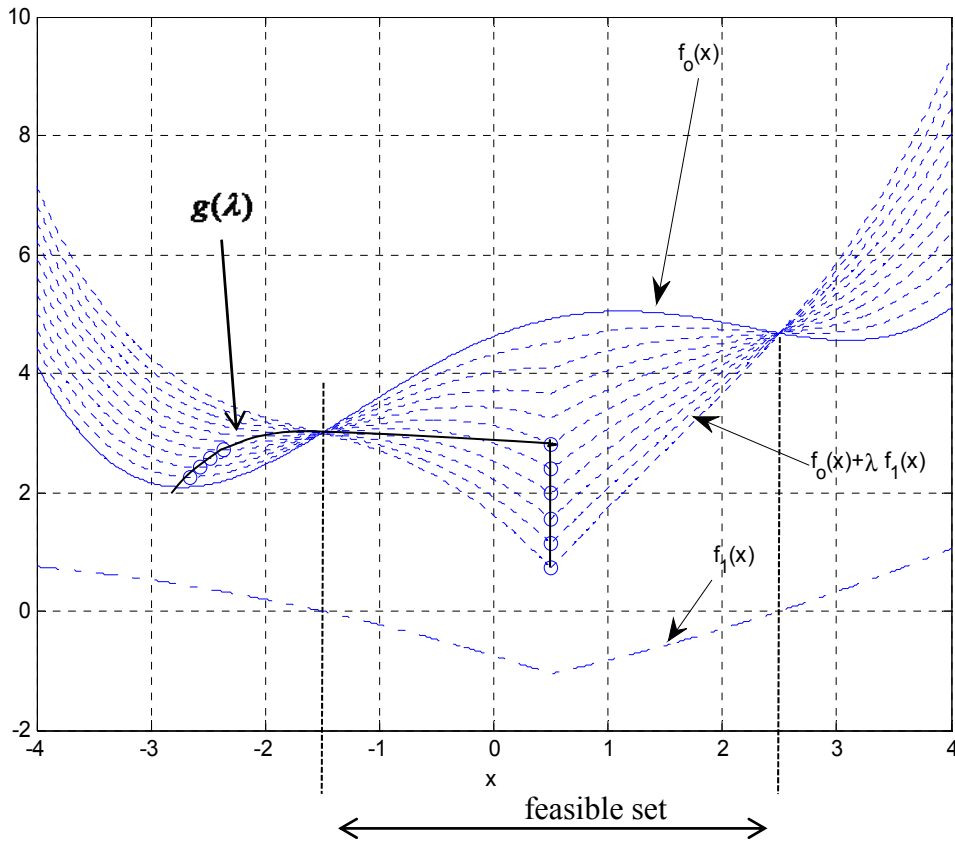


Figure 16: The trajectory of $L(x, \lambda)$ with λ is varied from 0.4 to 4.

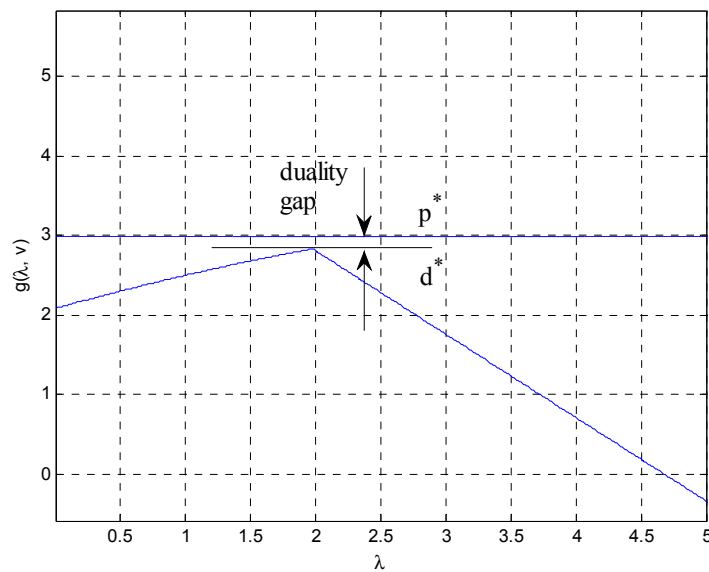


Figure 17: Illustration of the duality gap

The dual problem

Lagrange dual problem

$$\begin{array}{ll} \text{maximize} & g(\lambda, \nu) \\ \text{subject to} & \lambda \succeq 0 \end{array}$$

