Universidade Federal do Rio de Janeiro

Compressive Sensing

Claudio Mayrink Verdun

Rio de Janeiro

Dezembro de 2016

Universidade Federal do Rio de Janeiro

Compressive Sensing

Claudio Mayrink Verdun

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática Aplicada, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática Aplicada.

Orientador: Prof. Cesar Javier Niche Mazzeo.
Co-orientador: Prof. Bernardo Freitas Paulo da Costa.
Co-orientador: Prof. Eduardo Antonio Barros da Silva.

Rio de Janeiro
Dezembro de 2016

Universidade Federal do Rio de Janeiro

Compressive Sensing

Claudio Mayrink Verdun

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Matemática Aplicada, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática Aplicada.

Aprovada por:

Prof. Cesar Javier Niche Mazzeo, Ph.D.

Prof. Bernardo Freitas Paulo da Costa, Ph.D.

Prof. Eduardo Antonio Barros da Silva, Ph.D.

Prof. Amit Bhaya, Ph.D.

Prof. Carlos Tomei, Ph.D.

Prof. Fábio Antonio Tavares Ramos, Ph.D.

Prof. Roberto Imbuzeiro M. F. de Oliveira, Ph.D.

Rio de Janeiro
Dezembro de 2016

Verdun, Claudio Mayrink
V487c Compressive Sensing / Claudio Mayrink Verdun. –
Rio de Janeiro: UFRJ/ IM, 2016.
202 f. 29,7 cm.
Orientador: César Javier Niche Mazzeo.
Coorientadores: Bernardo Freitas Paulo da Costa,
Eduardo Ântonio Barros da Silva.
Dissertação (Mestrado) — Universidade Federal do
Rio de Janeiro, Instituto de Matemática, Programa de
Pós-Graduação em Matemática, 2016.

1. Compressive Sensing. 2. Sparse Modelling.
3. Signal Processing. 4. Applied Harmonic Analysis.
5. Inverse Problems I. Mazzeo, César Javier Niche,
orient. II. Costa, Bernardo Freitas Paulo da, coorient.
Silva, Eduardo Ântonio Barros da, coorient. III. Título

Del rigor en la ciencia

En aquel Imperio, el Arte de la Cartografía logró tal Perfección que el Mapa de una sola Provincia ocupaba toda una Ciudad, y el Mapa del Imperio, toda una Provincia. Con el tiempo, estos Mapas Desmesurados no satisficieron y los Colegios de Cartógrafos levantaron un Mapa del Imperio, que tenía el Tamaño del Imperio y coincidía puntualmente con él. Menos Adictas al Estudio de la Cartografía, las Generaciones Siguientes entendieron que ese dilatado Mapa era Inútil y no sin Impiedad lo entregaron a las Inclemencias del Sol y los Inviernos. En los Desiertos del Oeste perduran despedazadas Ruinas del Mapa, habitadas por Animales y por Mendigos; en todo el País no hay otra reliquia de las Disciplinas Geográficas.

Suárez Miranda: Viajes de varones prudentes
Libro Cuarto, cap. XLV, Lérida, 1658. In [Borges '60].

Acknowledgments

This is the only part of my dissertation written in my native language. I hope that all who are cited here may read clearly what follows. Se você é daquelas pessoas que só liga para a matemática e acha que ela não é feita por seres humanos que sorriem, choram, pensam coisas profundas mas também fazem coisas triviais, pule esta parte e vá diretamente para o Capítulo 1. Ali, vetores esparsos e medições compressivas te esperam.

Eu tentei ser preciso e não esquecer ninguém. Se eu esqueci, possivelmente foi de propósito, se não foi, peço desculpas, embora nunca irá transparecer qual foi o motivo do esquecimento. Além disso, veremos ao longo dessa dissertação que redundância é algo fundamental, então começaremos por aqui, citando algumas pessoas em mais de um lugar e por motivos diferentes.

Esta dissertação é dedicada a duas santíssimas trindades: A primeira delas, Felipe Acker, o pai, Fábio Ramos, o filho, e Luiz Wagner Biscainho, o espírito santo. A segunda, Roberto Velho, o pai, Filipe Goulart, o filho, e Hugo Carvalho, o espírito santo. Essas seis pessoas contribuíram de maneira não trivial para minha formação e introduziram muitas não-linearidades na minha vida ao longo dos últimos anos. A quantidade de coisas que elas me ensinaram e todas as vezes que fui ajudado por elas, mesmo que as três horas da manhã, as vezes desesperado, e com pedidos nada convencionais não pode ser descrito aqui em poucas palavras. Parte substancial da minha formação humana e científica deve-se aos laços estreitos e inúmeras conversas que mantenho com esses seis indivíduos bastante singulares.

Começo agradecendo três professores que tive na escola e que foram fonte de inspiração para a escolha da carreira científica: Andre Fontenele, José Carlos de Medeiros e José Alexandre Tostes Lopes (*in memoriam*). Aproveitei muitos os três anos de contato com eles, as conversas que iam de geometria euclidiana à suma teológica de São Tomás de Aquino e todos os ensinamentos sobre porque o mundo era muito maior que o IME/ITA e porque haviam opções melhores sobre o que fazer da vida.

Agradeço ao Vinícius Gripp por ter me mostrado a sociedade secreta que é a matemática aplicada da UFRJ. Se eu não tivesse passado mal um certo dia na escola e não tivesse precisado ir embora, não teria conhecido o Vinícius, todo o seu entusiasmo e tudo o que se seguiu a partir daí. Além disso, agradeço aos seis bandeirantes da matemática aplicada: Enio Hayashi, Pedro Maia, Rafael March, Roberto Velho, Rodrigo Targino e Yuri Saporito. Esses veteranos, que acabaram virando amigos, continuam por aí desbravando a mata fechada e eu sou muito feliz em aproveitar muitos dos conselhos por eles oferecidos.

Aos vários mestres que tive nos últimos sete anos. Cada lição, cada raciocínio, cada forma de fazer as coisas dentro e fora da sala de aula me mostrou um mundo de ideias interessantes que eu ainda vou utilizar muito nas minhas pesquisas e atividades didáticas. São eles Adan Corcho, Alexander Arbieto, Alexandra Schmidt, Amit Bhaya, Átila Pantaleão, Bernardo da Costa, Bruno Scárdua, César Niche, Didier Pilod, Eduardo Barros Silva, Fábio Ramos, Felipe Acker, Flávio Dickstein, Heudson Mirandola, Luca Moriconi, Marco Aurélio Cabral, Marcus Venícius Pinto, Luiz Wagner Biscainho, Nilson Bernardes, Paulo Diniz, Sergio Lima Netto, Severino Collier, Umberto Hryniewicz, Wagner Coelho, Wallace Martins e Wladimir Neves. Agradeço também à vários outros professores que conheci na universidade e que me ensinaram como *não* fazer as coisas, como *não* trabalhar, como *não* fazer pesquisa e dar aula, como *não* tratar os alunos, etc.

Aos muitos amigos e colegas que fiz aplicada e na UFRJ. Alguns se tornaram muito próximos, outros nem tanto, mas todos eles compartilharam momentos felizes e ajudaram a superar momentos tristes na universidade. Também me proporcionaram conversas fantásticas e muitas vezes me mostraram que eu estava errado, por mais que que fosse difícil dar o braço a torcer. São eles: Aloizio Macedo, Arthur Mitrano, Bruno Almeida, Bruno Oliveira, Bruno Santiago, Camilla Codeço, Carlos Lechner, Carlos Zanini, Cecília Mondaini, Claudio Heine, Daniel Soares, Danilo Barros, Davi Obata, Felipe Pagginelli, Felipe Senra, Flávio Ávila, Franklin Maia, Gabriel Castor, Gabriel Sanfins, Gabriel Vidal, Gabriela Lewenfus, Gabriella Radke, Ian Martins, Jhonatas Afradique, Leandro Machado, Leonardo Assumpção, Lucas Barata, Lucas Manoel, Lucas Stolerman, Marcelo Ribeiro, Marco Aurélio Soares, Milton Nogueira, Nicolau Sarquis, Paulo Cardozo, Pedro Fonini, Pedro Schmidt, Philipe de Fabritiis, Rafael Teixeira, Raphael Lourenço, Reinaldo de Melo e Souza, Rogério Lourenço, Thiago Degenring, Tiago Domingues, Tiago Vital, Washington Reis, Victor Coll e Victor Corcino.

Aos top's, a galera realmente truuu, que se tornaram super amigos e que sei poder contar para qualquer coisa em qualquer hora e lugar. Alguns deles estão longe mas sei que caso eu precise, eles não medirão esforços para estar perto. Obrigado Adriano Côrtes, Bruno Braga, Bruno Santiago, Cláudio Gustavo Lima, Cristiane Krug, Douglas Picciani, Guilherme Sales, Luís Felipe Velloso e Rafael Ribeiro por serem quem vocês são e pela amizade tão forte.

Aos meus amigos de infância Hugo Portuita, Pedro Goñi e João Paulo Siqueira que sempre me procuram, insistem em me trazer para um mundo distante do acadêmico, e me lembram que o samba não pode morrer. Obrigado pelo carinho e amizade ao longo deste muitos anos. E desculpem pelo sumiço. Minhas ausências estão parcialmente justificada nas páginas que seguem.

Aos meus amigos de conversas estranhas, em qualquer horário e sobre qualquer assunto, sempre regadas a muitas risadas, gritos e discussões intensas. Algumas delas, no batista às cinco da tarde, outras no 485 lotado às dez da noite que faziam as pessoas olharem com vergonha alheia para nós, seja falando de política, religião, mal dos outros ou qualquer outra coisa. Ou mesmo nos infinitos áudios do Whatsapp. Espero ainda continuar brigando e conversando com Adriano Côrtes, André Saraiva, Bruno Santiago, Bruno Braga, Camila Codeço, Carlos Lechner, Cristiane Krug, Daniel Soares, Douglas Picciani, Filipe Goulart, Flávio Ávila, Gabriel Vidal, Guilherme Sales, Hugo Carvalho, Luís Felipe Velloso, Nicolau Sarquis, Reinaldo de Melo e Souza, Tiago Domingues e Tiago Vital.

À Sonia Giacomo, professora da UFMS, com quem venho trabalhando no último ano e que eu espero continuar trabalhando. Sua energia, carisma e sinceridade são contagiantes. Além disso, todos os aprendizados com ela obtidos sobre como trabalhar em equipe, cumprir prazos e revisar trabalhos de forma cirúrgica têm sido essenciais para a minha formação. E agradeço ao Adan, por ter feito esta ponte entre nós.

Aos meus amigos do SMT, do lado aplicado da força. Isabela Apolinário e Renam Castro me ajudaram nas matérias com questões de computação e a entender a linguagem falada pelos engenheiros que soava tão estranha para mim. Ainda me assusto um pouco com um banco de filtros mas isso vai passar... Nossos trabalhos em equipe e seminários foram divertidos, por mais que muitas vezes eu não soubesse direito o que eu estava fazendo. Ao Flávio Ávila, grande amigo de seminários e de conversas político-ecônomicas, pela amizade aplicada. Espero finalmente ter tempo para nosso projetos começarem a andar. Meus colegas Hamed Yazdanpanah, Pedro Fonini e Rafael Zambrano, que trabalharam comigo no projeto do CENPES, me ajudaram a desbravar um mundo de engenharia de reservatórios completamente desconhecido para todos nós. Juro que tentarei não perturbar mais vocês as quatro da manhã para terminarmos um monte de relatórios e simulações. Aos professores Eduardo Barros Silva, Paulo Diniz e Sergio Lima Netto que compõem esta equipe e me ofereceram a oportunidade de trabalhar neste projeto, ajudando-me financeira e intelectualmente. A maneira como vocês encaram as situações de forma descontraída mas ao mesmo tempo profissional me encanta e fez com que todas as tarde de quarta fossem sempre engraçadas e leves mas ao mesmo tempo profundas. Aos professores Flávio Dickstein e Paulo Goldfeld por gentilmente cederem o espaço que utilizei para a escrita desta dissertação e para o trabalho no projeto nos últimos tempos. Foi muito interessante interagir com eles, mesmo que brevemente, e vê-los trabalhando com entusiasmo todos os dias. E ao Renan Vicente, pelas risadas, inúmeros favores com rodar simulações no IMEX para mim e pela companhia diária no LABMA.

Aos meus amigos Dingos da Montanha e companheiros de aventura Carolina Casals, Gabriel Vidal, Luis Felipe Velloso, Nicolau Sarquis, Petrônio Lima e Tiago Vital. Os dias ensolarados de trilha e as noites frias de travessia seriam menos engraçadas e intensos se vocês não estivessem juntos para dividir barraca, comida, roupa e as expências pelas quais passamos. Espero vocês para continuarmos as aventuras pela Europa!

Aos meus amigos de café, de Batista, de Anjinho, de almoço, etc, Bruno Scárdua, Fábio Ramos, Heudson Mirandola e Marcelo Tavares, que sempre estiveram presentes vibrando e compartilhando conquistas, que me proporcionaram ótimas conversas e me deram conselhos vitais sobre como não aumentar a entropia, resolver conflitos e como chegar ao sim comigo mesmo e com os outros. Em particular ao Bruno, mestre e amigo por quem nutro grande admiração, que me ensinou a importância dos slogans das propagandas de pneumáticos, dos alunos ilustres de Munique e que me ajudou em vários pontos fundamentais desde o curso de análise real.

Aos membros do seminário de Compressive Sensing: Amit Bhaya, Bernardo da Costa, César Niche, Hugo Carvalho, Iker Sobron, Luiz Wagner Biscainho, Marcello Campos e Wallace Martins. A paciência de vocês foi infinita ao me ouvir falar todas as terças, por mais de dois anos, sobre CS. Obrigado pelas discussões prolíficas e

por me fazer entender o assunto melhor.

Ao Bernardo da Costa por toda a amizade, ajuda, suporte, orientação, empolgação, por toda a matemática e ciência que aprendi, pelas dúvidas difíceis desta dissertação (e fora dela!) e pelas conversas que nunca têm fim. Tenho a sensação que toda vez que eu o encontro fico cada vez mais perdido sobre por onde começar a falar, já que o número de assunto entre nós está divergindo. Ao Carlos Tomei pelos almoços regados a histórias engraçadas e pelos conselhos de vida. Ele me ajudou de forma fundamental em momentos singulares onde eu achei que não dava para continuar em frente. Ao Roberto Imbuzeiro por conselhos envolvendo matemática pura e aplicada, pelas críticas fundamentais à esta dissertaçã e por toda a ajuda em utilizar a sala de videoconferência do IMPA. Sem ela, meu futuro talvez tivesse tomado outro rumo. Ao Amit Bhaya, pelo carinho, amizade, pela erudição e conversas que vão desde a União Soviética à ética científica, passando pelo controle dos sistemas fisiológicos e a culinária indiana. Ao Eduardo Silva por me ensinar muitas coisas de processamento de sinais e por arrumar tempo para me dar conselhos e estar sempre disposto a me ouvir mesmo tendo 73472416 alunos, projetos, defesas, seminários e reuniões. Aprendi com ele, mais do que com qualquer outra pessoa, a importância do trabalho duro e da disciplina para fazer as coisas funcionarem. Ao Fábio Ramos, por tudo. Pelas matérias que não tinham fim e que nos faziam vir ao fundão no dia 30 de dezembro, pelos conselhos diários envolvendo todos os aspectos da minha vida, pela quantidade de livros e assuntos que descobri graças a ele, pelas piadas e trocadilhos que faziam os almoços um ponto alto do dia e muitas outras coisas difíceis de enumerar aqui. Além disso, seu apoio e extrema confiança valeram mais do que qualquer orientação. E à todos eles, por prontamente aceitarem fazer parte da minha banca e por lerem com todo o cuidado o trabalho que segue.

Não posso deixar de citar os funcionários da secretaria Alan, Aníbal e Cristiano. Os três sempre foram exemplares para resolver todos os problemas buracráticos que tive e para regularizar meu histórico cheio de coisas estranhas. Agradeço a tia Deise (*in memoriam*) que estava todo dia no fundão, com bom humor ou sem, e além de entender como todo aquele conto kafkaniano funcionava, resolvia com primor tudo o que os alunos da aplicada precisavam. Espero, como ela, trabalhar até o último dia. E também ao Walter, que é a pessoa que faz o Instituto de Matemática funcionar e que mesmo com greve, cachorros, esgoto radioativo na linha vermelha, tiroteio na Vila do João, falta de água, ataque zumbi, etc, sempre esteve disposto a me ajudar com os problemas que surgiam na ABC-116 e a proporcionar o melhor ambiente possível para os alunos estudarem.

No que diz respeito à dissertação, agradeço ao Yuri Saporito que forneceu o template que serviu de molde para essa e inúmeras outras dissertações da aplicada, ao Luís Felipe Velloso pela ajuda e suporte com todas as figuras que precisei nos últimos tempos não somente nesta dissertação mas principalmente em apresentações diversas e pela rataria com o WinFIG e Xfig, ao Hugo Carvalho pela revisão ninja e por inúmeras melhorias no texto e ao Roberto Velho por um certo código mágico do Mathematica. Gostaria muito de ter descoberto isso antes...

À Natália Parentoni pelo suporte e por me ensinar a ver a vida de vários outros ângulos, repensando minhas posições e estruturas. O crescimento que você me proporciona é muito singular.

Aos criadores e colaboradores da Wikipedia, pelas inestimáveis consultas. Aos que trabalham duro para manter os servidores do Cazaquistão, de Neue, do Equador, da Samoa Ocidental, de Belize, do Território Britânico do Oceano Índico e da Rússia funcionando a todo vapor, apesar dos ataques constantes e que tornam o conhecimento livre de qualquer mordaça. Essa dissertação teria sido impossível sem os inúmeros esforços desses anônimos espalhados pelo mundo. Agradeço ao Donald Knuth, Leslie Lamport e à todos os que trabalham arduamente para ampliar este mundo incrível da formatação que é o LaTeX. Aos criadores do Detextify, esta ferramenta incrível que me ajudou em vários momentos. Sem todos vocês, este documento teria sido escrito naquele editor que *não pode ser nomeado* e as coisas não ficariam tão bonitas.

À Andrei Kolmogorov e Claude Shannon, duas das mentes mais privilegiadas da história da humanidade, pela inspiração e pelas ideias que originaram os primórdios desse trabalho. Sem eles, não é exagero dizer que tudo isso (e muitas outras coisas) teria demorado mais uns 500 anos para surgir.

Às minhas companhias da madrugada por proporcionarem múltiplos ensinamentos. Muitos de vocês foram essenciais em momentos recentes onde concentrações de medida ou teoremas malucos de russos sobre espaços de Banach não faziam mais nenhum sentido e eu estava de enjoado de olhar para o computador. Vivos ou mortos, carrego seus escritos como marcas indeléveis. Obrigado Atul Gawande, Carl Sagan, Douglas Adams, Edward Wilson, Fernando Pessoa, Fiodor Dostoievsky, Franz Kafka, George Orwell, José Saramago, Liev Tolstoi, Luiz

Felipe Pondé, Malcolm Gladwell, Nelson Rodrigues, Oliver Sacks, Otto Maria Carpeaux, Robert Sapolsky, Valter Hugo Mãe e Wislawa Szymborska.

Aos professores Felix Krahmer e Holger Boche pela oportunidade que terei daqui pra frente e por tudo que conversamos por email até agora. Estou sendo tratado de maneira íncrivel e espero retribuir isso trabalhando bastante e não decepcionando-os.

À minha família pelo suporte e pelo amor. Eu enganei vocês e na verdade eu nunca estudei psicologia dos anelídeos como muitas vezes eu disse em casa. Agradeço ao meu irmão Gabriel, minhas tias Marisa e Marilene e meu primo Flávio por tudo. Agradeço a minha avó Maria da Glória pelas conversas sobre a finitude da vida e por me ensinar, dentre muitas coisas, que em várias situações, a experiência é um pente que a gente ganha quando fica careca. E também ao Carlos e Irene, pessoas que acabaram por se tornarem meus segundos pais e que sei estarem presentes, de coração aberto, para o que eu precisar.

E finalmente, mas não menos importante...ao meu orientador César Niche, por tudo. Por uma orientação que virou amizade e que espero mantê-la sempre e pelo carinho e preocupação. Também pelos ótimos cursos, pela matemática aprendida, por topar me acompanhar nessa jornada nos últimos quatro anos, por acreditar em mim e nos meus projetos muitas vezes loucos e por sempre ter me dado liberdade de seguir em frente e estudar o que eu quisesse. Apesar da minha imensa teimosia e das muitas vezes em que discordei de suas ideias, sou grato a todas elas e a tudo o que aprendi com ele. Essa dissertação é fruto direto de toda essa ótima interação.

Ao meu pai Luiz Cláudio pelo entusiamo com ciência, por me ensinar o poder da comunicação e por querer sempre mais dos meus esforços. À minha mãe Marília, por ter me ensinado o poder das próprias escolhas, da leitura e busca incessante, por tudo o que enfrentamos juntos, no dia a dia, sempre com muita conversa e argumentação e por sempre ter me dito que isso ia dar certo. À ambos, pelo amor e carinho, por nunca terem medido esforços para me dar a melhor educação que estava ao alcance deles e por me ensinarem que, na vida, devemos escolher fazer aquilo que nos deixe feliz e que faça com que nunca olhemos para o relógio.

À Ivani Ivanova, revisora máxima desse texto (que em outras situações seria colocada como coautora...), lendo cada detalhe, cada demonstração, cada argumento e respondendo com um tratado completo sobre como melhorá-los. Se o texto ainda apresenta inúmeros erros (e eu sei que apresenta pois somente hoje achei uns três), foi porque eu não terminei suas revisões com a paciência e esmero suficientes e a culpa é inteiramente minha. Ela não só é a revisora desse texto mas fez com que minha vida passasse por uma revisão completa desde que saí por aí falando de números cardinais. Toda a gratidão à companhia, amor, carinho e dedicação de sua parte, ao longo de toda a montanha-russa, não podem ser descritos aqui.

Claudio Mayrink Verdun
Dezembro de 2016

“All that you touch
All that you see
All that you taste
All you feel.
All that you love
All that you hate
All you distrust
All you save.
All that you give
All that you deal
All that you buy,
beg, borrow or steal.
All you create
All you destroy
All that you do
All that you say.
All that you eat
And everyone you meet
All that you slight
And everyone you fight.
All that is now
All that is gone
All that’s to come and everything
under the sun is in tune but the
sun is eclipsed by the moon.

"There is no dark side of the
moon really.
Matter of fact it’s all dark." ."

Eclipse
Roger Waters - Pink Floyd

Resumo

Compressive Sensing

Claudio Mayrink Verdun

Resumo da dissertação de Mestrado submetida ao Programa de Pós-graduação em Matemática Aplicada, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática Aplicada.

Resumo: Vivemos em um mundo digital. Os aparelhos ao nosso redor interpretam e processam bits a todo instante. Para isso ser possível, a conversão de sinais analógicos para o domínio discreto se faz necessária. Ela se dá por meio dos processos de amostragem, quantização e codificação. Através de tais etapas, é possível processar e armazenar os sinais em dispositivos digitais. O paradigma clássico do Processamento Digital de Sinais nos diz que devemos realizar uma etapa de amostragem e, em seguida, uma etapa de compressão, por meio da quantização e codificação, para a aquisição de sinais. Entretanto, realizado desta forma, este processo pode ser custoso ou desnecessário, devido a uma alta taxa de amostragem ou a uma grande perda de dados na etapa de compressão. A partir desta observação, coloca-se a seguinte questão:

É possível realizar o processo de aquisição e compressão simultaneamente de forma a obter o mínimo de informação necessária para a reconstrução de um dado sinal?

Os trabalhos seminais de David Donoho, Emmanuel Candès, Justin Romberg e Terence Tao mostram que a resposta para esta pergunta é afirmativa em muitas situações envolvendo sinais naturais ou gerados pelo ser humano. O presente trabalho tem como objetivo estudar a área denominada Compressive Sensing, termo sem tradução que se refere à resposta positiva a esta pergunta, isto é, ao procedimento de realizar uma aquisição compressiva de dados. Esta dissertação é um guia do mochileiro para esta área que está em rápido crescimento. Os teoremas que a fundamentam são apresentados com detalhe e rigor. Do ponto de vista matemático, ela é uma interseção das áreas de Otimização, Probabilidade, Geometria de Espaços de Banach, Análise Harmônica e Álgebra Linear Numérica. Aqui são discutidas técnicas de Teoria de Frames, Matrizes Aleatórias e aproximações ótimas em espaços de Banach para demonstrar a viabilidade deste novo paradigma em Processamento de Sinais.

Palavras–chave. Compressive Sensing, Compressive Sampling, Compressed Sensing, Esparsidade, Representação Esparsa, Redundância, Teoria de Frames, Análise Harmônica Aplicada, Matrizes Aleatórias, Princípio da Incerteza, Processamento de Sinais.

Rio de Janeiro
Dezembro de 2016

Abstract

Compressive Sensing

Claudio Mayrink Verdun

Abstract da dissertaçao de Mestrado submetida ao Programa de Pós-graduação em Matemática Aplicada, Instituto de Matemática da Universidade Federal do Rio de Janeiro (UFRJ), como parte dos requisitos necessários à obtenção do título de Mestre em Matemática Aplicada.

Abstract: We live in a digital world. The devices around us interpret and process bits all the time. For this to be possible, a conversion of analog signals into the discrete domain is necessary. It happens through the processes of sampling, quantization and coding. Through such steps, it is viable to process and store such signals in digital devices. The classical paradigm of Digital Signal Processing for signal acquisition is a sampling step followed by a compression step, through quantization and coding. However, this way, the process can be costly or unnecessary due to a high sampling rate or a large loss of data in the compression stage. From this observation, the following question is posed:

Is it possible to perform the acquisition and compression process simultaneously in order to obtain the minimum information for the reconstruction of a given signal?

The seminal works of David Donoho, Emmanuel Candès, Justin Romberg and Terence Tao show that the answer to this question is affirmative in many circumstances involving natural or man-made signals. The present work aims to study the area called Compressive Sensing, which is a procedure to perform data acquisition and compression at the same time. This dissertation is a hitchhiker's guide to this rapidly growing field. The fundamental theorems will be explored with detail and rigor. From the mathematical point of view, it is an intersection of the Optimization, Probability, Geometry of Banach Spaces, Harmonic Analysis and Numerical Linear Algebra. Here we discuss the techniques of Frame Theory, Random Matrices and approximation in Banach Spaces to demonstrate the feasibility of this new paradigm in Signal Processing.

Keywords. Compressive Sensing, Compressive Sampling, Compressed Sensing, Sparsity, Sparse Representation, Redundance, Frame Theory, Applied Harmonic Analyis, Random Matrices, Uncertainty Principle, Signal Processing.

Rio de Janeiro
December of 2016

Contents

Preface

Measure what can be measured, and make measurable what cannot be measured.
Attributed to Galileo Galilei

Measure what should be measured.
Compressive Sensing version by Thomas Strohmer

In this dissertation we will explore one of the fascinating connections that emerged over the last 20 years among the communities of Applied Mathematics, Electrical Engineering and Statistics. It deals with several relevant ideas that have arisen about data acquisition and how we perform measurements.

We live in an era where data and information are spread everywhere and it is a major challenge getting them in the most economical way. Two studies sponsored by EMC corporation, [Gantz & Reinsel '10] and [Gantz & Reinsel '12], measured all digital data created, replicated, and consumed in a single year, and also did a projection of the digital universe size for the end of the decade. They argued that at the end of 2020, there will be 40000 exabytes or 40 trillion gigabytes of data in the world. They say *"From now (2012) until 2020, the digital universe will about double every two years"*. Also, it can be expensive in terms of time, money or (in case of, say, Computed Tomography) damage done to the object from which information is being acquired.

Based on these facts, more than ever, we need efficient ways to represent, store and manipulate data. The question of how to sample and compress datasets is of major importance. In the last twenty years, many researchers asked if there are classes of signals that can be encoded using just few informations about them without much perceptual loss.

Simultaneously, they realized that many signals seems to have few degrees of freedom, despite the fact that they "live", in principle, in high-dimensional spaces. Knowing this, one can potentially design efficient sampling and storage protocols that take the useful information content embedded in a signal and condense it into a small amount of data. For example, [Candès, Romberg & Tao I '06] performed an impressive numerical experiment and showed that a 512×512 pixel test image, known in the Image Processing community as the Logan-Shepp phantom, can be reconstructed from 512 Fourier coefficients sampled along 22 radial lines. In other words, more than 95% of the relevant data is missing yet a very efficient compression and reconstruction scheme was found.

At the same year, [Donoho '06] realized that these compressed data acquisition protocols are, in his own words *"nonadaptive and parallelizable"*. They do not require any prior knowledge of the data other than the fact that data has a parsimonious representation with few coefficients. Furthermore, the measurements can be performed all at the same time. He called this new simultaneous acquisition and compression process *Compressed Sensing*.

These two fundamental papers were followed by the works [Candès & Tao I '06], [Candès & Tao II '06] and [Candès, Romberg & Tao II '06]. These three works, in turn,have further developed the results and disseminate them in the communities of Statistics, Signal Processing and Mathematics.

Since these five founding papers were published, ten years ago, there was an explosion in the theory and applications of Compressive Sensing. Therefore it is important to highlight that ideas which emerged from the notions of sparsity, incoherence and randomness and that now permeate the world of data science will not be transient.

In Brazil, the emergence of Compressive Sensing was in 2009, in a course at the 27° Brazilian Mathematical Colloquium through the book [Schulz, da Silva & Velho '09]. I attended the classes and despite

the fact that I was in the first year of college, the impressions recorded in my mind were that it would be interesting to put together Mathematics, Signal Processing, computation and statistical skills and try to solve a real-life problem.

After I spent two years studying Fourier Analysis and Inverse Problems and also working on mathematical problems from medical images, I decided that I would dedicate myself to this subject and its new developments.

Now, a little over ten years after the articles that founded the theory have been published, there are some excellent books that are partially or entirely dedicated to it. See [Elad '10], [Damelin & Miller '12] [Eldar & Kutyniok '12], [Rish & Grabarnik '14] and [Eldar '15]. However, up to now, the comprehensive book about the subject is [Rauhut & Foucart '13]. It is the main reference of this dissertation, my companion in the last three years and most of the main ideas and proofs were extracted from it. Despite this, the influence of the other five books is also present in several passages of this text.

There are some excellent reviews and surveys about Compressive Sensing. See [Candès '06], [Holtz '08], [Baraniuk '07], [Romberg '08], [Strohmer '12], [Candès & Wakin '08], [Jacques & Vandergheynst '11]. In addition, the success of this theory has been described in four issues of important journals entirely devoted to the subject. The articles in these four issues served as a source of inspiration and clarification on some obscure and hard points of the theory.

- IEEE Signal Processing Magazine, Vol. 25, No. 2, Mar 2008.
- IEEE Journal on Selected Topics in Signal Processing, Vol. 4, No. 2, Apr 2010.
- Proceedings of the IEEE, Vol. 98, No. 6, Jun 2010.
- IEEE Journal on Emerging and Selected Topics in Circuits and Systems, Vol. 2, No. 3, Sep 2010.

Unfortunately, I did not solve a real problem in the last three years, as was my initial desire. Instead, I studied Compressive Sensing theory carefully. It is essential to say that in each chapter I have outlined important open problems connected to Compressive Sensing. The purpose of this dissertation is to state Compressive Sensing as a new paradigm in Signal Processing and to develop the main techniques necessary to completely understand it from the mathematical viewpoint. I hope the next three years will be enough to solve or, at least, to understand a real problem using the knowledge acquired from the study of these techniques.

Chapter 1

Sparse Solutions of Linear Systems

> Is there anything of which one might say, "See this, it is new"? Already it has existed for ages Which were before us.
> *Ecclesiastes 1:10*

1.1 The What, Why and How of Compressive Sensing

In the last century, we have witnessed the deployment of a wide variety of sensors that acquire measurements to represent the physical world. Microphones, telescopes, anemometers, gyroscopes, transducers, radars, thermometers, GPS, seismometers, microscopes, hydrometers, etc are everywhere. The purpose of all these instruments is to directly acquire measures in order to capture the meaning of the world around us. Besides, a world without digital photos, videos and sounds being stored or shared is unimaginable nowadays.

The increase in technology has allowed us to obtain more and more data in a short time. To deduce the state or structure of a system from this data is a fundamental task throughout the sciences and engineering. The first step to acquire the data is *sampling*. We need to sample the signals obtained from the real world in order to convert them from the analogical domain to a sequence of bits in the digital domain. The second step is *compression*, which involves encoding information using fewer bits than the original representation. Both steps are necessary to store, manipulate, process, transmit and interpret data.

However, in many situations it is difficult and costly to obtain a massive amount of data or even slow, such as medical image acquisition. Moreover, nowadays, in many analog-digital (A/D) converters, after the signal is sampled, significant computations are expended on lossy compression algorithms. A representative example of this paradigm is the digital camera of any smartphone. It acquires millions of measurements each time a picture is taken and most of the data is discarded after the acquistion through the application of an image compression algorithm. Therefore, when the signal is decoded, typically only an approximation of the original one is produced. The main question that motivates this text is

Can both operations, sampling and compression, be combined?

That is, instead of high-rate sampling followed by computationally expensive compression, is it possible to reduce to number of samples and produce a compressed signal directly? The classical sampling theory does not exploit the structure or any prior knowledge about the signal being sampled. All we need to know is its frequency information and use this to determine the sample rate.

The objective of this dissertation is the study of *Compressive Sensing*, also known as *Compressed Sensing* or *Compressive Sampling*. It takes its name from the fact that data acquisition and compression can be performed simultaneously in such a way that the reconstruction algorithms will exploit the structure of the data. This technique is, nowadays, an essential underpinning of data analysis.

We have two basic hypotheses behind the theory: *sparsity* and *incoherence*. The first one is related to the objects we are acquiring, i.e, we will use few measurements in order to capture signals by exploiting

their parsimonious representation in some basis. The second one is related to the measurement system. It extends the duality between time and frequency and the idea behind it is that an object that has sparse representation in a given basis must be spread out in any other basis, such as the one which represents the acquisition domain. We wil also see that *randomness* plays a key role in the design of optimal incoherent system. Thus, it can be considered as a third ingredient of the theory.

From a mathematical viewpoint, Compressive Sensing can be seen as a chapter of contemporary Linear Algebra since it deals with the pursuit of sparse solution of underdetermined linear systems. Nevertheless, mathematical techniques behind it are very sophisticated, going from Probability Theory, passing through Harmonic Analysis and Optimization and arriving at the Geometry of Banach Spaces. From the applied viewpoint, it is a fecund intersection of many areas from Electrical Engineering, Computer Science, Statistics and Physics.

Through the use of Compressive Sensing, we can not only provide theoretical guarantees for the minimal number of measurements but also efficient algorithms for practical reconstruction. And all this is based on a simple principle of parsimony, which was enunciated many times, but divulged by the medieval philosopher William of Ockham: *"Pluralitas non est ponenda sine neccesitate"*, i.e., entities should not be multiplied unnecessarily.

It is interesting to note that ideas of Compressive Sensing ideas spontaneously occur in nature due to natural selection: the mammalian visual system has been shown to behave using sparse and redundant representation of sensory input. Some researches, starting from the works of Barlow (great-grandson of Charles Darwin), Hubel and Wiesel argued that a dramatic reduction occurs from the information that hits the retina to the information present in the visual cortex. Typically, a neuron in the retina responds simply to whatever contrast is present at that point in space, whereas a neuron in the cortex would respond only to a specific spatial configuration of light intensities. This leads to an optimal representation of the visual information in the brain and is known nowadays in the field of Neuroscience as *Sparse Coding*. See [Olshausen & Field '04] and [Huang & Rao '11].

The first application of Compressive Sensing was in pediatric magnetic resonance imaging (MRI). See [Vasanawala, Alley, Hargreaves, Barth, Pauly & Lustig '10] and the blog post [Ellenberg - 02/22/10]. It enables a sevenfold speedup while preserving diagnostic quality and resolution. In this imaging modality, the traditional approaches to produce high-resolution imagines relies on Shannon Sampling Theorem (Theorem 1.2 demonstrated below), and may take several minutes. For a heart patient imaging, one cannot hold his breath for too long and the artificial induction of a cardiac arrest can be dangerous and irreversible, depending on the health of the patient. Thus, to get images without blurring is a difficult task. See [Lustig, Donoho & Pauly '07] for the problem statement and the use of Compressive Sensing.

After the launch of these preprints, Compressive Sensing appeared as a noticeable improvement, gaining the attention of the media and of several research groups interested in improving the performance of signal reconstruction. Nowadays, besides medical imaging, it leads to important contributions in Radar Technology, Error Correction Codes, Systems Recommendation, Wireless Communications, Learning Algorithms, DNA Microarrays, etc. See Section 1.2 of [Rauhut & Foucart '13] and references therein for these and more examples.

In mathematical terms, the problem of acquiring and compressing simultaneously the signal $x \in \mathbb{C}^N$ of interest is modeled by a (fat and short) matrix measurement, converting x to $y = Ax$. After, we need to reconstruct it using the underdetermined linear system

$$Ax = y,$$

where the measurement *fat and short*[1] matrix $A \in \mathbb{C}^{m \times N}$ models the information acquisition process and $y \in \mathbb{C}^m$ is the observed data. Since we are in a compression regime, we expect (or impose) that $m \ll N$. At a first moment, this system has an infinite number solutions. But the role played by sparsity will allow us to reduce it to a unique solution and despite a seeming lack of sufficient information in the acquisition, we will see that recovery algorithms have a wondrous performance. Therefore we must start by understanding what sparsity is and how and when it leads to unique solutions of underdetermined linear systems. It is important to note that linear measurements are, in principle, not necessary and

[1] For this name and other ideas, one should consult the blog by Dustin Mixon: `https://dustingmixon.wordpress.com/`.

we could develop a theory for non-linear measurements. Nevertheless, linear sensing is widely used and simpler than non-linear one and we study this framework throughout the text.

In this dissertation we begin with basic definitions of sparse and compressible vectors. Then, we will explore the algorithms that solve efficiently the problem of the recovery of sparse vectors. They are divided in three main families: optimization algorithms, greedy algorithms and thresholding algorithms. After this, we will discuss the three most important properties in the study of indeterminate linear systems: the *nullspace property*, the *coherence property* and the *restricted isometry property*. Later, we will proceed to the study of non-asymptotic probability techniques, specially the concentration of measure phenomena, and how to use them to design efficient measurement matrices with an optimal number of measurements. Finally, the Geometry of Banach Spaces will play a special role to understand the minimal amount of information to recover sparse signals and it will appear as the main tool to provide sharper general results for even more general reconstruction methods than those presented in the previous chapters. The flowchart of ideas developed through this dissertation can be seen in Figure 1.1 below.

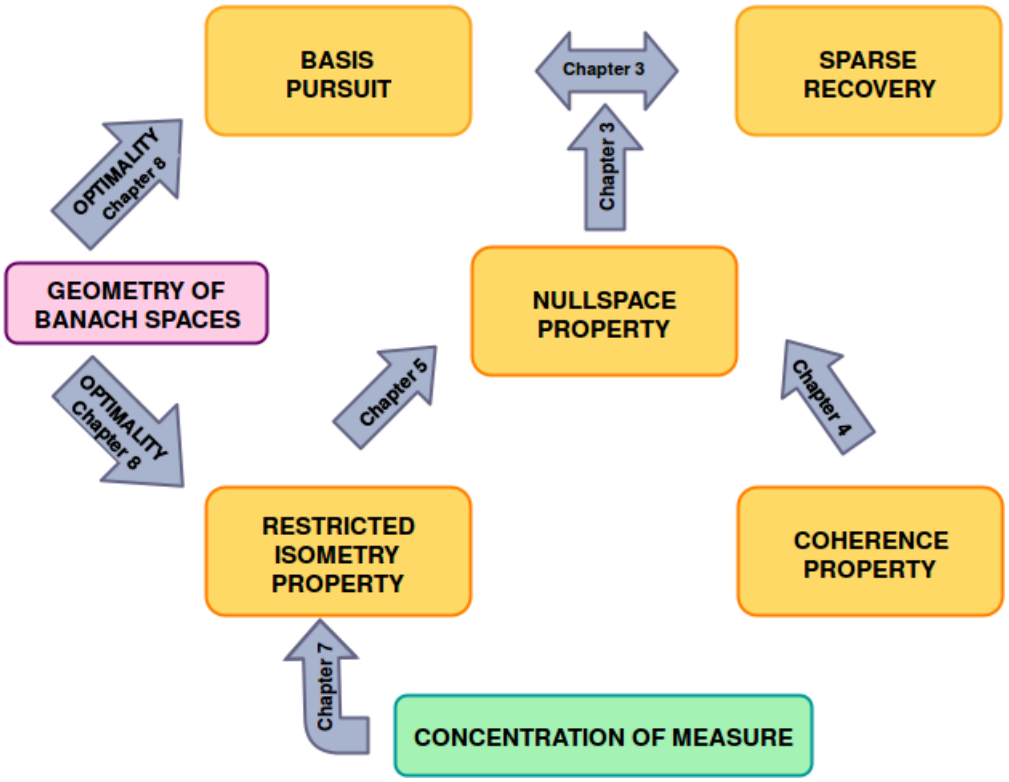


Figure 1.1: Flowchart of the dissertation

The core of this diagram is the "equivalence" between Basis Pursuit, an algorithm from convex optimization, and sparse recovery, a combinatorial problem. We will see that the Nullspace Property (NSP) plays a fundamental role in this equivalence. More specifically, Basis Pursuit allows sparse recovery if and only if the measurement matrix satisfies this property, as will be seen in Chapter 3. Since it is very hard to establish this property, we will replace it by two different sufficient conditions: the Coherence Property and the Restricted Isometry Property (RIP). This will be done in Chapters 4 and 5, respectively. The second will lead to better results in the number of measurements through the use of probability techniques, such as the concentration of measure. Therefore, in Chapters 6 and 7 we will explore probabilistic tools and how to use then in the Compressive Sensing framework. Finally, in Chapter 8 we introduce some concepts of the Geometry of Banach Spaces in order to establish the optimality of Basis Pursuit and other methods of sparse recovery discussed through this dissertation.