- finds best lower bound on p^* , obtained from Lagrange dual function
- a convex optimization problem; optimal value denoted d^*
- λ, ν are dual feasible if $\lambda \succeq 0, (\lambda, \nu) \in \text{dom } g$
- often simplified by making implicit constraint $(\lambda, \nu) \in \text{dom } g$ explicit

example: standard form LP and its dual (page 5–5)

$$\begin{array}{ll} \text{minimize} & c^T x \\ \text{subject to} & Ax = b \\ & x \succeq 0 \end{array} \qquad \begin{array}{ll} \text{maximize} & -b^T \nu \\ \text{subject to} & A^T \nu + c \succeq 0 \end{array}$$

Duality

5–9

Since, we now know that, the Lagrange dual function is concave and provides a lower bound to p^* for non negative λ , the dual problem makes sense. Namely, the maximization $g(\lambda, \nu)$ will narrow the gap between p^* and $g(\lambda, \nu)$. Along the same direction of discussion, the following results make sense.

Let $d^* := \max g(\lambda, \nu)$ s.t. $\lambda \succeq 0$ for this discussion. Then,

- Weak duality, $d^* \leq p^*$, holds always even for nonconvex problems. Thus, Lagrange dual optimization can be used to provide a non trivial lower bound to the optimal value for nonconvex and difficult problems.
- Strong duality, $d^* = p^*$, usually holds for convex problems. There are some conditions on the constraints that guarantee strong duality in convex problems.

(Convex optimization problem) A convex optimization problem must take the following form

$$\begin{array}{ll} \min & f_o(x) \\ \text{s.t.} & f_i(x) \leq 0, \quad i = 1, \dots, m \\ & Ax = b, \quad A \text{ is a } [p \times n] \text{ matrix} \end{array} \quad (171)$$

where the objective and the functions of the inequality constraints, i.e., f_i s, $i = 0, 1, \dots, m$, are convex functions and the inequality constraint functions h_i s are affine functions. In other words,

- the objective function is convex and
- the feasible set is convex as well

for an optimization problem to be convex. The first one should be obvious while the second one needs a little discussion.

The set of points $\{x \in f_i \mid f_i(x) \leq 0\}$ is called a sublevel set. A sublevel set of a function $f_i : \mathbb{R}^n \rightarrow \mathbb{R}$ is convex if the function is convex. The intersection of convex sets is a convex set as well.

(Strong duality holds if strictly feasible) Now we go back to the discussion of the constraint qualification which guarantees the strong duality. Namely, for a convex problem, the strong duality holds if it is strictly feasible. Strict feasibility is meant by the existence of a feasible point $x \in D$ such that

$$f_i(x) < 0, \quad i = 1, 2, \dots, m, \quad Ax = b. \quad (172)$$

Notation for there exists a feasible point is $x \in \text{int } D$. When some of the constraint functions are affine, say $i = 1, \dots, k$, then the condition (172) can be relaxed, such that $f_i(x) \leq 0$, the equality part, is o.k. for $i = 1, \dots, k$.

(Strong duality holds if (x^*, λ^*) is a saddle point for $L(x, \lambda)$) Consider Figure 16 again, and now observe that the primal optimal value and the dual optimal value can be compared as follows. For this discussion, we omit the contribution of the equality constraints without loss of generality, and the Lagrange function with one inequality constraint function $f_1(x) \leq 0$:

$$L(x, \lambda) := f_0(x) + \lambda f_1(x) \quad (173)$$

The figure explains how the Lagrange dual function $g(\lambda) := \inf_x L(x, \lambda)$ behaves as λ is varied from 0 to a larger value. Figure 17 shows the concavity of $g(\lambda)$ as a function of λ and the duality gap $p^* - d^*$. On the one hand, the primal optimal value p^* was obtained by visually inspecting the objective function $f_0(x)$ within the feasible set: one can move up and down the horizontal line, and read off the primal optimal value p^* and its primal optimal point x^* inside the feasible set. On the other hand, the dual optimal value d^* was obtained by maximizing the concave function $g(\lambda)$ with the constraint of non negative lambda, $\lambda \geq 0$, i.e.,

$$d^* := \sup_{\lambda \geq 0} \inf_{x \in D} L(x, \lambda) \quad (174)$$

Now let us consider obtaining the primal optimal value. We note that it can be obtained by taking the minimum of the maximum of $L(x, \lambda)$, i.e.,

$$p^* := \inf_{x \in D} \sup_{\lambda \geq 0} L(x, \lambda) \quad (175)$$

At each point x , there are infinitely many $L(x, \lambda)$ and we take the supremum first over λ . On the set of non-feasible points, $L(x, \lambda)$ tends to ∞ since $f_1 > 0$, i.e., $\sup_{\lambda \geq 0} L(x, \lambda) = \infty$ as $\lambda \rightarrow \infty$. On the set of feasible points, $\sup_{\lambda \geq 0} L(x, \lambda) = f_o(x)$ as $\lambda \rightarrow 0$. Second, we take the infimum of the supremums with respect to $x \in D$.