1.2 A Little (Pre)History of Compressive Sensing

Compressive Sensing has a long history. The first ideas of sparse estimation can be traced back to Gaspard Riche de Prony, a French mathematician and engineer. In 1795, he proposed an algorithm to estimate the parameters associated with a small number of complex exponentials sampled in the presence of noise. See Theorem 2.15 of [Rauhut & Foucart '13] and Section 15.2 of [Eldar '15].

After, in the first half of the 20th century, the works of Constantin Caratheodory, Arne Beurling and Benjamin Logan in Harmonic Analysis showed that it was possible to recover certain sinusoids and pieces of Fourier Transform by exploiting the notion of minimal extrapolation, which nowadays we recognize as ℓ_1-minimization. See [Donoho '10]. This will be further discussed in Chapter 2.

During World War II, the technique of *combinatorial group testing* was introduced. It was a time when there were many people infected with syphilis and the problem was to discover who was infected and who was not by using efficient testing schemes, considering that resources were scarce. The US Public Health Service did not want ill men serving in the military but it was not possible to test all individuals, hundreds of thousands of people, in order to identify the small proportion of infected people. Then, instead of test single individuals, the blood was combined in an efficient way to reveal that one person in this combination had the disease. Therefore the task was to identify all individuals with syphilis while minimizing the number of tests and again the parsimony principle applies. See [Dorfman '43] and [Gilbert, Iwen & Strauss '08].

In the late seventies and early eighties, researchers from Geology and Geophysics communities showed that the structure of the layered format in the earth's interior could be explored to increase the accuracy of seismic signal recovery. [Taylor, Banks & McCoy '79], [Levy & Fullagar '81] and [Walker & Ulrych '83], showed that very incomplete measurements can be used to recover a full wideband seismic signal, despite that no low-frequency can be acquired due to the nature of the seismic measurements. Also, the first paper to explicit state the use of ℓ_1-norm for signal reconstruction was written by people working on Geophysical Inverse Problems: [Santosa & Symes '86].

After these contributions, the Magnetic Resonance Spectroscopy and Radio Astronomy communities started to use sparsity concepts in the signal recovery. In particular, methods exploiting this parsimonious representation of data were faster and more efficient than the classical method of Maximum Entropy Inversion [Donoho, Johnstone, Stern & Hoch '92].

Essentially at the same time, the "Wavelets explosion" appeared through the works of Daubechies, Meyer, Mallat, Coifman, Wickerhauser and others. From their ideas, it became typical to describe effective media representation, such as image and video, by using Wavelets. Afterwards, Wavelets became part of the standard techniques in Signal Processing and standard technologies such as fingerprint databases. See [Daubechies '92] and [Burrus, Gopinath & Guo '97].

The first works that showed the general algorithmic principles which are used in compressive sensing were [Mallat & Zhang '93] and [Chen & Donoho '94]. In these papers, a systematic study was performed in order to develop a general theory of sparse representations of signals and to understand how this can be done in a efficient way.

Hitching on the paper [Chen & Donoho '94], it is important to point out that a large part of the history and evolution of ideas before Compressive Sensing, in the nineties, can be understood from the point of view of the works of David Donoho and his collaborators. Donoho, a very prolific mathematician, paved the way until to founding papers of Compressive Sensing. It starts with seminal paper [Donoho & Stark '89], where important ideas about the relation between the uncertainty principle and the acquisition of information from a signal where showed. After this, many of his works contained fundamental ideas in Statistics and Signal Processing that culminated in Compressive Sensing. One important example is [Donoho & Huo '01]. The goal of this paper was to establish the connections between convex optimization problems and optimal parsimonious representations of signals by using very few coefficients, what they called an *Ideal Atomic Decomposition*.

Parallel to advances in signal processing, similar ideas have emerged in the Statistics community. The development of LASSO by [Tibshirani '96], a linear regression that uses ℓ_1-norm as a regularizer, started a new era in Statistics and Machine Learning where many problems, previously intractable, could be solved.

Finally, the foundation of Compressive Sensing can be credited to [Donoho '06], [Candès & Tao I '06], [Candès & Tao II '06], [Candès, Romberg & Tao I '06] and [Candès, Romberg & Tao II '06]. They combined deep techniques from Probability Theory, Approximation Theory and Banach Spaces Geometry in order to achieve a theoretical breakthrough: they showed how to exponentially decrease the number of measurements for an accurate and computationally efficient recovery of signals, provided that they are sparse. These five works revolutionized the way we think about Signal Processing. Now, according to Google Scholar, they have now more than 52000 citations together.

Subsequently, Compressive Sensing turned into a rapidly growing field. Since 2006, many sharper theorems, more efficient algorithms, dedicated hardware and generalizations of the theory emerged. It will probably take a few more years for all contributions to be properly organized so we can tell the story of new ideas, developments and breakthroughs of the field in the last twenty years. Before we start with it, we need to come back to the fundamentals of sampling, specially to the classical Sampling Theorem.

1.3 Sampling Theory

Sampling lies at the core of Signal Processing and is the first step of analog to digital conversion. It is the gate between the discrete and the continuous world. Other two important steps are quantization and coding. The former is the reduction of the sample values from their continuous range to a discrete set. The latter generates a digital bitstream as the final representation of the continuous time signal. This dissertation is about sampling. We start by looking at the celebrated Shannon-Nyquist-Kotelnikov-Whittaker Theorem[2]. First, we need a definition.

Definition 1.1. The *Paley-Wiener Space* $PW_\Omega(\mathbb{R})$ is the subspace of $L^2(\mathbb{R})$ consisting of all functions with Fourier transforms supported on finite intervals symmetric around the origin. More precisely,

$$PW_\Omega(\mathbb{R}) = \big\{ f \in L^2(\mathbb{R}) : \hat{f}(\xi) = 0 \text{ for } |\xi| \geq \Omega \big\}.$$

As [Mishali & Eldar '11] points out, *"The band-limited signal model is a natural choice to describe physical properties that are encountered in many applications. For example, a physical communication medium often dictates the maximal frequency that can be reliably transferred. Thus, material, length, dielectric properties, shielding and other electrical parameters define the maximal frequency Ω. Often, bandlimitedness is enforced by a lowpass filter with cutoff Ω, whose purpose is to reject thermal noise beyond frequencies of interest".*

Theorem 1.2. (Theorem 13 of [Shannon '48]): *Let $f(t) \in PW_\Omega(\mathbb{R})$. Then $f(t)$ is completely determined by its samples at the points $\{t_n = n\pi/\Omega\}_{n\in\mathbb{Z}}$. Also, the following reconstruction formula holds*

$$f(t) = \sum_{n\in\mathbb{Z}} f\left(\frac{n\pi}{\Omega}\right) \frac{\sin(\Omega t - n\pi)}{\Omega t - n\pi},$$

where the series above converge in the L^2 sense and uniformly.

Remark 1. *Here we provide a proof of Sampling Theorem because it is difficult to find a rigorous one in the literature. For a proof in a abstract framework, with more general interpolation functions than $\frac{\sin t}{t}$, see Theorem 2.7 of [Güntürk '00]. Also, for an engineering interpretation of Theorem* 1.2 *as "sampling by modulation", see Section 4.2.2 of [Eldar '15].*

Proof. Since $f(t) \in L^2(\mathbb{R})$, then $\hat{f}(\xi) \in L^2(\mathbb{R})$. Also, as a consequence of the Paley-Wiener Theorem (see Section 4.3 from [Stein & Shakarchi '05]), $f(t)$ is the restriction to the real line of an analytic function. We can write the Fourier Series of $\hat{f}(\xi)$ in the interval $[-\Omega, \Omega]$ since the complex exponentials $\{e^{in\pi\xi/\Omega}\}_{n\in\mathbb{Z}}$ form an orthogonal basis for $L^2(-\Omega, \Omega)$. In particular, the coefficients will be given by

[2] In the literature of Approximation Theory, it is known as the *cardinal theorem of interpolation*. See [Higgins '85].

$$c_n = \frac{1}{2\Omega}\int_{-\Omega}^{\Omega} \hat{f}(\xi)e^{-in\pi\xi/\Omega}d\xi.$$

By hypothesis, $\hat{f}(\xi) = 0$ for all $|\xi| \geq \Omega$. This yields, for all $x \in \mathbb{Z}$,

$$c_{-n} = \frac{1}{2\Omega}\int_{-\Omega}^{\Omega} \hat{f}(\xi)e^{in\pi\xi/\Omega}d\xi = \frac{1}{2\Omega}\int_{-\infty}^{\infty} \hat{f}(\xi)e^{in\pi\xi/\Omega}d\xi = \frac{\pi}{\Omega}\frac{1}{2\pi}\int_{-\infty}^{\infty} \hat{f}(\xi)e^{i\xi\frac{n\pi}{\Omega}}d\xi = \frac{\pi}{\Omega}f\left(\frac{n\pi}{\Omega}\right),$$

where we used the Inversion Formula for the Fourier Transform in the last equality. Thus, we obtain

$$\hat{f}(\xi) = \frac{\pi}{\Omega}\sum_{n=-\infty}^{\infty} f\left(\frac{n\pi}{\Omega}\right)e^{-i\xi\frac{n\pi}{\Omega}} = \lim_{N\to\infty}\sum_{n=-N}^{N}\frac{\pi}{\Omega}f\left(\frac{n\pi}{\Omega}\right)e^{-i\xi\frac{n\pi}{\Omega}}. \tag{1.1}$$

The convergence above is in the $L^2(-\Omega,\Omega)$ sense. We have also that

$$||\hat{f}||_1 = \int_{-\Omega}^{\Omega}|\hat{f}(\xi)|dx = \int_{-\Omega}^{\Omega} 1\cdot|\hat{f}(\xi)|dx \leq ||1||_2||\hat{f}||_2 = \sqrt{2\Omega}||\hat{f}||_2.$$

This leads to the fact that $\hat{f}(\xi) \in L^1(\mathbb{R})$ and that the convergence in $L^2(\mathbb{R})$ implies convergence in $L^1(\mathbb{R})$. Using that $\hat{f}(\xi) = 0$ for $|\xi| \geq \Omega$, we can multiply Equation (1.1) on both sides by $\chi_{[-\Omega,\Omega]}(\xi)$, the characteristic function of the interval $[-\Omega,\Omega]$. This yields

$$\left|\left|\hat{f}(\xi) - \frac{\pi}{\Omega}\sum_{n=-N}^{N} f\left(\frac{n\pi}{\Omega}\right)e^{-i\xi\frac{n\pi}{\Omega}}\chi_{[-\Omega,\Omega]}(\xi)\right|\right|_p \xrightarrow[N\to\infty]{} 0.$$

for $p = 1$ and $p = 2$. We known, by Parseval's Theorem, that the Inverse Fourier Transform is a bounded continuous operator $\mathcal{F}^{-1} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$. Also, From Riemann-Lebesgue Theorem, it is a bounded continuous operator from $L^1(\mathbb{R})$ to $C_0(\mathbb{R})$, the space of continuous functions vanishing at infinity equipped with the supremum norm. Therefore we conclude

$$f = \mathcal{F}^{-1}(\hat{f}) = \mathcal{F}^{-1}\left(\lim_{N\to\infty}\frac{\pi}{\Omega}\sum_{n=-N}^{N} f\left(\frac{n\pi}{\Omega}\right)e^{-i\xi\frac{n\pi}{\Omega}}\chi_{[-\Omega,\Omega]}(\xi)\right)$$

$$= \lim_{N\to\infty}\frac{\pi}{\Omega}\sum_{n=-N}^{N} f\left(\frac{n\pi}{\Omega}\right)\mathcal{F}^{-1}\left(e^{-i\xi\frac{n\pi}{\Omega}}\chi_{[-\Omega,\Omega]}(\xi)\right),$$

where the convergence occurs both in the $L^2(\mathbb{R})$ norm and in $C_0(\mathbb{R})$. To finish, note that

$$\mathcal{F}^{-1}\left(e^{-i\xi\frac{n\pi}{\Omega}}\chi_{[-\Omega,\Omega]}(\xi)\right) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-i\xi\frac{n\pi}{\Omega}}\chi_{[-\Omega,\Omega]}(\xi)e^{i\xi t}d\xi = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{-i\xi\frac{n\pi}{\Omega}}e^{i\xi t}d\xi$$

$$= \frac{1}{2\pi}\int_{-\Omega}^{\Omega} e^{-\xi t - i\xi\frac{n\pi}{\Omega}}d\xi = \frac{\Omega}{\pi}\frac{\sin(\Omega t - n\pi)}{\Omega t - n\pi}.$$

□

This is a modern translation of the theorem. We can compare it with the original words of Shannon: *"If a function $f(t)$ contains no frequencies higher than Ω cycles-per-second, it is completely determined by giving its ordinates at a series of points spaced $1/2\Omega$ seconds apart".*

He also wrote *"it is a fact which is common knowledge in the communication art..but in spite of its evident importance it seems not to have appeared explicitly in the literature of communication theory".*

Many signals of interest are modeled as *bandlimited* functions, i.e., real valued functions on the real line with compactly supported Fourier transforms. Then Theorem 1.2 tells us that it is possible to reconstruct

a band-limited function using uniform samples. It says that an enumerable amount of information is sufficient to reconstruct any (band-limited) function on the continuum. The heuristic behind it is that band-limited functions have limited time variation, and can therefore be perfectly reconstructed from equalispaced samples with a rate at least Ω/π, its maximum frequency, called the *Nyquist rate*.

Remark 2. *Theorem* 1.2 *has a long and prolific history. It is a striking case of multiple discoveries and many names beyond Shannon, Nyquist, Kotelnikov and Whittaker are associated with it. Also, it forms the basis of Signal Processing. See [Jerri '77], [Unser '00], [Meijering '02], [Ferreira & Higgins '11] and references therein. Also, many generalizations, including results for not necessarily band-limited functions or nonuniform sampling can be found in the literature. See [Butzer & Stens '92] and [Marvasti '01].*

Although it forms the basis of modern Signal Processing, Theorem 1.2 has two major drawbacks. The first one is that not all signals of the real world are band-limited. The second is that if the bandwidth is too large, then we must have too many samples in order to perform the acquisition of the signal. Since, in many situations, it is very difficult to sample at high rate, some alternatives began to be considered. These alternatives are known as *sub-Nyquist sampling* strategies.

Incredibly, the first overcoming of the Sampling Theorem occurred before its existence. It was proved in [Carathéodory 1911] that if a signal is a positive linear combination of *any* N sinusoids, it is uniquely determined by its value at $t = 0$ and *any* other $2N$ points. Depending on the highest frequency sinusoid in the signal, this can be much better than Nyquist rate. While the number of samples in the Nyquist rate increases with Ω, in Caratheodory's argument it increases with N.

Among all sub-Nyquist sampling alternatives, Compressive Sensing is one of the most important [Mishali & Eldar '11]. It focuses on efficiently measuring a discrete signal through a measurement matrix $A \in \mathbb{C}^{m\times N}$ with fewer measurements than the ambient dimension. Although it is a theory for discrete signals, recently some researches generalized ideas from Compressive Sensing to analog signals. See [Tropp, Laska, Duarte, Romberg & Baraniuk '10] and [Adcock, Hansen, Roman & Teschke '13].

This sensing with fewer measurements is only possible (and useful) because many natural signals are *sparse* or *compressible*. Using this priori information will allow us to design efficient reconstruction schemes. These, in turn, will be able to recover the vectors as the unique solution of some optimization problems.

1.4 Sparse and Compressible Vectors

In this section we define the main objects of our study: sparse and compressible vectors. The fundamental hypothesis we will explore is that many real-world signals have most of their components being zero (or approximately zero).

Definition 1.3. The support of a vector $x \in \mathbb{C}^N$ is the index set of its nonzero entries, i.e.,

$$\operatorname{supp}(x) = \{j \in [N] : x_j \neq 0\}.$$

A vector $x \in \mathbb{C}^N$ is called s-sparse if at most s of its entries are nonzero, i. e., if $||x||_0 = \#(\operatorname{supp}(x)) \leq s$. The set of all s-sparse signals is denoted by $\Sigma_s = \{x : ||x||_0 \leq s\}$.

Despite $||x||_0$ not being a norm, it is abusively called this way throughout the literature on Sparse Recovery, Signal Processing or Computational and Applied Harmonic Analysis. It can be seen as the limit of ℓ_p-quasinorms as p decreases to zero through the following observation.

$$||x||_p^p = \sum_{j=1}^{N} |x_j|^p \xrightarrow{p\to 0} \sum_{j=1}^{N} \mathbf{1}_{\{x_j\neq 0\}} = \#\{j \in [N] : x_j \neq 0\}.$$

Remark 3. *Sparsity has no linear structure. Given any $x, y \in \Sigma_s$, we do not necessarily have $x+y \in \Sigma_s$, although we do have $x + y \in \Sigma_{2s}$. The set Σ_s consists in the union of all possible $\binom{N}{s}$ canonical subspaces of $\mathbb{C}^N$.*

As discussed in Section 1.3.2 of [Eldar & Kutyniok '12], in certain applications it is possible to have more concise models by restricting the feasible signal supports to a small subset of the possible $\binom{N}{s}$ selection of nonzero coefficient. This leads to the concept of *sparse union of subspaces*. See [Duarte & Eldar '11].

In the Signal Processing community, it is said that we can model the sparse signal $x \in \mathbb{C}^m$ as the linear combination of N elements, called *atoms*, that is

$$x = \Phi\alpha = \sum_{i=1}^{N} \alpha_i \varphi_i,$$

where α_i are the representation coefficients of x in the *dictionary* $\Phi = [\varphi_1, \ldots, \varphi_N]$. We will require that the elements of the dictionary $\{\varphi\}_{i=1}^N$ span the entire signal space. Here it will be $\mathbb{C}^m$ but in other situations it can be some specific subset or proper subspace of $\mathbb{C}^m$. This representation is often *redundant* since we allow $N > m$. Due to redundancy, this representation is not unique, but we typically consider (or look for) the one with the smallest number of terms. A signal may be not sparse in a first moment but it may be "sparsified" just by using a specific dictionary.

What is behind the concept of sparsity is the philosophy of *Occam's razor*: when we are faced with many possible ways to represent a signal, the simplest choice is best one. This is so because there is a cost to describe the basis we are using for this representation. However, sparsity is a theoretical abstraction. In the real world, one does not seek for sparse vectors but, instead, for *approximately* sparse vectors. Even more, we want to measure how close our signals of interest are to a true sparse one. This measure is given by the next concept.

Definition 1.4. For $p > 0$, the *ℓ_p-error of best s-term approximation* to a vector $x \in \mathbb{C}^N$ is defined by

$$\sigma_s(x)_p = \inf\{||x - z||_p, z \in \mathbb{C}^N \text{ is } s\text{-sparse}\}.$$

Remark 4. *The infimum in Definition* 1.4 *will always be achieved by an s-sparse vector $z \in \mathbb{C}^N$ whose nonzero entries are equal to the s largest absolute entries of x. There is no reason for this vector to be unique. Nevertheless, when it attains the infimum, it occurs independently of $p > 0$.*

The concept of *approximately sparse* or *compressible* vector is a little more ingenious and realistic. Instead of asking for the number of nonzero component to be small, we ask for the error of its best s-term approximation to decay fast in s.

This can be modeled in two different ways. The first one is by using ℓ_p balls for some small $p > 0$. The second one is by exploring the concept of significant components of a vector itself. This will lead to the concept of weak ℓ_p-spaces.

Proposition 1.5 tells us that discrete signals which belong to the unit ball in some ℓ_p-norm, for some small $p > 0$ are good models for compressible vectors. Here we will denote the nonconvex ball in $\mathbb{R}^N$ equipped with ℓ_p-norm by $B_p^N = \{z \in \mathbb{C}^N : ||z||_p \leq 1\}$.