Thus, the weak duality is simply

$$d^* := \sup_{\lambda \geq 0} \inf_{x \in D} L(x, \lambda) \leq p^* := \inf_{x \in D} \sup_{\lambda \geq 0} L(x, \lambda). \quad (176)$$

The equality is obtained if the sup and inf can be switched without affecting the result. This happens when the function $L(x, \lambda)$ satisfies the strong max-min property or the saddle point property. For the proof on this, please refer to Theorem 1 and Theorem 2 of Lasdon [16] on page 84 and 85 respectively.

Definition. (A slack variable s) A variable is called a *slack* variable when it is used to convert an inequality into equality. For example, the inequality, $x + y \leq 20$, can be made to become $x + y + s = 20$ and then, s is a slack variable.

Complementary slackness

assume strong duality holds, x^* is primal optimal, (λ^*, ν^*) is dual optimal

$$\begin{aligned} f_0(x^*) = g(\lambda^*, \nu^*) &= \inf_x \left(f_0(x) + \sum_{i=1}^m \lambda_i^* f_i(x) + \sum_{i=1}^p \nu_i^* h_i(x) \right) \\ &\leq f_0(x^*) + \sum_{i=1}^m \lambda_i^* f_i(x^*) + \sum_{i=1}^p \nu_i^* h_i(x^*) \\ &\leq f_0(x^*) \end{aligned}$$

hence, the two inequalities hold with equality

- x^* minimizes $L(x, \lambda^*, \nu^*)$
- $\lambda_i^* f_i(x^*) = 0$ for $i = 1, \dots, m$ (known as complementary slackness):

$$\lambda_i^* > 0 \implies f_i(x^*) = 0, \quad f_i(x^*) < 0 \implies \lambda_i^* = 0$$

Karush-Kuhn-Tucker (KKT) conditions

the following four conditions are called KKT conditions (for a problem with differentiable f_i, h_i):

1. primal constraints: $f_i(x) \leq 0, i = 1, \dots, m, h_i(x) = 0, i = 1, \dots, p$
2. dual constraints: $\lambda \succeq 0$
3. complementary slackness: $\lambda_i f_i(x) = 0, i = 1, \dots, m$
4. gradient of Lagrangian with respect to x vanishes:

$$\nabla f_0(x) + \sum_{i=1}^m \lambda_i \nabla f_i(x) + \sum_{i=1}^p \nu_i \nabla h_i(x) = 0$$

from page 5–17: if strong duality holds and x, λ, ν are optimal, then they must satisfy the KKT conditions

Duality

5–18

Note that the KKT conditions here are necessary conditions such that if the strong duality holds then the optimal parameters must satisfy the KKT conditions.

KKT conditions for convex problem

if $\tilde{x}, \tilde{\lambda}, \tilde{\nu}$ satisfy KKT for a convex problem, then they are optimal:

- from complementary slackness: $f_0(\tilde{x}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$
- from 4th condition (and convexity): $g(\tilde{\lambda}, \tilde{\nu}) = L(\tilde{x}, \tilde{\lambda}, \tilde{\nu})$

hence, $f_0(\tilde{x}) = g(\tilde{\lambda}, \tilde{\nu})$

if **Slater's condition** is satisfied:

x is optimal if and only if there exist λ, ν that satisfy KKT conditions

- recall that Slater implies strong duality, and dual optimum is attained
- generalizes optimality condition $\nabla f_0(x) = 0$ for unconstrained problem

Duality

5–19

For convex problem, those parameters that satisfy the KKT conditions are the optimal parameters.

example: water-filling (assume $\alpha_i > 0$)

$$\begin{aligned} &\text{minimize} && -\sum_{i=1}^n \log(x_i + \alpha_i) \\ &\text{subject to} && x \succeq 0, \quad \mathbf{1}^T x = 1 \end{aligned}$$

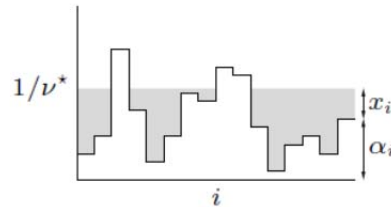
x is optimal iff $x \succeq 0$, $\mathbf{1}^T x = 1$, and there exist $\lambda \in \mathbf{R}^n$, $\nu \in \mathbf{R}$ such that

$$\lambda \succeq 0, \quad \lambda_i x_i = 0, \quad \frac{1}{x_i + \alpha_i} + \lambda_i = \nu$$

- if $\nu < 1/\alpha_i$: $\lambda_i = 0$ and $x_i = 1/\nu - \alpha_i$
- if $\nu \geq 1/\alpha_i$: $\lambda_i = \nu - 1/\alpha_i$ and $x_i = 0$
- determine ν from $\mathbf{1}^T x = \sum_{i=1}^n \max\{0, 1/\nu - \alpha_i\} = 1$

interpretation

- n patches; level of patch i is at height α_i
- flood area with unit amount of water
- resulting level is $1/\nu^*$



3. Some Useful Linear Algebra Results

A. Some Facts on the Matrix Norm

Let us consider the following equalities:

Let A be an $M \times K$ matrix, $M \geq K$. Note the following equalities:

$$\begin{aligned} \max_{x \neq 0} \frac{\|Ax\|_2^2}{\|x\|_2^2} &= \max_{x \neq 0} \frac{x^T A^T A x}{\|x\|_2^2} \\ &= \frac{x^T \lambda_{\max} x}{\|x\|_2^2} \quad (x \text{ the eigenvector for } \lambda_{\max}) \\ &= \lambda_{\max} \end{aligned} \quad (177)$$

Let λ_{\max} be the largest eigenvalue of the non-negative definite $K \times K$ matrix $A^T A$.

The maximum gain

$$\max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\max}} \quad (178)$$

is called the matrix norm or spectral norm of A .

Then, the matrix norm of $A^T A$ is λ_{\max} .

Similarly, let λ_{\min} be the minimum eigenvalue of the non-negative definite matrix. Then, the minimum gain is

$$\min_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\lambda_{\min}}. \quad (179)$$

B. Other useful matrix norm properties

Suppose $A = A^T$ be $K \times K$ matrix. Then, $A = U \Sigma U^T$ is the eigenvalue decomposition where U is the unitary matrix, and the diagonal matrix Σ 's entries are the sorted eigenvalues, such as $\lambda_1 \geq \dots \geq \lambda_K$. The following are true for any x :

- $x^T A x \leq \lambda_1 x^T x$
- $x^T A x \geq \lambda_K x^T x$
- For any x , $\|Ax\| \leq \|A\| \|x\|$.
- $\|A + B\| \leq \|A\| + \|B\|$ (Triangle inequality)
- $\|A\| = 0 \Leftrightarrow A = 0$