Proposition 1.5. *For any $q > p > 0$ and any $x \in \mathbb{C}^N$,*

$$\sigma_s(x)_q \leq \frac{1}{s^{1/p-1/q}} ||x||_p.$$

Definition 1.6. The nonincreasing rearrangement of the vector $x \in \mathbb{C}^N$ is the vector $x^* \in \mathbb{R}^N$ for which

$$x_1^* \geq x_2^* \geq \cdots \geq x_N^* \geq 0.$$

and there is a permutation $\pi : [N] \to [N]$ with $x_j^* = |x_{\pi(j)}|$ for all $j \in [N]$.

Proof. (of Proposition 1.5): $x^* \in \mathbb{R}_+^N$, the nonincreasing rearrangement of $x \in \mathbb{C}^N$, yields

$$\sigma_s(x)_q^q = \sum_{j=s+1}^{N} (x_j^*)^q \leq (x_s^*)^{q-p} \sum_{j=s+1}^{N} (x_j^*)^p \leq \left(\frac{1}{s}\sum_{j=1}^{s}(x_j^*)^p\right)^{\frac{q-p}{p}} \left(\sum_{j=s+1}^{N}(x_j^*)^p\right)$$

$$\leq \left(\frac{1}{s}||x||_p^p\right)^{\frac{q-p}{p}} ||x||_p^p = \frac{1}{s^{q/p-1}}||x||_p^q.$$

□

The next result is a stronger version, with optimal constant, of Proposition 1.5. Its importance relies on the fact that sharp results on error reconstruction analysis depends on it. Also, the method for proving it appears many times in the Compressive Sensing literature. We use optimization ideas to prove some inequalities with sharp constant. This technique will be invoked again in Lemma 4.25 and Lemma 5.21.

Theorem 1.7. *For any $q > p > 0$ and any $x \in \mathbb{C}^N$, the inequality*

$$\sigma_s(x)_q \leq \frac{c_{p,q}}{s^{1/p-1/q}}||x||_p$$

holds with

$$c_{p,q} = \left[\left(\frac{p}{q}\right)^{p/q}\left(1-\frac{p}{q}\right)^{1-p/q}\right]^{1/p} \leq 1.$$

Proof. Let $x^* \in \mathbb{R}_+^N$ be the nonincreasing rearrangement of $x \in \mathbb{C}^N$. If we set $\alpha_j = (x_j^*)^p$, we will transform the problem into a equivalent one given by

$$\begin{cases} \alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_N \geq 0 \\ \alpha_1 + \alpha_2 + \cdots + \alpha_N \leq 1 \end{cases} \implies a_{s+1}^{q/p} + a_{s+2}^{q/p} + \cdots + a_{s+N}^{q/p} \leq \frac{c_{p,q}^q}{s^{q/p-1}}.$$

Therefore, using $r = q/p > 1$, we need to maximize the convex function

$$f(\alpha_1, \alpha_2, \ldots, \alpha_N) = \alpha_{s+1}^r + \alpha_{s+s}^r + \cdots + \alpha_N^r,$$

over the convex polytope $C = \{(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N : \alpha_1 \geq \cdots \geq \alpha_N \geq 0 \text{ and } \alpha_1 + \cdots + \alpha_N \leq 1\}$. As any point of C is a convex combination of its vertices and because the function f is convex, the maximum is attained at a vertex of C (see Theorem B.16 from [Rauhut & Foucart '13] or Theorem 2.65 from [Ruszczynski]). Besides, the vertices are intersections of N hyperplanes when we force N of the $N+1$ inequalities to become equalities. From this we obtain the following possibilities:

1. If $\alpha_1 = \cdots = \alpha_N = 0$, then $f(\alpha_1, \alpha_2, \ldots, \alpha_N) = 0$.
2. If $\alpha_1 + \cdots + \alpha_N = 1$ and $\alpha_1 = \cdots = \alpha_k > \alpha_{k+1} = \cdots = \alpha_N = 0$ for some $1 \leq k \leq s$, then one has $f(\alpha_1, \alpha_2, \ldots, \alpha_N) = 0$.
3. If $\alpha_1 + \cdots + \alpha_N = 1$ and $\alpha_1 = \cdots = \alpha_k > \alpha_{k+1} = \cdots = \alpha_N = 0$ for some $s+1 \leq k \leq N$ then one has $\alpha_1 = \cdots = \alpha_k = 1/k$, and consequently $f(\alpha_1, \alpha_2, \ldots, \alpha_N) = (k-s)/k^r$.

Thus we have obtained

$$\max_{(\alpha_1,\ldots,\alpha_s)\in C} f(\alpha_1, \alpha_2, \ldots, \alpha_s) = \max_{s+1\leq k\leq N} \frac{k-s}{k^r}.$$

Now, consider $f(k)$ as a function of a continuous variable k, we observe that the function $g(k) = (k-s)/k^r$ is increasing until the critical point $k^* = (r/(r-1))s$ and decreasing thereafter. Using this information yields

$$\max_{(\alpha_1,\ldots,\alpha_s)\in C} f(\alpha_1, \alpha_2, \ldots, \alpha_s) \leq g(k^*) = \frac{1}{r}\left(1-\frac{1}{r}\right)^{r-1}\frac{1}{s^{r-1}} = \frac{c_{p,q}^q}{s^{q/p-1}}.$$

□

The other way of reasoning about compressible vectors is to require that the number of its significant components be small. This can be done with the introduction of the weak ℓ_p-spaces.

Definition 1.8. For $p > 0$, the weak ℓ_p space $w\ell_p^N$ denotes the space $\mathbb{C}^N$ equipped with the quasinorm

$$||x||_{p,\infty} = \inf\Big\{M \geq 0 : \#\{j \in [N] : |x_j| \geq t\} \leq \frac{M^p}{t^p} \text{for all } t > 0\Big\}.$$

Remark 5. *In the definition above, the quaisnorm comes from the fact that a constant appears in the triangular inequality:* $||x+y||_{p,\infty} \leq 2^{\max\{1,1/p\}}(||x||_{p,\infty} + ||y||_{p,\infty})$. *For definitions, see Section 1.7.*

There is an alternative expression for the weak ℓ_p-quasinorm of a vector $x \in \mathbb{C}^N$.

Proposition 1.9. *For $p > 0$, the weak ℓ_p-quasinorm of a vector $x \in \mathbb{C}^N$ can be expressed as*

$$||x||_{p,\infty} = \max_{k\in[N]} k^{1/p} x_k^*,$$

where $x^* \in \mathbb{R}_+^N$ denotes the noincreasing rearrangement of $x \in \mathbb{C}^N$.

Proof. Given $x \in \mathbb{C}^N$, we clearly have $||x||_{p,\infty} = ||x^*||_{p,\infty}$. Then we need to prove that $||x|| := \max_{k\in[N]} k^{1/p} x_k^*$ equals $||x^*||_{p,\infty}$. For $t > 0$ we have two possibilities:

$$\{j \in [N] : x_j^* \geq t\} = [k] \text{ for some } k \in [N] \qquad \text{or} \qquad \{j \in [N] : x_j^* \geq t\} = \emptyset.$$

In the first case, $t \leq x_j^* \leq ||x||/k^{1/p}$. This implies $\#\{j \in [N] : x_j^* \geq t\} = k \leq ||x||^p/t^p$. Note that this inequality holds trivially in the case that $\{j \in [N] : x_j^* \geq t\} = \emptyset$. By the definition of the weak ℓ_p-quasinorm, this leads to $||x^*||_{p,\infty} \leq ||x||$. Now, suppose that $||x|| > ||x^*||_{p,\infty}$, so that $||x|| \geq (1+\varepsilon)||x^*||_{p,\infty}$ for some $\varepsilon > 0$. This can be translate into $k^{1/p} x_k^* \geq (1+\varepsilon)||x^*||_{p,\infty}$ for some $k \in [N]$. Therefore we have the inclusion

$$[k] \subset \{j \in [N] : x_j^* \geq (1+\varepsilon)||x^*||_{p,\infty}/k^{1/p}\}.$$

Again, by the definition of the weak ℓ_p-quasinorm, we obtain

$$k \leq \frac{||x^*||_{p,\infty}^p}{\big((1+\varepsilon)||x^*||_{p,\infty}/k^{1/p}\big)^p} = \frac{k}{(1+\varepsilon)^p},$$

which is a contradiction. We conclude that $||x|| = ||x^*||_{p,\infty}$.

□

With the alternative expression of the weak ℓ_p-quasinorm, we can establish a proposition similar to Proposition 1.5.

Proposition 1.10. *For any $q > p > 0$ and $x \in \mathbb{C}^N$, we have the inequality*

$$\sigma_s(x)_q \leq \left(\frac{p}{q-p}\right) \frac{1}{s^{1/p-1/q}} ||x||_{p,\infty}.$$

Proof. Without loss of generality, we assume that $||x||_{p,\infty} \leq 1$, so that $x_k^* \leq 1/k^{1/p}$ for all $k \in [N]$. This yields

$$\sigma_s(x)_p^p = \sum_{k=s+1}^{N} (x_k^*)^q \leq \sum_{k=s+!}^{N} \frac{1}{k^{q/p}} \leq \int_s^\infty \frac{1}{t^{q/p}} dt = -\frac{1}{q/p-1} \frac{1}{t^{q/p-1}}\bigg|_{t=s}^{t=N} \leq \frac{p}{q-p} \frac{1}{s^{q/p-1}}.$$

Taking the power $1/q$ leads to the desired inequality. □

Proposition 1.5 and Proposition 1.10 show that if we have vectors $x \in \mathbb{C}^N$ for which $||x||_p \leq 1$ or $||x||_{p,\infty} \leq 1$ for small $p > 0$, than they will be compressible vectors in the sense that their errors of best s-term approximation decay quickly with s.

1.5 How Many Measurements Are Necessary?

The fundamental problem in Compressive Sensing is to reconstruct an s-sparse vectors from few measurements. In mathematical terms, this is represented by solving the linear system $Ax = y$, where $x \in \mathbb{C}^N$ is the s-sparse signal being acquired, $A \in \mathbb{C}^{m\times N}$ represents the measurement matrix and $y \in \mathbb{C}^m$ represents the acquired information.

Since we want to acquire the signal in a compressible fashion, we have fewer measurements than the ambient dimension, i.e., $m < N$. This results in an underdetermined system, which has, without any other assumption, an infinite number of solutions. As described in Section 1.3, it is natural to assume that the signals are sparse in some basis. With this regularization assumption, we expect to identify the original vector x. Thus, the recovery of sparse signals has two meanings:

1. *Uniform recovery*: the reconstruction of *all* s-sparse vectors $x \in \mathbb{C}^N$ simultaneously.

2. *Nonuniform recovery*: the reconstruction of an *specific* s-sparse vector $x \in \mathbb{C}^N$.

In case we are dealing with measurements matrices described by stochastic models, we can translate conditions above by finding a lower bound of the form

1. *Uniform recovery*: $\mathbb{P}(\forall s\text{-sparse } x, \text{ recovery of } x \text{ is successful using } A) \geq 1 - \varepsilon$.

2. *Nonuniform recovery*: $\forall$ s-sparse, $\mathbb{P}(\text{recovery of } x \text{ is successful using } A) \geq 1 - \varepsilon$.

In both cases, the probability is over the random draw of A, described by a certain model. In this dissertation we will deal with the first case. We will answer when it is possible to recover all signals from few measurements, provided that all of them have some sparsity and that we have access to the acquired vector y (this acquired vector will be different for every signal x). For the nonuniform case, see Sections 12.2 and 14.2 of [Rauhut & Foucart '13] and the discussion in [Candès & Plan '11].

Remark 6. *For a matrix $A \in \mathbb{C}^{m\times N}$ and a subset $S \subset [N]$, we denote by A_S the column submatrix of A consisting of the columns indexed by S. For a vector $x \in \mathbb{C}^N$, we denote by x_S the vector in $\mathbb{C}^S$ consisting of the entries indexed by S or the vector in $\mathbb{C}^N$ which coincides with x on the entries in S and is zero on the entries outside S. It will be clear from the context which one of the two options we are dealing with.*

The way we recover signals is by supposing that they are as sparse as possible. Therefore, an algorithmic approach for recovering them is through ℓ_0-minimization. In other words, we search for the sparsest vector consistent with the measured data $y = Ax$.

$$\min_{z\in\mathbb{C}^N} ||z||_0 \qquad \text{subject to } Az = y. \tag{P_0}$$

In case that measurements are corrupted by noise or are slightly inaccurate, the natural generalization is given by

$$\min_{z\in\mathbb{C}^N} ||z||_0 \qquad \text{subject to } ||Az - y||_2 \leq \eta. \tag{$P_{0,\eta}$}$$

The main point about Compressive Sensing is that we can recover compressible signals corrupted by noise. Then, this problem has stability and robustness. See Sections 3.4 and 3.5. Therefore, in real situations, we will deal with the problem $(P_{0,\eta})$. We can understand the recovery of sparse vectors through the properties of the measurement matrix A as shown in the next theorem.

Theorem 1.11. *Given $A \in \mathbb{C}^{m\times N}$, the following properties are equivalent*

a) Every s-sparse vector $x \in \mathbb{C}^N$ is the unique s-sparse solution of $Az = Ax$, that is, if $Ax = Az$ and both x and z are s-sparse, then $x = z$.

b) The null space $\ker A$ *does not contain any* $2s$*-sparse vector other than the zero vector, that is,* $\ker A \cap \{z \in \mathbb{C}^N : ||z||_0 \leq 2s\} = \{0\}$.

c) Every set of $2s$ *columns of* A *is linearly independent.*

d) For every $S \subset [N]$ *with* $\#S \leq 2s$*, the submatrix* A_S *is an injective map from* $\mathbb{C}^S$ *to* $\mathbb{C}^m$ *and the Gram matrix* $A_S^* A_S$ *is invertible.*

Proof. *a)* $\implies$ *b)*: Assume that for every vector $x \in \mathbb{C}^N$, we have $\{z \in \mathbb{C}^N : Az = Ax, ||z||_0 \leq s\} = \{x\}$. Let $v \in \ker A$ be $2s$-sparse. We write $v = x - z$ for s-sparse vectors x, z with $\operatorname{supp} x \cap \operatorname{supp} z = \emptyset$. Then $Ax = Az$ and by assumption $x = z$. Since the supports of x and z are disjoint, it follows that $z = x = 0$ and $v = 0$.

b) $\implies$ *c)*: Let S be a set of indexes $1 \leq i_1 < i_2 < \cdots < i_{2s} \leq N$ with $\#S = 2s$ and let $x \in \mathbb{C}^N$ such that $\operatorname{supp}(x) \subset S$. If $Ax = \sum_{\ell=1}^{2s} a_{i_\ell} x_{i_\ell} = 0$, then $x = 0$. Thus the $2s$ columns vectors indexed by S must be linearly independent.

c) $\implies$ *d)*: If every set of $2s$ columns of A is linearly independent then every set of s columns is also linearly independent. Thus, clearly, A_S is injective as a map from $\mathbb{C}^S$ to $\mathbb{C}^m$. Also, consider S and x as in the previous case. We have

$$\langle x, A_S^* A_S x \rangle = \langle A_S x, A_S x \rangle = \left|\left| \sum_{s,t=1}^{2s} x_{i_s} a_{i_s} \right|\right|_2^2 > 0,$$

since, by c), the $2s$ column vectors a_{i_s} are linearly independent. Also, the matrix $A_S^* A_S$ is self-adjoint, therefore all of its eigenvalues are real and they will be positive if and only if the quadratic form $\langle x, A_S^* A_S x \rangle$ is positive. Furthermore, a matrix with all positive eigenvalues is invertible.

d) $\implies$ *a)*: Suppose that x_1 and x_2 are two s-sparse vectors satisfying $Ax_1 = Ax_2 = Az$. Setting $x = x_1 - x_2$ we have that it is $2s$-sparse and also that $Ax = Ax_1 - Ax_2 = Az - Az = 0$. Let S be the index set $\{i_s\}$ for the support of x. Since x is in the kernel of A, we have

$$\langle x, A_S^* A_S x \rangle = \left|\left| \sum_{s,t=1}^{2s} x_{i_s} a_{i_s} \right|\right|_2^2.$$

Since $A_S^* A_S$ in invertible, the last sum vanishes only if $x = 0$. This implies $x_1 = x_2$. □

From Theorem 1.11, we conclude that if it is possible to reconstruct *every* s-sparse, so that the statement c) above holds, then we have $\operatorname{rank}(A) \geq 2s$. On the other hand, since $A \in \mathbb{C}^{m \times N}$, $\operatorname{rank}(A) \leq m$. We conclude that the number of measurements necessary to reconstruct every s-sparse vector satisfies $m \geq 2s$. The next theorem shows that $m = 2s$ measurements suffice. After, we will discuss why this bound is a theoretical one and typically not stable for practical implementation. In particular, due to a lack of knowledge about the location of the support, in Chapter 8 we will see that, in fact, a few more than $2s$ measurements will be necessary.

Theorem 1.12. *For any integer* $N \geq 2s$*, there exists a measurements matrix* $A \in \mathbb{C}^{m \times N}$ *with* $m = 2s$ *rows such that every* s*-sparse vector* $x \in \mathbb{C}^N$ *can be recovered from its measurement vector* $y = Ax \in \mathbb{C}^m$ *as a solution of* (P_0).

Proof. For fixed $t_n > \cdots > t_2 > t_1 > 0$, we can define the matrix $A \in \mathbb{C}^{m \times N}$ with $m = 2s$ as

$$A = \begin{bmatrix} 1 & 1 & \dots & 1 \\ t_1 & t_2 & \dots & t_N \\ \vdots & \vdots & \dots & \vdots \\ t_1^{2s-1} & t_2^{2s-1} & \dots & t_N^{2s-1} \end{bmatrix}.$$

Let $S = \{j_1 < \cdots < j_{2s}\}$ be an index set of cardinality $2s$. Define the square matrix $A_S \in \mathbb{C}^{2s \times 2s}$. It will be the transpose of a Vandermonde matrix. It is well known that $\det(A_S) = \prod_{k<\ell}(t_{j_\ell} - t_{j_k}) > 0$. This

shows that A_S is invertible and, in particular, injective. The hypotheses of Theorem 1.11 are fulfilled, thus every s-sparse vector $x \in \mathbb{C}^N$ is the unique s-sparse vector satisfying $Az = Ax$. Therefore it can be recovered as the solution of (P_0).

□

Remark 7. *As discussed in Section 2.2 of [Rauhut & Foucart '13], it is interesting to note that many matrices fulfill the hypotheses of Theorem* 1.11. *We can take any matrix $M \in \mathbb{R}^{N\times N}$ that is totally positive, i.e., that satisfies $\det(M_{I,J}) > 0$ for any sets $I, J, \subset [N]$ with same cardinality, where $M_{I,J}$ represents the submatrix of M with rows indexed by I and columns indexed by J. After, we select any $m = 2s$ rows of M, indexed by a set I, to form the measurement matrix A.*
Since the original matrix M is totally positive, for any index $S \subset [N]$ with $\#S = 2s$, the matrix A_S is the same as $M_{I,S}$, hence it is invertible and we can apply Theorem 1.11. *In particular, the partial Discrete Fourier Transform*

$$A = \begin{bmatrix} 1 & 1 & 1 & 1\dots & 1 \\ 1 & e^{2\pi i/N} & e^{2\pi i 2/N} & \dots & e^{2\pi i(N-1)/N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & e^{2\pi i(2s-1)/N} & e^{2\pi i(2s-1)2/N} & \dots & e^{2\pi i(2s-1)(N-1)/N} \end{bmatrix},$$

allows for the reconstruction of every s-sparse vector $x \in \mathbb{C}^N$ from $y = Ax \in \mathbb{C}^{2s}$.

The reasoning above can be generalized and we can prove that, from a measure-theoretical viewpoint, the set of matrices that can not recovery s-sparse vector is small.

Theorem 1.13. *The set of $2s \times N$ matrices such that $\det(A_S) = 0$ for some $S \subset [N]$ with $\#S \leq 2s$ has Lebesgue measure zero. Therefore, most $2s \times N$ matrices allow the reconstruction of every s-sparse vector $x \in \mathbb{C}^N$ from $y = Ax \in \mathbb{C}^{2s}$.*

Theorem 1.13 is a simples consequence of Sard's Theorem[3] and tells us that we can draw a matrix at random and it will be, in principle, a good measurement matrix. However, we do not address the problem of robustness to error measurements and stability for compressible vectors recovery. These points will be further discussed in Chapter 3 and Chapter 8. In the latter, we will prove that *any* stable reconstruction for s-sparse vectors requires at least $m = Cs\ln(eN/s)$ linear measurements, for a constant $C > 0$. From this perspective, matrices such as the Vandermonde matrix from Theorem 1.12 are not adequate as measurement matrices.