- $\|AB\| \leq \|A\|\|B\|$

Chapter VIII. REFERENCES

- [1] Claude E. Shannon, "Communication in the Presence of Noise," *Proceeding of the I.R.E.*, vol. 37, pp. 10-21, January, 1949.
- [2] J. M. Whittaker, "Interpolatory Function Theory," Cambridge Tracts in Mathematics and Mathematical Physics, No. 33, Cambridge University Press, Chapt. IV; 1935.
- [3] David L. Donoho, "Compressed Sensing," *IEEE Trans. Information Theory*, vol. 52, no. 4, pp. 1289-1306, Apr. 2006.
- [4] David L. Donoho and Jared Tanner, "Precise Undersampling Theorems," *Proceedings of the IEEE*, vol. 98, pp. 913-924, May, 2010.
- [5] Richard Baraniuk, "Lecture Notes: Compressive Sensing," *IEEE Signal Processing Magazine*, p. 118-121, July, 2007.
- [6] Justin Romberg, "Imaging via compressive sampling," *IEEE Signal Processing Magazine*, 25(2), pp. 14 - 20, March 2008.
- [7] Ashikhmin, A. and Calderbank, A.R., "Grassmannian Packings from Operator Reed-Muller Codes," *IEEE Trans. Information Theory*, vol. 56, Issue: 11, pp. 5689-5714, 2010.
- [8] Emmanuel Candès and Terence Tao, Near optimal signal recovery from random projections: Universal encoding strategies? (*IEEE Trans. on Information Theory*, 52(12), pp. 5406 - 5425, December 2006)
- [9] Emmanuel Candès and Terence Tao, "Decoding by linear programming," *IEEE Trans. on Information Theory*, 51(12), pp. 4203 - 4215, December 2005.
- [10] Emmanuel Candès, Justin Romberg, and Terence Tao, [Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information](#). (*IEEE Trans. on Information Theory*, 52(2) pp. 489 - 509, February 2006)
- [11] Dimitri P. Bertsekas, A. Nedic and A.E. Ozdaglar, *Convex Analysis and Optimization*, Athena Scientific, 2003.
- [12] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004. D.
- [13] Donoho and X. Huo, "Uncertainty Principles and Ideal Atomic Decomposition," *IEEE Trans. on Info. Theory*, vol.47, no.7, Nov. 2001.
- [14] M. Elad and A. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Info. Theory*, vol. 48, no. 9, Sept. 2002.
- [15] J. Romberg, E. Candes, and T. Tao, "Robust Uncertainty Principles and Optimally Sparse Decompositions," Presentation Downloaded from http://www.ipam.ucla.edu/publications/mgaws1/mgaws1_5184.pdf.
- [16] Leon S. Lasdon, *Optimization Theory for Large Systems*, Dover publication, 2002.
- [17] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders, "Atomic Decomposition by Basis Pursuit," Vol. 43, no.1, pp. 129-159, *SIAM Review*, 2001.
- [18] D.L. Donoho and Y. Tsaig, "Fast Solution of l-1 Norm Minimization Problems When the Solution May Be Sparse," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789-4812, Nov. 2008.
- [19] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Trans. Inf. Theory*, vol. 50, no. 10, pp. 2231-2242, Oct. 2004.
- [20] J. A. Tropp, "Just relax: Convex programming methods for subset selection and sparse approximation," *IEEE Trans. Inf. Theory*, vol. 51, no. 3, pp. 1030-1051, Mar.

- 2006.
- [21] M. R. Osborne, B. Presnell, and B. A. Turlach, “A new approach to variable selection in least squares problems,” *IMA J. Numer. Anal.*, vol. 20, pp. 389–403, 2000.
- [22] M. R. Osborne, B. Presnell, and B. A. Turlach, “On the LASSO and its dual,” *J. Comput. Graph. Stat.*, vol. 9, pp. 319–337, 2000.
- [23] R. Tibshirani, “Regression shrinkage and selection via the lasso,” *J. Roy. Statist. Soc.*, vol. 58, no. 1, pp. 267–288, 1996.
- [24] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” vol.106, no.45, *Proceedings of National Academy of Science*, Nov. 10, 2009.
- [25] H. Rauhut, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, volume 9 of *Radon Series Comp. Appl. Math.*, pages 1-92. deGruyter, 2010. <http://rauhut.ins.uni-bonn.de/research/pub.php?list=rauhut>.
- [26] Y.C. Chen, et. al, “Terahertz Pulsed Spectroscopic Imaging Using Optimized Binary Masks,” *Applied Physics Letters* 95, 231112, 2009.
- [27] W.L. Chan, M.L. Moravec, R.G. Baraniuk, and D. Mittleman, “Terahertz imaging with compressed sensing and phase retrieval,” *Optics Letters*, vol. 33, no. 9, May 1, 2008.
- [28] Michael Elad, *Sparse and Redundant Representations: From theory to applications in signal processing and image processing*, Springer Science, New York, 2010.
- [29] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. submitted to *IEEE Trans. Inform. Theory*, February 2004.
- [30] E. J. Candès, J. Romberg and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59 1207-1223.
- [31] Lawson and Hanson, "Solving Least Squares Problems", Prentice-Hall, 1974, Chapter 23, p. 161.
- [32] Donoho and M. Elad, “Optimally Sparse Representation in General Dictionaries via L_1 Minimization, ...
- [33] E. J. Candès, “Compressed Sampling,” *Proc. of the Int. Congress of Mathematicians*, Madrid, Spain, 2006.
- [34] S.J. Park and Heung-No Lee, “On the Derivation of RIP for Random Gaussian Matrices and Binary Sparse Signals,” *Proc. of International Conference on ICT Convergence 2011 (ICTC 2011)*, Seoul, Korea, Sept. 28-30, 2011.
- [35] G. Tang and A. Nehorai, “Performance Analysis for Sparse Support Recovery,” *IEEE Trans. Info. Theory*, vol. 56, no. 3, pp. 1383-1399, March, 2010.