1.6 Computational Complexity of Sparse Recovery

In this section we focus on a brief digression about computational complexity theory. This is a branch of mathematics/computer science where one tries to quantify the amount of computational resources required to solve a given task. We do not intend to be rigorous here. For a rigorous treatment, see the classical [Garey & Johnson '79] or the modern [Arora & Barak '07].

For example, if one tries to solve a linear system, the classic Gaussian elimination algorithm uses essentially n^3 basic arithmetic operations to solve n equations over n variables. However, in 1960s, a more efficient algorithm was created, see [Strassen '69]. The latter is a highly nonintuitive algorithm.

Therefore one might ask if, in some cases, there are more efficient (but nonintuitive) algorithms than the ones used for many years. In the field of computational complexity, researchers try to answer such questions and prove that, sometimes, there is no better algorithm than the existing ones. Even more, for some problems they try to prove that *there is no* efficient algorithm.

Computational complexity starts with the work [Hartmanis & Stearns '65]. For an historical account, see [Fortnow & Homer '03]. It is philosophically similar to taxonomy, where typically one tries to establish patterns and group problems into classes according to its inherent difficulty, in the same way naturalists

[3]For a proof of this theorem, see pages 205-207 of [Guillemin & Pollack].

try to classify the living beings according to physiological properties. After, they relate those classes to each other. In this classification, a problem is regarded as inherently difficult if its solution requires significant resources, whatever the algorithm used.

A polynomial/exponential-time algorithm is an algorithm performing its tasks in a number of steps bounded by a polynomial/exponential expression in the size of the input. The first distinction between polynomial-time and exponential-time algorithm was given by [von Neumann '53]. It is standard in complexity theory to identify polynomial-time with feasible. Clearly, this is not always true, as pointed by [Cook]. He says: *"for example, a computer program requiring n^{100} steps could never be executed on an input even as small as $n = 10$."* Even so, let us assume that this is the case and let us consider the *Feasibility Thesis*:

"A natural problem has a feasible algorithm if and only if it has a polynomial-time algorithm."

There is a classification of computational problems: they are divided in decision problems, search problems, counting problems, optimization problems and function problems. See [Goldreich '08]. Here we are interested in decision and optimization problems. A decision problem is a question in some formal system with a yes-or-no answer, depending on the values of some input parameters. An optimization problem asks to find the "best possible" solution among the set of all feasible solutions. Every optimization problem can be transformed into a decision problem just by asking if the best possible solution exists or not. Now we introduce some terminology following [Rauhut & Foucart '13]. For rigorous definitions, see Definitions 1.13, 2.1 and 2.7 from [Arora & Barak '07].

Definition 1.14. We have four basic complexity classes:

- The class $\mathcal{P}$ of P-problems consists of all decision problems for which there exists a polynomial-time algorithm finding a solution.
- The class $\mathcal{NP}$ of NP-problems consists of all decision problems for which there exists a polynomial-time algorithm certifying a solution.
- The class *$\mathcal{NP}$-hard* of NP-hard problems consist of all problems (not necessarily decision problems) for which a solving algorithm could be transformed in polynomial time into a polynomial time solving algorithm for any NP-problem. Roughly speaking, this is the class of problems at least as hard as any NP-problem.
- The class *$\mathcal{NP}$-complete* of NP-complete problems consist of all problems that are both NP and NP-hard; in other words, it consists of all the NP-problems at least as hard as any other NP-problem.

Note that the class $\mathcal{P}$ is contained in the class $\mathcal{NP}$. One can ask if the converse is also true. This is the important conjecture $\mathcal{P} = \mathcal{NP}$, see [Cook]. If this conjecture is shown to be false, which is widely believed, then there will exist problems for which potential solutions can be certified, but no solution can be found in polynomial time.

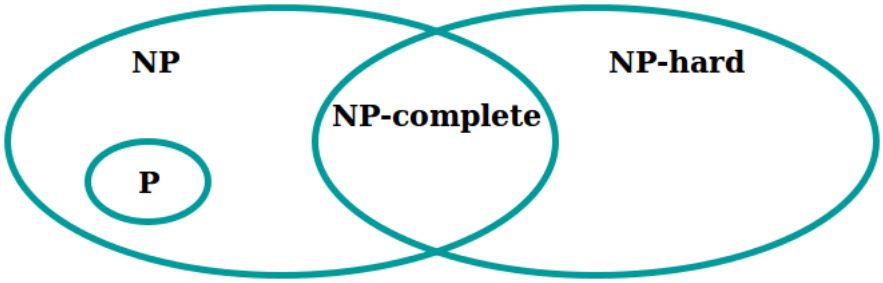


Figure 1.2: Relation between the four main complexity classes.

As [Arora & Barak '07] points out, *Recognizing the correctness of an answer is often much easier than coming up with the answer. Appreciating a Beethoven sonata is far easier than composing the sonata; verifying the solidity of a design for a suspension bridge is easier (to a civil engineer anyway!) than coming up with a good design; verifying the proof of a theorem is easier than coming up with a proof itself.*

Since nowadays we know that there are many $\mathcal{NP}$*-complete* problems, these problems are probably intrinsically intractable. Some of the first examples were given by a list of 21 problems described in [Karp '72]. He says *"In this paper we give theorems that suggest, but do not imply, that these problems, as well as many others, will remain intractable perpetually"*. In particular, we are interested in the fourteenth problem of the list.

Remark 8. *(Exact cover by 3-sets problem or X3C problem): Given a collection $\{C_i, i \in [N]\}$ of 3-element subsets of $[m]$, does there exist an exact cover (or a partition) of $[m]$, i.e., a set $J \subset [N]$ such that $\cup_{j\in J} C_j = [m]$ and $C_j \cap C_k = \emptyset$ for all $j, k \in J$ with $j \neq k$?*

Clearly we see that m must be a multiple of 3. Let us use one example in order to understand. Suppose we have $[m]$, i.e., $[m] = \{1, 2, 3, 4, 5, 6\}$. If we have the collection of 3-sets elements given by $\{\{1, 2, 3\}, \{2, 3, 4\}, \{1, 2, 5\}, \{2, 5, 6\}, \{1, 5, 6\}\} = \{C_1, C_2, C_3, C_4, C_5\}$ then we could choose $\{C_2, C_5\} = \{\{2, 3, 4\}, \{1, 5, 6\}\}$ as an exact cover because each element in $[m]$ appears exactly once.

If instead, the collection of 3-sets was given by $\{\{1, 2, 3\}, \{2, 4, 5\}, \{2, 5, 6\}\}$, then any subcover from this that we choose will not be an exact cover (we need all 3 subsets to cover all elements in $[m]$ at least once, but then the element 2 appears three times).

In Section 1.5 we stated the main problem of Compressive Sensing: the optimization problem of sparse recovery. The straightforward approach for solving it is to solve every square linear system given by $A_S^* A_S u = A_S^* y$ for $u \in \mathbb{C}^S$ where S runs though all possible subsets of $[N]$ with size s. In this brute force approach, one needs to solve a number of linear systems of order $\binom{N}{s}$.

Example 1.15. As described in Section 2.3 of [Rauhut & Foucart '13], suppose that we have a small size problem of sparse recovery with $N = 1000$ and $s = 10$. We would have to solve $\binom{1000}{10} \geq (\frac{1000}{10})^{10} = 10^{20}$ linear systems of size 10×10. If each linear system require 10^{-10} seconds to be solved, the time required for solving (P_0) will be 10^{10} seconds, i.e., more than 300 years.

The heuristic of Example 1.15 will be confirmed by Theorem 1.16, i.e., this problem is in fact intractable for any possible approach. We will prove that (P_0) belongs to the NP-hard class, i.e., assuming that the exact cover by 3-sets problem is NP-complete, we can now prove that the main Compressive Sensing problem is as hard as this one.

Theorem 1.16. (Theorem 1 of [Natarajan '96]): *For any $\eta \geq 0$, the ℓ_0-minimization problem $(P_{0,\eta})$ for general $A \in \mathbb{C}^{m\times N}$ and $y \in \mathbb{C}^m$ is NP-hard.*

Proof. Without loss of generality, we may assume that $\eta < 1$, by a rescaling argument. Using that the *exact cover by 3-sets problem* is NP-complete, we will reduce it, in polynomial time, to our ℓ_0-minimization problem. Let us consider $\{C_i : i \in [N]\}$, the collection of 3-element subsets of $[m]$. Using this collection, we define vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n \in \mathbb{C}^m$ by

$$(\boldsymbol{a}_i)_j = \begin{cases} 1 & \text{if } j \in C_i, \\ 0 & \text{if } j \notin C_i. \end{cases}$$

With these vectors, we can define a matrix $A \in \mathbb{C}^{m\times N}$ and a vector $y \in \mathbb{C}^m$ by

$$A = \left[\begin{array}{c|c|c|c} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \ldots & \boldsymbol{a}_N \end{array}\right], \qquad y = [1, 1, \ldots, 1].$$

It is possible to make this construction in polynomial time since, by construction, we have $N \leq \binom{m}{3}$. Using the vector y defined above, for any $z \in \mathbb{C}^N$ satisfying $||Az - y|| \leq \eta$, we have that any of its

components must be distant to 1 at most by η. Hence, these components are nonzero and $||Az||_0 = m$. On the other hand, by definition, each $\boldsymbol{a}_i$ has exactly 3 nonzero components and then $Az = \sum_{j=1}^{N} z_j \boldsymbol{a}_j$ has at most $3||z||_0$ nonzero components, i.e., $||Az||_0 \leq 3||z||_0$.
Therefore, we conclude that if a vector satisfies $||Az - y||_2 \leq \eta$ then it must satisfy $||z||_0 \geq m/3$. Suppose that we ran the ℓ_0-problem and that it returned $x \in \mathbb{C}^N$ as the output. We have two possibilities.

1. If $||x||_0 = m/3$, then suppose the collection $\{C_j, j \in \text{supp}(x)\}$ is not an exact cover of $[m]$. In this case, the m components of $Ax = \sum_{j=1}^{N} x_j \boldsymbol{a}_j$ would not be all be nonzero. Since this is impossible, we conclude that in this case the exact cover exists. Therefore we can answer positively the question about the existence of an exact cover of $[m]$ using the solution of $(P_{0,\eta})$.

2. If $||x||_0 > m/3$, suppose that an exact cover of $[m]$ exists, i.e., we can cover the set $[m]$ with a partition $\{C_j, j \in J\}$. In this case, we can define a vector $z \in \mathbb{C}^N$ by $z_j = 1$ if $j \in J$ and $z_j = 0$ if $z_j \notin J$. This vector satisfies $Az = \sum_{j=1}^{N} z_j \boldsymbol{a}_j = \sum_{j \in J} \boldsymbol{a}_j = y$ and $||z||_0 = m/3$. This is a contradiction, since in this case x would be no more the solution of $(P_{0,\eta})$. Thus, in the case $||x||_0 > m/3$, we can answer negatively about the existence of an exact cover by using the solution of $(P_{0,\eta})$.

In both cases, we showed that through the solution of the ℓ_0-minimization problem we can solve the exact cover by the 3-sets problem.

□

It is important to highlight what this theorem is and what it is not. It concerns the intractability in the *general* case. This means that for a general matrix A and a general vector y, we cannot have a smart strategy to solve (P_0), in other words, no algorithm is able to solve the problem for *any* choice of A and y. This does not mean that for special choices of A and y we do not have tractable algorithms for sparse recovery. This will be the content of Chapter 2. Also, as we shall see in Theorem 3.30, that ℓ_q-minimization, for $q < 1$, is also an NP-hard problem.

1.7 Some Definitions

In this section we give a *pot-pourri* of definitions used along this dissertation.

Definition 1.17. A nonnegative function $||.|| : X \to [0, \infty)$ on a vector space X is called a norm if

(a) $||x|| = 0$ if and only if $x = 0$.
(b) $||\lambda x|| = |\lambda|||x||$ for all scalars λ and all vectors $x \in X$.
(c) $||x + y|| \leq ||x|| + ||y||$.

If only (b) and (c) hold,so that $||x|| = 0$ does not necessarily imply $x = 0$, then $||.||$ is called a *seminorm*. If (a) and (b) hold, but (c) is replacedby the weaker quasitriangular inequality

$$||x + y|| \leq C(||x|| + ||y||)$$

for some $C \geq 1$, then $||.||$ is called a *quasinorm*. The smallest constant C is called its quasinorm constant.

Definition 1.18. Let $||.||$ be a norm on $\mathbb{R}^n$ or $\mathbb{C}^n$. Its dual norm $||.||_*$ is defined by

$$||x||_* = \sup_{||y|| \leq 1} |\langle x, y \rangle|.$$

In the real case we have

$$||x||_* = \sup_{y \in \mathbb{R}^n, ||y|| \leq 1} \langle x, y \rangle.$$

In the complex case, we have

$$||x||_* = \sup_{y \in \mathbb{R}^n, ||y|| \leq 1} \text{Re}\langle x, y \rangle.$$

The dual of the dual norm $||.||_*$ is the norm $||.||$ itself. In particular, we have

$$||x|| = \sup_{||y||_* \leq 1} |\langle x, y \rangle| = \sup_{||y||_* \leq 1} \text{Re}\langle x, y \rangle$$

Definition 1.19. Let $A : X \to Y$ be a linear map between two normed vector spaces $X, ||.||$ and $(Y, |||.|||)$ with possible different norms. The operator norm of A is defined as

$$||A|| = \sup_{||x|| \leq 1} |||Ax||| = \sup_{||x||=1} |||Ax|||.$$

In particular, for a matrix $A \in \mathbb{C}^{m \times n}$ and $1 \leq p, q \leq \infty$, we define the matrix norm (or operator norm) between ℓ_p^n and ℓ_q^m as

$$||A||_{p \to q} = \sup_{||x||_p \leq 1} ||Ax||_q = \sup_{||x||_p = 1} ||Ax||_q.$$

Definition 1.20. We say that two functions $f, g \in \mathbb{R}$ are comparable if there exists absolute constants $c_1, c_2 > 0$ such that $c_1 f(t) \leq g(t) \leq c_2 A f(t)$ for all t. We denote it by $f \asymp g$.

We have a notational convention for asymptotic analysis. It was introduced by Paul Bachmann in 1894 and popularized in subsequent years by Edmund Landau in the field of Number Theory.

Definition 1.21. i.) (Big O): We say that $f(x) = O(g(x))$ if there exists a constant $C > 0$ such that

$$|f(x)| \leq C|g(x)| \qquad \text{for all } x. \tag{1.2}$$

and when $O(g(x))$ stands in the middle of a formula it represents a function $f(x)$ that satisfies Equation (1.2). This notation is typically used in the asymptotic context, specially when x is not integer. See Chapter 9 of [Graham, Knuth & Patashnik '94].

ii.) (Big Omega) : For lower bounds, we say that $f(x) = \Omega(g(x))$ if there exists a constant $C > 0$ such that

$$|f(x)| \geq C|g(x)| \qquad \text{for all } x. \tag{1.3}$$

We have $f(x) = \Omega(g(x))$ if and only if $g(x) = O(f(x))$.

iii.) (Big Theta): To specify an exact order of growth, we have the Big Theta notation. We denote by $f(x) = \Theta(g(x))$ if the following statements hold

$$f(x) = O(g(x)) \qquad \text{and} \qquad f(x) = \Omega(g(x)). \tag{1.4}$$

Big Theta definition equivalent to Definition 1.20. We emphasize it here due to the fact that Computational Complexity community uses it while Approximation Theory community uses $f \asymp g$. Both of them will be used along this text.

Chapter 2

Some Algorithms and Ideas

All models are wrong, some are useful.
George E. P. Box in [Box '79].

2.1 Introduction

For general measurement matrices A and vectors y, the problem (P_0) of recovering sparse vectors is NP-hard, as Theorem 1.16 shows. Then there is no hope of solving it in a computationally efficient way. However, in the last twenty years a lot of efficient algorithms were developed to deal with this problem. Techniques from Convex Optimization and from Approximation Theory were combined in order to find heuristics that solve the problem, at least for particular instances.

The early development of Compressive Sensing was based on the assumption that the solution to the ℓ_1-minimization problem provides the recovery of the correct sparse vector, serving as a proxy to (P_0), and also that it is feasible to solve this problem with a computer. We will see why this is true but also that a lot of work has been done in order to find alternative algorithms that are faster or give superior reconstruction performance in some situations.

The purpose of this chapter is to present the three most popular classes of algorithms used for sparse vectors recovery: optimization methods, greedy methods and thresholding methods. Here we will not analyze any of them. This will be postponed to later chapters, after we introduce the Coherence Property and the Restricted Isometry Property. We do not intend to be exhaustive in the description of these classes of algorithms and our goal now is just to give some intuitive explanation and to develop some ideas about the computational aspects of Compressive Sensing.

After presenting the three major classes of algorithms in Sections 2.2 to 2.4, we will briefly discuss the problem of choosing an algorithm for a particular application in Section 2.5, along with a review of recent numerical work related to this issue.

2.2 Basis Pursuit

Basis Pursuit is the main strategy for solving P_0 we will deal with along this dissertation. It was introduced by [Chen '95] as part of his Ph.D. thesis and was fully explored by [Chen, Donoho & Saunders '01]. Since then, it has been widely used by the Signal Processing, Statistics and Machine Learning communities. As this strategy belongs to the optimization methods family, we need to start by defining what is an optimization problem. It is not our intention to be encyclopedic because Mathematical Programming is a vast and deep subject, see [Ruszczynski], [Boyd & Vanderberghe '04] or [Bertsekas '16][1] for further

[1] There is a whole volume of Documenta Mathematica dedicated to the history of Optimization: `https://www.math.uni-bielefeld.de/documenta/vol-ismp/vol-ismp.html` for some references. Also, the book [Gass & Assad '04] has some interesting historical notes.

information. For the use in Machine Learning of Optimization ideas related to Compressive Sensing, see [Sra, Nowozin & Wright '11].

Definition 2.1. An optimization problem is described by

$$\text{minimize } F_0(x) \qquad \text{subject to } F_i(x) \leq b_i,\ i = 1, 2, \ldots \tag{2.1}$$

where the function $F_0 : \mathbb{R}^N \to (-\infty, \infty]$ is the *objective function* and the functions $F_1, \ldots, F_n : \mathbb{R}^N \to (-\infty, \infty]$ are the *constraint functions*. A point $x \in \mathbb{R}^N$ is called *feasible* if it satisfies the constraints. A feasible point $x^\#$ for which the minimum is attained, that is, $F_0(x^\#) \leq F_0(x)$ for all feasible point x is called an *optimal point* and $F_0(x^\#)$, an optimal value. If $F_0, F_1, \ldots, F_n$ are all convex (linear) functions, this is called a *convex (linear) optimization problem.*

Remark 9. *This framework, only with inequality constraints, is the most general one and encompasses equality constraints too. Every time we have some constraint of the form $F_i(x) = c_i$, we can write the equivalent inequalities $F_i(x) \leq c_i$ and $-F_i(x) \leq -c_i$.*

In our setting, the measurement constraint $Ax = y$ must be satisfied, therefore some of the inequalities will be given by $F_j(x) := \langle A_j, x\rangle \leq y_j$ and $-F_j(x) := -\langle A_j, x\rangle \leq -y_j$, where $A_j \in \mathbb{R}^N$ is the jth row of A. Here, we will distinguish the measurement constraint as a *special* constraint. Also, when modeling some signals of interest, other constraints $F_j(x)$ may appear, thus we will describe the set of feasible points by

$$K = \{x \in \mathbb{R}^N \mid Ax = y,\ F_j(x) \leq b_j,\ j \in [M]\}. \tag{2.2}$$

Therefore, we are interested in problems of the form

$$\min_{x \in K} F_0(x). \tag{2.3}$$

Another important class of optimization problems is the class of *conic problem*. This kind of problem will appear in the context of complex vectors recovery. Here is the formal definition:

Definition 2.2. A conic optimization problem is of the form

$$\min_{x \in \mathbb{R}^N} F_0(x) \qquad \text{subject to } x \in K \text{ and } F_i(x) \leq b_i,\ i \in [n],$$

where K is a convex cone and all F_i are convex functions. If $K = \left\{x \in \mathbb{R}^{N+1} : \sqrt{\sum_{j=1}^N x_j^2} \leq x_{N+1}\right\}$, then we have a *second-order cone problem* and if K is the cone of positive semidefinite matrices, we have a *semidefinite program.*

The study of many optimization problems is carried on through the notion of duality. As sometimes it is hard to find the minimum of a problem, we transform it into another (hopefully easier) problem and maximize the new one. This new problem, called the *dual problem*, will provide a bound for the optimal value of the original problem, called *primal problem.* If we are lucky enough, the solution of both problems will agree. In order to describe the dual problem, we need another definition.

Definition 2.3. The *Lagrange function* of an optimization problem given by (2.2) is defined for $x \in \mathbb{R}^N, \xi \in \mathbb{R}^m, v \in \mathbb{R}^M$ with $v_\ell \geq 0$ for all $\ell \in [M]$, by

$$L(x, \xi, v) = F_0(x) + \langle \xi, Ax - y\rangle + \sum_{\ell=1}^M v_\ell (F_\ell(x) - b_\ell).$$

The variables ξ and v are called *Lagrange Multipliers.* Here, we will denote that $v_\ell \geq 0$ for all $\ell \in [M]$ by $v \succcurlyeq 0$. The *Lagrange dual function* is defined by

$$H(\xi, v) = \inf_{x \in \mathbb{R}^N} L(x, \xi, v), \qquad \xi \in \mathbb{R}^m, v \in \mathbb{R}^M, v \succcurlyeq 0.$$

If $x \mapsto L(x, \xi, v)$ is unbounded from below, we set $H(\xi, v) = -\infty$. Also, if there are no inequality constraints, then

$$H(\xi) = \inf_{x \in \mathbb{R}^N} L(x, \xi) = \inf_{x \in \mathbb{R}^N} \{F_0(x) + \langle \xi, Ax - y \rangle\}, \ \xi \in \mathbb{R}^m. \tag{2.4}$$

The dual function provides a bound on the optimal value of $F_0(x^\#)$ for the minimization problem describe by the feasible set (2.2), that is, $H(\xi, v) \leq F_0(x^\#)$ for all $\xi \in \mathbb{R}^m, v \in \mathbb{R}^M, v \succcurlyeq 0$. To see why, taking x as a feasible point for (2.3), then $Ax - y = 0$ and $F_\ell(x) - b_\ell \leq 0$ for all $\ell \in [M]$, so we have, for all $\xi \in \mathbb{R}^m$ and $v \succcurlyeq 0$, that

$$\langle \xi, Ax - y \rangle + \sum_{\ell=1}^{M} v_\ell (F_\ell(x) - b_\ell) \leq 0.$$

And then

$$L(x, \xi, v) = F_0(x) + \langle \xi, Ax - y \rangle + \sum_{\ell=1}^{M} v_\ell (F_\ell(x) - b_\ell) \leq F_0(x).$$

Using that $H(\xi, v) \leq L(x, \xi, v)$ and taking the minimum over all feasible points $x \in \mathbb{R}^N$, we obtain a lower bound for the primal problem. This leads to the new optimization problem

$$\max_{\xi \in \mathbb{R}^m, \ v \in \mathbb{R}^M} H(\xi, v) \qquad \text{subject to } v \succcurlyeq 0. \tag{2.5}$$

The main point of this transformation is that $H(\xi, v)$ is always concave, even if the original function is not convex. Therefore the dual problem is equivalent to minimizing $-H$, a convex function, subject to $v \succcurlyeq 0$. A feasible maximizer of (2.5) is called a *dual optimal*. We always have $H(\xi^\#, v^\#) \leq F(x^\#)$ and this is what we call *weak duality*. In case of equality, we have *strong duality*. We can interpret Lagrange duality via a saddle-point argument.

Remark 10. *(Saddle-point interpretation): We will consider here the problem*

$$\textit{minimize } F_0(x) \qquad \textit{subject to } Ax = y, \tag{2.6}$$

that is, for the sake of simplicity we will consider the original problem without inequality constraints. The general case with inequality constraints (or for conic programs) has a similar derivation. Using the definition of Lagrange function yields

$$\sup_{\xi \in \mathbb{R}^m} L(x, \xi) = \sup_{\xi \in \mathbb{R}^m} F_0(x) + \langle \xi, Ax - y \rangle = \begin{cases} F_0(x) & \textit{if } Ax = y, \\ \infty & \textit{otherwise.} \end{cases}$$

Therefore, is x is not feasible, than the above supremum is infinite. On one hand, a feasible minimizer of the primal problem will satisfies $F_0(x^\#) = \inf_{x \in \mathbb{R}^N} \sup_{\xi \in \mathbb{R}^m} L(x, \xi)$ On the other hand, a dual optimal satisfies $H(\xi^\#) = \sup_{\xi \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^N} L(x, \xi)$ by definition of the Lagrange dual function. Hence, weak duality reads as

$$\sup_{\xi \in \mathbb{R}^m} \inf_{x \in \mathbb{R}^N} L(x, \xi) \leq \inf_{x \in \mathbb{R}^N} \sup_{\xi \in \mathbb{R}^m} L(x, \xi),$$

whereas the same reasoning holds for strong duality with equality instead of inequality. This tells us that we can change maximization and minimization if strong duality holds. This is the ***saddle-point property****. This is the same as says that for a primal-dual optimal pair $(x^\#, \xi\#)$ holds that*

$$L(x^\#, \xi) \leq L(x^\#, \xi^\#) \leq L(x, \xi^\#) \qquad \textit{for all } x \in \mathbb{R}^N, \xi \in \mathbb{R}^m.$$

Saddle-property says that it is equivalent to finding a saddle point of the Lagrange function or to jointly optimize primal and dual problems, provided that strong duality holds. We will use it in Theorem 2.5. The next Theorem, stated here in a simplified version and known as the *Slater condition*, provides a sufficient condition for strong duality to hold, for a proof see Section 5.3 of [Boyd & Vanderberghe '04].

Theorem 2.4. ([Slater '50]): *Assume that $F_0, F_1, \ldots, F_M$ are convex functions with $dom(F_0) = \mathbb{R}^N$. If there exists $x \in \mathbb{R}^N$ such that $Ax = y$ and $F_\ell(x) < b_\ell$ for all $\ell \in [M]$, strong duality holds for the optimization problem (2.3). In the absence of inequality constraints, strong duality holds if there exists $x \in \mathbb{R}^N$ with $Ax = y$.*

Remark 11. *Duality theory for conic problems is more involved, see Section 4.3 of [Ruszczynski].*

After this digression about optimization, we need to state what is the main strategy for solving the problem of finding the sparsest vector satisfying the linear system of measurements. This is called *Basis Pursuit* or *ℓ_1-minimization* and is interpreted as the convex relaxation of (P_0).

Basis Pursuit (BP)

Input: measurement matrix A, measurement vector y.
Instructions:

$$x^\# = \text{argmin} \, ||z||_1 \qquad \text{subject to } Az = y. \tag{P_1}$$

Output: the vector $x^\#$.

Note we called Basis Pursuit a *strategy* and not an *algorithm* because we did not say how to implement the strategy above. In fact, there are many ways of doing that. Basis Pursuit is a linear program in the real case and a second-order cone program in the complex case, as we will show below. Therefore general purpose optimization algorithms can be used such as the Simplex Method or Interior-Point Methods. In particular, [Chen & Donoho '94] points out that *"Basis Pursuit is only thinkable because of recent advances in linear programming via "interior point" methods"*. We refer to [Nesterov & Nemirovskii '94][2] for details about this technique and to [Kim, Koh, Lustig, Boyd & Gorinevsky '08] for its applications to ℓ_1-minimization and sparse recovery.

We also have specifically design algorithms developed for ℓ_1-minimization. There is a discussion in the community about how faster they are compared to general purpose algorithms. We can cite the Homotopy Method, developed by [Donoho & Tsaig, '08][3], Iteratively Reweighted Least Squares, developed by [Daubechies, DeVore, Fornasier & Güntürk '10] or Primal-Dual Algorithms, developed by [Chambolle & Pock '11] among many other direct or iterative methods. These three main algorithms are described in Chapter 15 of [Rauhut & Foucart '13].

Besides, it is important to note that now there are many solvers available for Basis Pursuit. One can cite [ℓ_1-MAGIC], [YALL1], [Fast ℓ_1], [NESTA], [GPSR] and [L1-LS][4]. A user-friendly package for Convex Optimization in Python is [CVXPY], where many of the techniques mentioned above are implemented. The work [Lorenz, Pfetsch & Tillmann '15] studies and compares some solvers and heuristics for Basis Pursuit.

When we have measurement errors, we replace $Ax = y$ by $||Ax - y||_2 \leq \eta$ and then the problem

$$\min_{z \in \mathbb{C}^N} ||z||_1 \qquad \text{subject to } ||Az - y||_2 \leq \eta, \tag{$P_{1,\eta}$}$$

[2] The papers [Wright '04] and [Gondzio '12] are very interesting for a historical perspective about Interior-Point Methods.

[3] In the Statistics community, this algorithm was independently developed by [Efron, Hastie, Johnstone & Tibshirani '04] and is known as Least Angle Regression (or LARS).

[4] The most complete (albeit not up-to-date and disorganized) list of ℓ_1-minimization solvers can be found at Section 4 of `https://sites.google.com/site/igorcarron2/cs`.

is known as *Quadratically Constrained Basis Pursuit*. There are also two very relevant, related problems. The first one is *Basis Pursuit Denoising*(BPDN), which consists in solving, for some parameter $\lambda \geq 0$,

$$\min_{z \in \mathbb{C}^N} \lambda ||z||_1 + ||Az - y||_2^2. \tag{2.7}$$

The second one is LASSO, the Least Absolute Shrinkage and Selection Operator, which consists in solving for some parameter $\tau \geq 0$,

$$\min_{z \in \mathbb{C}^N} ||Az - y||_2 \qquad \text{subject to } ||z||_1 \leq \tau. \tag{2.8}$$

Basic Pursuit Denoising was introduced in [Chen & Donoho '94] and LASSO in [Tibshirani '96]. These three problems are related, as the next Theorem shows. For a unified treatment of them, one can see [Hastie, Tibshirani & Wainwright '15].

Theorem 2.5. *We have an equivalence between LASSO, Basis Pursuit Denoising and Quadratically Constrained Basis Pursuit as the following three statements show.*

i. If x is a minimizer of the Basis Pursuit Denoising with $\lambda > 0$, then there exists $\eta = \eta(x)$ such that x is a minimizer of the Quadratically Constrained Basis Pursuit.

ii. If x is a unique minimizer of the Quadratically Constrained Basis Pursuit with $\eta \geq 0$, then there exists $\tau = \tau(x) \geq 0$ such that x is a unique minimizer of the LASSO.

iii. If x is a unique minimizer of the LASSO with $\tau > 0$, then there exists $\lambda = \lambda(x) \geq 0$ such that x is a minimizer of the Basis Pursuit Denoising.

Proof. i.) Let $\eta = ||Ax - y||_2$ and consider $z \in \mathbb{C}^N$ such that $||Az - y||_2 \leq \eta$. Using the fact that x is a minimizer of (2.7), we have

$$\lambda ||x||_1 + ||Ax - y||_2^2 \leq \lambda ||z||_1 + ||Az - y||_2^2 \leq \lambda ||z||_1 + ||Ax - y||_2^2.$$

After some simplifications, we obtain $||x||_1 \leq ||z||_1$ and then x is a minimizer of the Quadratically Constrained Basis Pursuit.

ii.) Set $\tau = ||x||_1$ and consider $z \in \mathbb{C}^N$, $z \neq x$ such that $||z||_1 \leq \tau$. Since x is, by hypothesis, the unique minimizer of Quadratically Constrained Basis Pursuit, this implies that z cannot satisfy the constraint of $(P_{1,\eta})$. Hence, $||Az - y||_2 > \eta > ||Ax - y||_2$. This shows that x is the unique minimizer of (2.8).

iii.) This part is more elaborate and we will prove a more general result, where we replace the ℓ_2-norm by any other norm. The Theorem will be proved for the real case but it also holds for the complex setting just by interpreting $\mathbb{C}^N$ as $\mathbb{R}^{2N}$.

Let $||.||$ be a norm on $\mathbb{R}^m$ and $|||.|||$ a norm on $\mathbb{R}^N$. For $A \in \mathbb{R}^{m \times N}, y \in \mathbb{R}^m$ and $\tau > 0$, the general LASSO is given by

$$\min_{x \in \mathbb{R}^N} ||Ax - y|| \qquad \text{subject to } |||x||| \leq \tau, \tag{2.9}$$

while the general Basis Pursuit Denoising is given by

$$\min_{x \in \mathbb{R}^N} \lambda |||x||| + ||Ax - y||^2. \tag{2.10}$$

Therefore we will prove that if $x^\#$ is a minimizer of the general LASSO, then there exists $\lambda = \lambda(x) \geq 0$ such that x is a minimizer of the general Basis Pursuit Denoising. First of all, the general LASSO is obviously equivalent to

$$\min_{x\in\mathbb{R}^N} ||Ax-y||^2 \qquad \text{subject to } |||x||| \leq \tau. \tag{2.11}$$

The Lagrange function for this problem is given by

$$L(x,\xi) = ||Ax-y||^2 + \xi\big(|||x||| - \tau\big) \tag{2.12}$$

For $\tau > 0$, there exist vectors x with $|||x||| < \tau$. Then, by Theorem 2.4, we have strong duality for (2.9), so we guarantee the existence of a dual optimal $\xi^\# \geq 0$. The *saddle-point property* ensures that $L(x^\#,\xi^\#) \leq L(x,\xi^\#)$ for all $x \in \mathbb{R}^N$. Therefore $x^\#$ is also a minimizer of $x \mapsto L(x,\xi^\#)$. Since the constant term $-\xi^\#\tau$ does not affect the minimizer, $x^\#$ is also a minimizer of $||Ax-y||^2 + \xi^\#|||x|||$ and the conclusion follows with $\lambda = \xi^\#$. □

The parameters η, λ and τ can be seen as regularization parameters for the solution of the linear system. Theorem 2.5 shows that the transformation of the parameters depends on the minimizer. Therefore we need to solve the optimization problems before we come to know them. For this reason, this equivalence is useless for practical purposes. Finding appropriate parameters is a highly nontrivial task and the topic is widely discussed in the literature, see [Yu & Feng '14] and references therein. Also, problems related to how this kind of optimization technique can lead to false discoveries are discussed in [Su, Bogdan & Candès '15].

Basis Pursuit is a linear program in the real case and a second-order cone program in the complex case. For the real case, let us introduce the variables $x^+, x^- \in \mathbb{R}^N$. For $x \in \mathbb{R}^N$, let

$$x_j^+ = \begin{cases} x_j & \text{if } x_j > 0 \\ 0 & \text{if } x_j \leq 0 \end{cases} \qquad \text{and} \qquad x_j^- = \begin{cases} 0 & \text{if } x_j > 0 \\ -x_j & \text{if } x_j \leq 0. \end{cases}$$

Hence, problem (P_1) is equivalent to the following linear optimization problem for the variables $x^+, x^- \in \mathbb{R}^N$

$$\min_{x^+,x^-\in\mathbb{R}^N} \sum_{i=1}^N (x^+ + x^-) \qquad \text{subject to } [A|-A]\begin{bmatrix} x^+ \\ x^- \end{bmatrix} = y, \quad \begin{bmatrix} x^+ \\ x^- \end{bmatrix} \geq 0. \tag{P_1'}$$

With a solution $(x^+)^\#, (x^-)^\#$ of this problem in hand, we get the solution of the original problem (P_1) just by taking $x^\# = (x^+)^\# - (x^-)^\#$. This consideration allows us to conclude that we have, in fact, a linear problem.

In the complex setting, we will be looking, instead, to the more general Quadratically Constrained Basis Pursuit. Then, given a vector $z \in \mathbb{C}^N$, we need to introduce its real and imaginary parts $u, v \in \mathbb{R}^N$ and also a vector $c \in \mathbb{R}^N$ such that $c_j \geq |z_j| = \sqrt{u_j^2 + v_j^2}$ for all $j \in [N]$. The problem $P_{1,\eta}$ will be equivalent to the following optimization problem with variables $c, u, v \in \mathbb{R}^N$:

$$\min_{c,u,v,\in\mathbb{R}^N} \sum_{j=1}^N c_j \quad \text{subject to} \quad \begin{cases} \left|\left| \begin{bmatrix} \mathrm{Re}(A) & -\mathrm{Im}(A) \\ \mathrm{Im}(A) & \mathrm{Re}(A) \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix} - \begin{bmatrix} \mathrm{Re}(y) \\ \mathrm{Im}(y) \end{bmatrix} \right|\right|_2 \leq \eta \\ \sqrt{u_1^2+v_1^2} \leq c_1, \ldots, \sqrt{u_N^2+v_N^2} \leq c_N. \end{cases} \tag{$P_{1,\tau}'$}$$

This is a *second-order cone problem*. Then given a solution $(c^\#, u^\#, v^\#)$, the solution of the original problem is $x^\# = u^\# + iv^\#$. When we take $\eta = 0$, we recover the solution of (P_1) in the complex case.

Why is Basis Pursuit good alternative to the original combinatorial problem (P_0)? Because in the real case ℓ_1-minimizers are sparse, as we will show now. In Chapter 3 this will by fully explored through the notion of the Null Space Property for measurement matrices.

Theorem 2.6. *Let $A \in \mathbb{R}^{m\times N}$ be a measurement matrix with columns $a_1, \ldots, a_N$. Assuming the uniqueness of a minimizer $x^\#$ of*

$$\min_{x\in\mathbb{R}^N} ||x||_1 \qquad \textit{subject to } Ax = y,$$

then the set $\{a_j, j \in supp(x^\#)\}$ *is linearly independent, and in particular* $||x^\#||_0 = \#(supp(x^\#)) \leq m$.

Proof. Suppose, by contradiction, that the set $\{a_j, j \in S\}$ is linearly dependent, where $S = \text{supp}(x^\#)$. This means that there exists a nonzero vector $v \in \mathbb{R}^N$ supported on S such that $Av = 0$. By uniqueness of $x^\#$, for any $t \neq 0$,

$$||x^\#||_1 < ||x^\# + tv||_1 = \sum_{j\in S} |x_j^\# + tv_j| = \sum_{j\in S} \text{sgn}(x_j^\# + tv_j)(x_j^\# + tv_j).$$

Now, we need to understand what happens with a sign of a number when we add another number to it. If $|b| < |a|$ the $\text{sgn}(a + b) = \text{sgn}(a)$. Therefore, when $|t| < \min_{j\in S} |x_j|^\#/||v||_\infty$, we have

$$\text{sgn}(x_j^\# + tv_j) = \text{sgn}(x_j^\#) \qquad \text{for all } j \in S.$$

It follows that, for $t \neq 0$ with $|t| < \min_{j\in S} |x_j|^\#/||v||_\infty$,

$$||x^\#||_1 < \sum_{j\in S} \text{sgn}(x_j^\# + tv_j)(x_j^\# + tv_j) = \sum_{j\in S} \text{sgn}(x_j^\#)(x_j^\# + tv_j) = \sum_{j\in S} \text{sgn}(x_j^\#)x_j^\# + t\sum_{j\in S} \text{sgn}(x_j^\#)v_j$$

$$= ||x_j^\#||_1 + t\sum_{j\in S} \text{sgn}(x_j^\#)v_j.$$

This is a contradiction because we can always take a very small $t \neq 0$ such that $t\sum_{j\in S} \text{sgn}(x_j^\#)v_j \leq 0$. □

Remark 12. *In the complex case, the situation is more delicate. Consider the vector* $x = [1, e^{i2\pi/3}, e^{i4\pi/3}]$ *and the measurement matrix described by*

$$A = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \end{bmatrix}.$$

This vector is the unique *minimizer of the problem* $\min_{z\in\mathbb{C}^3} ||z||_1$ *subject to* $Az = Ax$. *See Section 3.1 of [Rauhut & Foucart '13]. Hence, in the complex setting,* ℓ_1*-minimization does not necessarily provides* m*-sparse solutions, where* m *is the number of measurements. In order to better understand when this occurs, we need the theory of the next chapters, such as the Null Space Property.*

Quadratically Constrained Basis Pursuit relies on $||Ax - y||_2 \leq \eta$, that is, the noise is bounded in the ℓ_2-norm regardless of structure or prior information about it. So, in the case of Gaussian noise, which is unbounded, the error in sparse vector recovery typically does not decay as the number of measurement increases, see pages 29 and 30 of [Boche et at. '15]. In order to deal which this problem, another type of ℓ_1-minimization algorithm was proposed by [Candès & Tao III, 06], called *Dantzig Selector*[5]. It is described by

$$\min_{z\in\mathbb{C}^N} ||z||_1 \qquad \text{subject to} \quad ||A^*(Az - y)||_\infty \leq \tau. \tag{DS}$$

The heuristics of this method is based on the fact that the residual $r = Az - y$ should have small correlation with all columns a_j of the matrix A. Indeed, the constraint is $||A^*(Az - y)||_\infty = \max_{j\in[N]} |\langle r, a_j\rangle|$. There is a theory for the Dantzig Selector similar to the one developed through this dissertation for Basis Pursuit, which is sometimes more complicated, see Chapter 8 in [Elad '10], [Hastie, Tibshirani & Wainwright '15], [Meinshausen, Rocha & Yu '07], [James, Radchenko & Lv '09], [Lv & Fan '09], [Zhang '09] and [Bickel, Ritov & Tsybakov '09].

[5] As Tao points out in [Tao's Blog - 22/03/2008], *"we called the Dantzig selector, due to its reliance on the linear programming methods to which George Dantzig, who had died as we were finishing our paper, had contributed so much to".*

2.3 Greedy Algorithms

The basic idea of a greedy algorithm in Compressive Sensing is to update the target vector in such a way that in each iteration, the algorithm adds one index to the target support and then finds the vector that best fits the measurements with this given support. This kind of strategy justifies the name of such class of algorithms.

The most famous algorithm in this context is called the *Orthogonal Matching Pursuit*, sometimes called Orthogonal Greedy Algorithm. This algorithm has been rediscovered many times in different fields. We can trace its origins back to [Chen, Billings & Luo '89], in the context of system identification in Control Theory. Independently, it was introduced and analyzed in the context of Time-Frequency Analysis by [Davies, Mallat & Zhang '94] and [Pati, Rezaiifar & Krishnaprasad '93]. Here is the formal description of the algorithm.

Orthogonal Matching Pursuit (OMP)

Input: measurement matrix A, measurement vector y.
Initialization: $S^0 = \emptyset, x^0 = 0$.
Iteration: repeat until a stopping criterion is met at $n = \overline{n}$:

$$S^{n+1} = S^n \cup \{j_{n+1}\}, \qquad j_{n+1} = \underset{j \in [N]}{\operatorname{argmax}} \{|(A^*(y - Ax^n))_j|\}, \tag{OMP$_1$}$$

$$x^{n+1} = \underset{z \in \mathbb{C}^N}{\operatorname{argmin}} \{||y - Az||_2, \operatorname{supp}(z) \subset S^{n+1}\}. \tag{OMP$_2$}$$

Output: the $\overline{n}$-sparse vector $x^{\#} = x^{\overline{n}}$.

To identify the signal x, we need to determine which columns of A participate in the measurement vector. Informally speaking, OMP is a greedy algorithm that selects at each step the column that is most correlated with the current residuals. This column is then included into the set of selected columns. The algorithm updates the residuals by projecting the observation onto the linear subspace spanned by the columns that have already been selected and the algorithm then iterates.

The main goal of the introduction of this algorithm was to improve the already existing Matching Pursuit[6] by forcing it to converge, for finite-dimensional signals, in a finite number of steps, which does not happen for the original (Nonorthogonal) Matching Pursuit, as Theorem 4.1 of [DeVore & Temlyakov '96] and the discussion in Section 2.3.2 of [Chen, Donoho & Saunders '01] show. In general, the residuals of the Orthogonal MP decrease faster than the Nonorthogonal MP. However, this improvement possess some disadvantages, as, for example, orthogonal projection procedure can yield unstable expansions by selecting ill-conditioned elements through the running of the algorithm. Moreover, Orthogonal MP also requires much more operations than Nonorthogonal Pursuits due to the Gram-Schmidt orthogonalization process.

Remark 13. *[Davies, Mallat & Avellaneda '97] showed exponential convergence and made a detailed comparison of both greedy algorithms in terms of theoretical complexity and practical performance. For a proof and discussion of the decay rate, see Chapter 3 of [Elad '10].*

It is important to note that the connection between sparse approximation and greedy algorithms was first done by [Tropp '04]. His results showed that under some conditions on the coherence of the matrix (see Theorem 4.41), OMP works for sparse recovery in the noiseless case. [Cai, Wang & Xu III '10] proved that this condition is sharp and the performance of OMP on noisy case was analyzed by [Cai & Wang '11].

[6] This algorithm was introduced by [Mallat & Zhang '93] in the Signal Processing community and by [Friedman & Stuetzle '81] in the Statistics community under the name Projection Pursuit Regression.

As we can see in the description of OMP, the costlier part is the projection step (OMP_2). Typically a QR-decomposition of A_{S_n} is used in each step and then efficient methods for updating the QR-decomposition when a column is added to the matrix are used, see Section 3.2 of [Björck '96]. Moreover, fast vector-matrix multiplication via FFT can be performed in (OMP_1) for the computation of $A^*(y - Ax^n)$. We will briefly describe how this is done for a general least-square problem. For any matrix $A \in \mathbb{C}^{m\times n}$ with $m \geq n$, there exists a unitary matrix $Q \in \mathbb{C}^{m\times m}$ and an upper triangular matrix $R \in \mathbb{C}^{n\times n}$ such that

$$A = Q\begin{pmatrix} R \\ 0 \end{pmatrix}.$$

So, for a general least-squares problem of minimizing $||Ax - y||_2$ for all $x \in \mathbb{C}^n$, using that Q is unitary, we have the following

$$||Ax - y||_2 = ||Q^*Ax - Q^*y||_2 = \left|\left|\begin{pmatrix} R \\ 0 \end{pmatrix}x - Q^*y\right|\right|_2. \tag{2.13}$$

Partitioning $b = (b_1, b_2) = Q^*y$ with $b_1 \in \mathbb{C}^n$, the right-hand side of (2.13) will be minimized when we solve the triangular system $Rx = b_1$ and this can be done with a simple backward elimination. Therefore, at each iteration of OMP, we need to solve a least-square problem with a new added column to the matrix A.

We discuss now the choice of the index j_{n+1}, as stated in (OMP_1). It is given by a greedy strategy where the objective is to reduce the ℓ_2-norm of the residual $y - Ax^n$ as much as possible at each iteration. Lemma 2.7 explains why an index j maximizing $|(A^*(y - Ax^n))_j|$ is a good candidate for this large decrease of the residual. We just need to apply it to $S = S^n$ and $v = x^n$.

Lemma 2.7. *Let $A \in \mathbb{C}^{m\times N}$ be a matrix with ℓ_2-normalized columns. Given $S \subset [N]$, v supported on S and $j \in [N]$, if*

$$w = \underset{z\in\mathbb{C}^N}{\text{argmin}}\{||y - Az||_2, \textit{supp}(z) \subset S \cup \{j\}\},$$

then

$$||y - Aw||_2^2 \leq ||y - Av||_2^2 - |(A^*(y - Av))_j|^2.$$

Proof. Note that any vector $v + te_j$, with $t \in \mathbb{C}$, is supported on $S \cup \{j\}$. By the definition of w, we have

$$||y - Aw||_2^2 \leq \min_{t\in\mathbb{C}} ||y - A(v + te_j)||_2^2.$$

The idea is then to work in polar coordinates and choose proper radial and angular numbers to represent t. Indeed, writing $t = \rho e^{i\theta}$, with $\rho \geq 0$ and $\theta \in [0, 2\pi)$, we estimate

$$||y - A(v + te_j)||_2^2 = ||y - Av - tAe_j||_2^2 = ||y - Av||_2^2 + |t|^2||Ae_j||_2^2 - 2\text{Re}\left(\bar{t}\langle y - Av, Ae_j\rangle\right)$$

$$= ||y - Av||_2^2 + \rho^2 - 2\text{Re}\left(\rho e^{-i\theta}(A^*(y - Av))_j\right) \geq ||y - Av||_2^2 + \rho^2 - 2\rho|(A^*(y - Av))_j|$$

with equality for a properly chosen θ. The right-hand side of this inequality is a quadratic polynomial in ρ and this expression is minimized when $\rho = |(A^*(y - Au))_j|$. Therefore this leads to

$$\min_{t\in\mathbb{C}} ||y - A(v + te_j)||_2^2 \leq ||y - Av||_2^2 - |(A^*(y - Av))_j|^2$$

which concludes the proof.

□

Another interesting point is that (OMP_2) is equivalent to solving the normal equations of least squares. In fact, it also reads as $x^{n+1}_{S^{n+1}} = (A^*_{S^{n+1}} A_{S^{n+1}})^{-1} A^*_{S^{n+1}} y = A^\dagger_{S^{n+1}} y$, where $x^{n+1}_{S^{n+1}}$ denotes the restriction of x^{n+1} to the support set S^{n+1}. This is justified by the following lemma.

Lemma 2.8. *Given an index set $S \subset [N]$, if*

$$v = \underset{z \in \mathbb{C}^N}{\operatorname{argmin}} \{||y - Az||_2, \mathit{supp}(z) \subset S\},$$

then

$$(A^*(y - Av))_S = 0.$$

Proof. The definition of v tells us that the vector Av is the orthogonal projection of y onto the space $\{Az, \text{supp}(z) \subset S\}$. This means that

$$\langle y - Av, Az \rangle = 0 \qquad \text{for all } z \in \mathbb{C}^N \text{ with } \text{supp}(z) \subset S.$$

The orthogonality condition is equivalent to $\langle A^*(y - Av), z \rangle = 0$ for all $z \in \mathbb{C}^N$ with $\text{supp}(z) \subset S$ which is the same as $(A^*(y - Av))_S = 0$. □

Lemma 2.8 is important as it may be useful for solving normal equations instead of using QR decomposition. In this case, a direct method, such as Cholesky decomposition, or an iterative method, such as Conjugate Gradient method, may be used. For details of when the use of such methods has advantages, see Chapters 6 and 7 of [Björck '96].

The most natural stopping criterion for OMP is $Ax^{\bar{n}} = y$. In applications, where we have measurement and rounding errors, this needs to be changed to $||y - Ax^{\bar{n}}||_2 \leq \varepsilon$ or $||A^*(y - Ax^{\bar{n}})||_\infty \leq \varepsilon$ for some tolerance parameter $\varepsilon > 0$.

If we have a priori estimates for the sparsity of the signals, we can use this information to provide $\bar{n} = s$ as a stopping criterion because the target vector $x^{\bar{n}}$ is n-sparse. In fact, in the case where A is an orthogonal matrix, this forces the algorithm to successfully recover the sparse $x \in \mathbb{C}^N$ from $y = Ax$. Indeed, from $x^n_{S^n} = A^\dagger_{S^n} y$, we see that the vector x^n will be the n-sparse vector consisting of the n largest entries of x. In the general case, the success of OMP for recovering s-sparse vectors using s iterations is given by the following theorem.

Theorem 2.9. *Given a matrix $A \in \mathbb{C}^{m \times N}$, every nonzero vector $x \in \mathbb{C}^N$ supported on a set S of size s is recovered from $y = Ax$ after at most s iterations of orthogonal matching pursuit if and only if the matrix A_S is injective and*

$$\max_{j \in S} |(A^* r)_j| > \max_{\ell \in \overline{S}} |(A^* r)_\ell|, \tag{2.14}$$

for all nonzero $r \in \{Az, \mathit{supp}(z) \subset S\}$.

Proof. First, we will assume that the OMP works and recovers *every* vector supported on a set S in at most $s = \#S$ iterations. Therefore, any two vectors x_1, x_2 having S as support and the same measurement vector $y = Ax_1 = Ax_2$ must be the same. This implies that A_S is injective. Moreover, we saw that an index chosen at the first iteration will always remain in the target support and then, if $y = Ax$ for some $x \in \mathbb{C}^N$ exactly supported on S, then an index $\ell \in \overline{S}$ cannot be chosen at the first iteration and $\max_{j \in S} |(A^* y)_j| > |(A^* y)_\ell|$. We use this to conclude

$$\max_{j \in S} |(A^* y)_j| > \max_{\ell \in \overline{S}} |(A^* y)_\ell| \qquad \text{for all nonzero } y \in \{Az,\ \text{supp}(z) \subset S\}.$$

We established the two necessary conditions. Now, we proceed to prove that they are also sufficient. Assume that $Ax^1 \neq y, \ldots, Ax^{s-1} \neq y$, because otherwise there is nothing to do. We need to prove that S^n is a subset of S of size n for any $0 \leq n \leq s$ and from this we will deduce that $S^s = S$. Using this with $Ax^s = y =: Ax$, given by (OMP_2), allows us to conclude that $x = x^s$ because A_S is

injective by hypothesis. Therefore we need to establish our claim by using (2.14). We will prove by induction that $S^n \subset S$. We begin with $S^0 = \emptyset \subset S$. Given $0 \leq n \leq s-1$, we see that $S^n \subset S$ yields $r^n = y - Ax^n \in \{Az, \text{supp}(z) \subset S\}$. Then, by equation (2.14), the index j_{n+1} must lie in S. Hence, $S^{n+1} = S^n \cup \{j_{n+1}\} \subset S$. This proves that $S^n \subset S$ for an $0 \leq n \leq s$.

Now, for $1 \leq n \leq s-1$, Lemma 2.8 implies that $(A^* r^n)_S^n = 0$. If $\{j_{n+1}\} \subset S^n$, then $A^* r^n = 0$ and by equation (2.14) we could conclude that $r^n = 0$, since the strict inequality must occur for any *nonzero* vector. This cannot happen since we assumed $Ax^1 \neq y, \ldots, Ax^{s-1} \neq y$. Therefore $\{j_{n+1}\} \not\subset S^n$, that is, in each iteration, a new index, different from all previous indexes is added to the target support. This inductively proves that S^n is a set of size n. □

As discussed by [Chen, Donoho & Saunders '01], *"because the algorithm is myopic, one expects that, in certain cases, it might choose wrongly in the first few iterations and, in such cases, end up spending most of its time correcting for any mistakes made in the first few terms. In fact this does seem to happen."*. This phenomenon will be illustrated in Section 5.5, after Theorem 5.22.

Despite these pathological negative results, there is theoretical evidence and empirical results which suggest that OMP can recover an s-sparse signal when the number of measurements m is nearly proportional to s, as the next Theorem shows.

Theorem 2.10. (Theorem 2 of [Tropp & Gilbert '07]): *Fix $\delta \in (0, 0.36)$ and choose $m > Cs\log(n/\delta)$. Assume that $x \in \mathbb{R}^N$ is s-sparse and let $A \in \mathbb{R}^{m \times N}$ have i.i.d entries from the Gaussian distribution $N(0, 1/m)$. Then, given the data $y = Ax$, orthogonal matching pursuit can reconstruct the signal x with probability exceeding $1 - 2\delta$. The constant C satisfies $C \leq 20$. For large s it can be shown that $C \approx 4$ is enough.*

These are not the unique greedy algorithms for sparse approximation. Many corrections and modifications were proposed and we can cite Subspace Pursuit [Dai & Milenkovic '09], Gradient Pursuit [Blumensath & Davies '08], Stagewise Orthogonal Matching Pursuit [Donoho, Tsaig, Drori & Starck '12] and Compressive Sampling Matching Pursuit, known as CoSaMP [Needell & Tropp '08]. There is a whole monograph devoted to topic of greedy approximations and algorithms, see [Temlyakov '11], specially Chapter 5, where the problem of Compressive Sensing is treated.

2.4 Thresholding Algorithms

In order to recover sparse vectors, we need to understand the action of the measurement matrix on them. Thresholding algorithms rely on the fact that we will approximate the inversion of the action of A onto sparse vectors by the action of its adjoint A^*. Thus, the basic thresholding strategy is to determine the support of the s-sparse vector $x \in \mathbb{C}^N$ to be recovered from $y = Ax \in \mathbb{C}^m$ as the indices of s largest absolute entries of A^*. After finding the support, we perform some least-square technique in order to find the vector having this support that best fits the measurements. To describe the strategy, we need to introduce two operators. The first one denotes the best s-term approximation of a vector and the second denotes the support of this approximation,

$$H_s(z) = z_{L_s(z)} \tag{2.15}$$

$$L_s(z) = \text{index set of } s \text{ largest absolute entries of } z \in \mathbb{C}^N \tag{2.16}$$

Next, we have the strategy using this *cut-off* idea.

Basic Thresholding Strategy

Input: measurement matrix A, measurement vector y, sparsity level s.
Instruction:

$$S^{\#} = L_s(A^*y). \tag{BT$_1$}$$

$$x^{\#} = \underset{z\in\mathbb{C}^N}{\operatorname{argmin}}\{||y - Az||_2, \operatorname{supp}(z) \subset S^{\#}\}. \tag{BT$_2$}$$

Output: the s-sparse vector $x^{\#} = x^{\overline{n}}$.

As Theorem 2.9 for greedy strategy, we have a necessary and sufficient condition for the recovery of s-sparse vectors through the use of Thresholding strategy.

Proposition 2.11. *A vector $x \in \mathbb{C}^N$ supported on a set S is recovered from $y = Ax$ via Basic Thresholding Strategy if and only if* $\min_{j\in S} |(A^*y)_j| > \max_{\ell\in\overline{S}} |(A^*y)_\ell|$.

Proof. The vector x will be recovered if and only if the index set $S^{\#}$ defined in the iteration of the strategy coincides with the set S and this happens if and only if any entry of A^*y on S is greater than any entry of A^*y on $\overline{S}$. □

This strategy seems too simple to have any chance of working. We can reformulate it and solve the system $Az = y$ using the prior information that the solution is s-sparse for some s. Actually, we will not solve the fat and short linear system $Az = y$ but instead we will solve the square system $A^*Az = A^*y$. This system can be reformulated as the fixed-point equation $z = (\mathrm{Id} - A^*A)z + A^*y$. This, in turn, can be solved by using the iteration $x^{n+1} = (\mathrm{Id} - A^*A)x^n + A^*y$.

The thresholding strategy, applied at this fixed-point iteration, tells us to keep the s largest absolute entries of $(\mathrm{Id} - A^*A)x^n + A^*y = x^n + A^*(y - Ax^n)$. If the s largest entries are not uniquely defined, this algorithm selects the smallest possible indices. Formally, that is the content of the next algorithm. It was developed by [Blumensath & Davies II '08] and analyzed in [Blumensath & Davies '09]. Its analysis relies on the fact that the matrix A^*A behaves like the identity when its domain and range are restricted to small support sets (as we shall see in Chapter 5) and then $x^{n+1} = x^n + A^*A(x - x^n) \approx x^n + x - x^n = x$.

The authors of the algorithm argue that it has many good features such as robustness to observation noise, near-optimal error guarantees, the requirement of a fixed number of iterations depending only on the logarithm of a form of signal to noise ratio of the signal, a memory requirement which is linear in the problem size, among others.

Iterative Hard Thresholding (IHT)

Input: measurement matrix A, measurement vector y, sparsity level s.
Initialization: s-sparse vector x^0, typically $x^0 = 0$.
Iteration: repeat until a stopping criterion is met at $n = \overline{n}$:

$$x^{n+1} = H_s(x^n + A^*(y - Ax^n)). \tag{IHT}$$

Output: the s-sparse vector $x^{\#} = x^{\overline{n}}$.

In the class of greedy algorithms, the difference between Matching Pursuit and Orthogonal Matching Pursuit is that in the later we decide to pay the price for an orthogonal projection. Here, we can do the same and look, at each iteration, to the vector with the same support of x^{n+1} that best fits the measurements. This idea was explored by [Foucart '11], where the *Hard Thresholding Pursuit* was developed as a fusion of IHT and the greedy algorithm CoSaMP. In fact, this work did more and created a

family of thresholding algorithms indexed by an integer k. There, Iterative Hard Thresholding and Hard Thresholding Pursuit correspond to the cases $k = 0$ and $k = \infty$, respectively.

Hard Thresholding Pursuit(HTP)

Input: measurement matrix A, measurement vector y, sparsity level s.
Initialization: s-sparse vector x^0, tipically $x^0 = 0$.
Iteration: repeat until a stopping criterion is met at $n = \overline{n}$:

$$S^{n+1} = L_s(x^n + A^*(y - Ax^n)), . \tag{HTP$_1$}$$

$$x^{n+1} = \underset{z \in \mathbb{C}^N}{\operatorname{argmin}}\{||y - Az||_2, \operatorname{supp}(z) \subset S^{n+1}\}. \tag{HTP$_2$}$$

Output: the s-sparse vector $x^\# = x^{\overline{n}}$.

Recently, some theoretical results concerning HTP have been proved. [Bouchot, Foucart & Hitczenko '16], for example, proved that the number of iterations can be estimated independently of the shape of x and it is at most proportional to the sparsity s. For a precise statement, see Theorem 5 in that article. Also, it provided some benchmarks for HTP: in terms of success rate, its performance is roughly similar to OMP, however, in terms of number of iterations HTP does perform best, owing to the fact that we have prior knowledge of the sparsity s.

Clearly, in these thresholding algorithms we need an estimate of vector sparsity. The *Soft Thresholding Algorithms* class of algorithms was developed to deal with situations in which we have no such estimates. There, we substitute the hard thresholding operator by the soft thresholding operator with threshold $\tau > 0$. This operator maps each entry z_j of a vector $z \in \mathbb{C}^N$ to

$$S_\tau(z_j) = \begin{cases} \operatorname{sgn}(z_j)(|z_j| - \tau), & \text{if } |z_j| \geq \tau, \\ 0, & \text{otherwise.} \end{cases}$$

One can cite as an example of such algorithms, the Iterative Shrinkage-Thresholding Algorithm developed by [Daubechies, Defrise & De Mol '04]. This is a variation of the classical Landweber iteration, well-known to the Inverse Problems community, see section 2.3 of [Kirsch '11]. It is typically used in the context of image restoration and in this context, we are able to use it to deal with non-differentiable regularizers such as total-variation regularization and wavelet-based regularization [Bioucas-Dias & Figueiredo '07].

2.5 The Search For the Perfect Algorithm

> Unfortunately, there is little guidance available on
> choosing a good technique for a given parameter regime.
> [Tropp & Gilbert '07]

After presenting these three major classes of algorithms, we should ask which of them is better suited for a certain application. In other words, we need to answer the following question: *Given the sparsity level s, the number of measurements m and the ambient dimension N, which algorithm should we use?* In Compressive Sensing we seek for uniform recovery results, that is, we aim to reconstruct *every* sparse vector with the same measurement matrix. Thus, if we have some prior information about the signals we are looking for, that is, if we have more structure beyond sparsity, can we take advantage of it in order to carefully select the algorithms?

These questions have not been not fully answered yet. The number of articles related to numerical issues of sparse recovery is small compared to the development of theoretical guarantees. Important references in this area are [Elad '10], which is devoted to numerical investigations in Compressive Sensing, as well as [Pope '09], [Maleki & Donoho '10], [Lorenz, Pfetsch & Tillmann '15] and

[Blanchard & Tanner '15]. This last one was the first to perform large-scale empirical testing with more realistic, application sized problems, typically with ambient dimension $N = 2^{18}$ or $N = 2^{20}$. This is particularly important in evaluating the behavior of algorithms in the extreme undersampling regime of $m \ll N$.

These analysis are out of the scope of this dissertation. Even so, we can provide some elementary rules of thumb. For small sparsity s, OMP is typically the fastest option because its speed depends on the number of iterations, which will be equal to s if the algorithm succeeds. For mild sparsity (compared to N) thresholding algorithms are preferred because its runtime is typically not affected by the sparsity level. Assigning rules of thumb for Basis Pursuit is more difficult because it depends of how we implement this strategy.

Due to the lack of a comprehensive numerical study, the following project remains open.

Open Project: Provide a computational investigation for worst and average case behavior of all algorithms in the three major classes: optimization methods, greedy methods and thresholding methods. Identify the regions of the problem size where these algorithms are able to reliably recover the sparsest solution. Construct phase transition diagrams to the probability of recovery for a given algorithm. Perform all the tests with matrices drawn from some probability distribution like Gaussian or Bernoulli and also with deterministic matrices.

Chapter 3

The Null Space Property

It is the hallmark of any deep truth that its negation is also a deep truth.[1].
Niels Bohr

3.1 Introduction

As discussed in Chapter 2, in this dissertation we are interested in the solution of fat and short linear systems. This is so because they model the simultaneous acquisition and (high) compression of data. As the matrices of these systems have nontrivial kernel, we must define a proper notion of solution, i.e., we must specify what properties this solution should have. This is the same as saying that find a unique solution of an underdetermined linear system is impossible unless some additional information about the solution is given. Therefore, after this regularization procedure, there is hope in finding a unique (regularized) solution. As we are looking for parsimonious representation of signals, the natural approach is to look for the sparsest solution. In previous chapters, we argued that this approach makes sense, since many natural signals contain little information, as they are a combination of only a handful of vector from an appropriate basis.

In mathematical terms, we need to recover a sparse vector $x \in \mathbb{C}^N$ from its measurements $y = Ax \in \mathbb{C}^m$, where $m < N$. In order to do so, we need to solve the following optimization problem:

$$\min_{x \in \mathbb{C}^N} ||z||_0 \qquad \text{subject to } Az = y. \tag{P_0}$$

This combinatorial problem is NP-hard, as Theorem 1.16 shows, so we need to develop some smart strategies in order to solve it. The most popular strategy is Basis Pursuit, or ℓ_1-minimization, presented in Chapter 2. Therefore, our approach will be to solve the problem

$$\min_{x \in \mathbb{C}^N} ||z||_1 \qquad \text{subject to } Az = y. \tag{P_1}$$

This chapter is devoted to understanding one property the matrix A must have in order to guarantee that solutions from problem (P_1) are solutions of the problem (P_0). It is the *null space property* (NSP), a necessary and sufficient condition for exact reconstruction of a sparse vector x from its measurements $y = Ax$, which we present in the next section. Even more, we will define two further criteria which we expect from a recovery algorithm, namely Stability and Robustness and show that NSP also works for stable reconstruction with respect to sparsity defect and that it is robust to measurement error. Then, the problem of low-rank matrix recovery will be explored. This can be seen as a variation of the Compressive Sensing problem and a generalization of the null space property will appear during the analysis of low-rank matrix recovery.

[1]Quoted by Max Delbruck, "Mind from Matter? An Essay on Evolutionary Epistemology", Blackwell Scientific Publications, Palo Alto, CA, 1986; page 167.

3.2 The Null Space Property

As we are interested in recovering the sparsest solution, we need to know when, i.e., for which matrices, this can be done. Moreover, we also need to know when we have uniqueness of the recovered solution. However, finding a property of the matrix A that implies that a solution to the (P_1) problem is a solution of the (P_0) problem was a nontrivial task. This took a little more than a decade, from the beginning of Basis Pursuit studies in the Thesis [Chen '95], until the formal definition of the Null Space Property in the paper [Cohen, Dahmen & DeVore '09]. It is important to mention that this property had already implicitly appeared in [Donoho & Huo '01], [Elad & Bruckstein '02] and [Donoho & Elad ' 03]. The first place where this property was isolated is in Lemma 1 of [Gribonval & Nielsen '03].

Definition 3.1. A matrix $A \in \mathbb{K}^{m\times N}$ is said to satisfy the *null space property*[2] relative to a set $S \subset [N]$ if

$$||v_S||_1 < ||v_{\overline{S}}||_1 \qquad \forall v \in \ker A \backslash \{0\} \tag{NSP}$$

It is said to satisfy the *null space property of order* s if it satisfies the null space property relative to any set $S \subset [N]$ with $\#S = s$.

Remark 14. *This property tells us how the null space of the matrix A should be oriented in order to touch the ℓ_1-ball in just one point. This geometric interpretation will be explored in Theorem* 3.12 *and can be seen in Figure 3.1 below.*

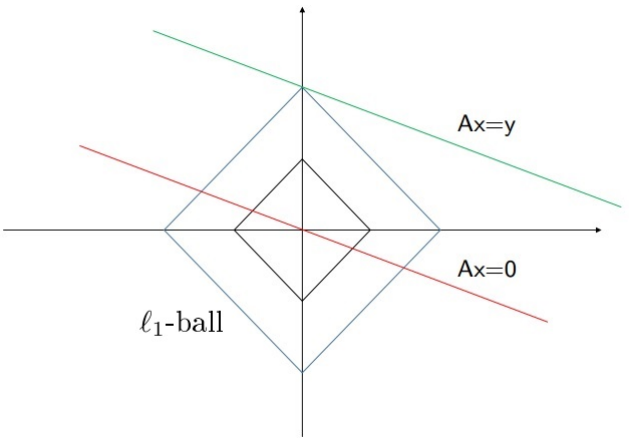


Figure 3.1: Proper null space of A for uniqueness in the reconstruction of sparse vector.

First of all, it is important to observe that if this condition holds for the index set S of s largest (in absolute value) entries of v, then it holds for any other set S with $\#S \leq s$. Also, we can reformulate this definition of the null space property in two ways. The first one is by adding $||v_S||_1$ to both sides of the inequality in the definition. Then, the null space relative to a set S become

$$2||v_S||_1 < ||v||_1 \qquad \forall v \in \ker A \backslash \{0\}. \tag{3.1}$$

For the second, one can add $||v_{\overline{S}}||_1$ to both sides of the inequality. Thus, the null space property of order s will be

$$||v||_1 < 2\ \sigma_s(v)_1 = 2 \inf_{||z||_0 \leq s} ||x - z||_p \qquad \forall v \in \ker A \backslash \{0\}$$

[2][Cohen, Dahmen & DeVore '09] defined NSP as a slightly more general property, namely, $||v||_1 \leq C\sigma_s(x)_1$ for all $v \in \ker A$, where $C \geq 1$ in an unspecified constant.

We can now state the main Theorem concerning the NSP, which is a necessary and sufficient condition for exact recovery of sparse vectors through ℓ_1-norm minimization. The statement given here is a small modification form the original reference.

Theorem 3.2. (Essentially Theorem 3.2 of [Cohen, Dahmen & DeVore '09]): *Given a matrix $A \in \mathbb{K}^{m\times N}$, every vector $x \in \mathbb{K}^N$ supported on a set S is the unique solution of (P_1) with $Ax = y$ if and only if A satisfies the null space property relative to S.*

Proof. Suppose that every vector $x \in \mathbb{K}^{m\times N}$ supported on S is the unique minimizer of $||z||_1$ subject to $Az = Ax$. Thus, for any $v \in \ker A\backslash\{0\}$, the vector v_S is the unique minimizer of $||z||_1$ subject to $Az = Av_S$. But, as we have $Av = 0$, then $A(v_S + v_{\overline{S}}) = Av = 0$ and so $A(-v_{\overline{S}}) = Av_S$. Furthermore, as $v \neq 0$, we must have $-v_{\overline{S}} \neq v_S$. Then, we conclude that $||v_S||_1 < ||v_{\overline{S}}||_1$.

Conversely, assume that the null space property relative to S holds. For a given vector $x \in \mathbb{K}^{m\times N}$ supported on S and a vector $z \in \mathbb{K}^{m\times N}$, with $z \neq x$ and satisfying $Az = Ax$, we consider the vector $v = x - z \in \ker A\backslash\{0\}$. Due to the null space property, we have

$$||x||_1 \leq ||x - z_S||_1 + ||z_S||_1 = ||v_S||_1 + ||z_S||_1 < ||v_{\overline{S}}||_1 + ||z_S||_1 = || - z_{\overline{S}}||_1 + ||z_S||_1 = ||z||_1.$$

□

Now, if we let the set S vary, we obtain as a obvious consequence the following result.

Theorem 3.3. *Given a matrix $A \in \mathbb{K}^{m\times N}$, every s-sparse vector $x \in A \in \mathbb{K}^N$ is the unique solution of (P_1) if and only if A satisfies the null space property of order s.*

This impressive and simple theorem tells us that if we have a sparse vector x and a measurement vector $y = Ax$, Basis Pursuit actually solves (P_0), provided that the NSP holds for the matrix A. In fact, suppose that we are able to recover the vector x via ℓ_1-norm minimization from $y = Ax$. Then, let $\tilde{x}$ be the minimizer of (P_0) with $y = Ax$. Therefore, of course, we have $||\tilde{x}||_0 \leq ||x||_0$ and so $\tilde{x}$ is also s-sparse, since by hypothesis we have that x is s-sparse. Since every s-sparse vector is the unique ℓ_1-minimizer, it follows that $\tilde{x} = x$.

The Null Space Property has an interesting feature: it is preserved if we rescale and reshuffle the measurements or even if some new measurements were added. From a mathematical point of view, this is the same as replacing the measurement matrix A

$$A_1 = GA, \qquad \text{where } A \text{ is some invertible } m \times m \text{ matrix.}$$

$$A_2 = \left[\frac{A}{B}\right], \qquad \text{where } B \text{ is some } m' \times N \text{ matrix.}$$

In the second case, A_2 is represented as a block matrix and B represents the additional measurements. This is reasonable because the measurement process should not get worse if: new information is acquired; if the order in which the measurements are obtained changes; or if we change the scale of all measurements of the signal. The justification for such facts is simple, note that $\ker A_1 = \ker A$ and $\ker A_2 \subset \ker A$ hence A_1 and A_2 have the NSP property if A does.

In the definition of NSP, $\mathbb{K}$ could be either $\mathbb{R}$ or $\mathbb{C}$. However, when the system has real coefficients, sparse solutions can be considered either as real or complex vectors, leading to two apparently distinct null space properties. This is so because there is a distinction between the real null space $\ker_{\mathbb{R}} A$ and the complex null space $\ker_{\mathbb{C}} A = \ker_{\mathbb{R}} A + i \ker_{\mathbb{R}} A$. Nevertheless, the real and complex NSP are equivalent, as proved by [Foucart & Gribonval '10]. While the original proof links it with a problem about convex polygons in the real plane, in Theorem 3.4, we prove this equivalence following the elementary argument of [Lai & Liu '11].

Theorem 3.4. (Theorem 1 of [Foucart & Gribonval '10]): *Given a matrix $A \in \mathbb{R}^{m\times N}$, the real null space property relative to a set S, that is,*

$$\sum_{j\in S}|v_j| < \sum_{i\in\overline{S}}|v_i| \qquad \forall v \in \ker_{\mathbb{R}} A,\ v \neq 0. \tag{3.2}$$

is equivalent to the complex null space property relative to this set S, that is,

$$\sum_{j\in S}\sqrt{v_j^2+w_j^2} < \sum_{i\in\overline{S}}\sqrt{v_i^2+w_i^2} \qquad \forall v,w \in \ker_{\mathbb{R}} A,\ (v,w) \neq (0,0). \tag{3.3}$$

In particular, the real null space property of order s is equivalent to the complex null space property of order s.

As [Foucart & Gribonval '10] notes, *"Before stating the theorem, we point out that, for a real measurement matrix, one may also recover separately the real and imaginary parts of a complex vector using two real ℓ_1-minimizations - which are linear programs - rather than recovering the vector directly using one complex ℓ_1-minimization - which is a second order cone program."*.

Proof. Clearly equation (3.2) follows from equation (3.3) when we set $w = 0$. Then we need to prove the converse. Assume that (3.2) holds and consider v, w *in* $\ker_{\mathbb{R}} A$ with $(v,w) \neq (0,0)$. Suppose first that v and w are linearly dependent, e.g. $w = kv$ for some $k \in \mathbb{C}$. Therefore, equation (3.3) is deduced from

$$\sum_{j\in S}|v_j| < \sum_{i\in\overline{S}}|v_i| \Rightarrow \sum_{j\in S}\sqrt{k^2+1}|v_j| < \sum_{i\in\overline{S}}\sqrt{k^2+1}|v_i| \Rightarrow \sum_{j\in S}\sqrt{k^2v_j^2+v_j^2} < \sum_{i\in\overline{S}}\sqrt{k^2v_i^2+v_i^2}$$
$$\Rightarrow \sum_{j\in S}\sqrt{w_j^2+v_j^2} < \sum_{i\in\overline{S}}\sqrt{w_i^2+v_i^2}.$$

Now assume u and v are independent. In this case, $u = v\cos\theta + w\sin\theta \in \ker_{\mathbb{R}} A$ will be nonzero. So, the real null space property (3.2) leads, for any $\theta \in \mathbb{R}$, to

$$\sum_{j\in S}|v_j\cos\theta + w_j\sin\theta| < \sum_{i\in\overline{S}}|v_i\cos\theta + w_i\sin\theta|. \tag{3.4}$$

If we define, for each $k \in [N]$, $\theta_k \in [-\pi,\pi]$ through the equations

$$\cos\theta_k = \frac{v_k}{\sqrt{v_k^2+w_k^2}}, \qquad \sin\theta_k = \frac{w_k}{\sqrt{v_k^2+w_k^2}}.$$

Equation (3.4) can be reformulate as

$$\sum_{j\in S}\sqrt{v_j^2+w_j^2}|\cos(\theta-\theta_j)| < \sum_{i\in\overline{S}}\sqrt{v_i^2+w_i^2}|\cos(\theta-\theta_i)|.$$

We can integrate over $\theta \in [-\pi,\pi]$ and obtain

$$\sum_{j\in S}\sqrt{v_j^2+w_j^2}\int_{-\pi}^{\pi}|\cos(\theta-\theta_j)|d\theta < \sum_{i\in\overline{S}}\sqrt{v_i^2+w_i^2}\int_{-\pi}^{\pi}|\cos(\theta-\theta_i)|d\theta.$$

Since $\int_{-\pi}^{\pi}|\cos(\theta-\theta_j)|d\theta = 4$ is independent of θ_i we have the desired inequality. Then the real null space property implies the complex null space property.

□

Sometimes it makes sense to use ℓ_q-norm minimization, for $0 < q < 1$ instead of ℓ_1-norm minimization. In order to analyze this case, the Null Space Property should be generalized. This was done by [Gribonval & Nielsen '07], see also [Gao, Peng, Yue & Zhao '15]. The same problem of the equivalence between real and complex Null Space Property arises in this nonconvex case. The proof that they are indeed equivalent notions was given by [Lai & Liu '11].

3.3 Recovery via Nonconvex Minimization

The ℓ_q-quasinorm of a vector can be used to approximate the number of nonzero entries of a vector, as

$$\sum_{j=1}^{N} |x_i|^q \xrightarrow{q\to 0} \sum_{i=1}^{N} \mathbf{1}_{\{x_i\neq 0\}} = ||x||_0.$$

Therefore, we would like to know whether we can obtain results for sparse recovery using ℓ_q-norm minimization, with $0 < q < 1$. Similarly to the other cases, let us call the problem of minimizing $||x||_q$ subject to $Ax = y$ of (P_q). Despite being a nonconvex problem, it can be shown that exact reconstruction is possible with substantially fewer measurements while maintaining robustness to noise and stability to signal non-sparsity, see [Chartrand '07] for numerical experiments and [Saab, Chartrand & Yilmaz '08] for the robustness analysis in the nonconvex case.

We will prove that the problem (P_q) does not provide a worse approximation to (P_0) when we make q smaller. In order to do this, we need to show an equivalence similar to Theorem 3.3 between sparse recovery with the ℓ_q-norm and some form of the Null Space Property.

Theorem 3.5. *Given a matrix $A \in \mathbb{C}^{m\times N}$ and $0 < q \leq 1$, every s-sparse vector $x \in \mathbb{C}^N$ is the unique solution of (P_q) with $y = Ax$ in and only if, for any $S \subset [N]$ with $\#S \leq s$,*

$$||v_S||_q < ||v_{\overline{S}}||_q \qquad \forall v \in \ker A\backslash\{0\}.$$

The proof of this theorem is analogous to that of Theorem 3.2. We just need to use the fact that the qth power of the ℓ_q-quasinorm satisfies the triangular inequality. Using Theorem 3.5, [Gribonval & Nielsen '07] were able to establish that sparse recovery via ℓ_q-minimization implies sparse recovery via ℓ_p-minimization for $0 < p < q \leq 1$.

Theorem 3.6. ([Gribonval & Nielsen '07]): *Given a matrix $A \in \mathbb{C}^{m\times N}$ and $0 < p < q \leq 1$, if every s-sparse vector $x \in \mathbb{C}^N$ is the unique solution of (P_q) with $y = Ax$, then $x \in \mathbb{C}^N$ is also the unique solution of (P_p) with $y = Ax$.*

Proof. Due to Theorem 3.5, we just need to prove that, if $v \in \ker A\backslash\{0\}$ and if S is an index set of s largest absolute entries of v, then

$$\sum_{i\in S} |v_i|^p < \sum_{i\in\overline{S}} |v_i|^p. \tag{3.5}$$

We have, by hypothesis, that the same inequality is valid with q in place of p. Then we have $v_{\overline{S}} \neq 0$ since S is an index set of the largest absolute entries and $v \neq 0$. So we can rewrite (3.5) dividing by $\sum_{i\in\overline{S}} |v_i|^p$. So, the inequality we want to prove is equivalent to

$$f(p) := \sum_{i\in S} \frac{1}{\sum_{j\in\overline{S}} (|v_j|/|v_i|)^p} < 1. \tag{3.6}$$

Clearly we have that $|v_j|/|v_i| \leq 1$ for $j \in \overline{S}$ and $i \in S$. So, $f(p)$ is a nondecreasing function of p. Hence, for $p < q$, we have $f(p) \leq f(q)$. By hypothesis, we have $f(q) < 1$ and this implies the result. □

In these results about sparse recovery via convex (and also nonconvex) optimization problem, nothing is said about the computational tractability of the Null Space Property. In Section 3.7 we will study some computational complexity issues related to ℓ_q-norm minimization for $0 < q \leq 1$. We first need to analyze what happen if the measurements are corrupted or if we have vectors which are not, in fact, sparse, but approximately sparse.

3.4 Stable Measurements

In idealized situations, all the vectors we deal with are sparse. This may not be the case in more realistic scenarios. Instead, we can have approximately sparse vectors and we want to recover them with an error controlled by its distance to the closest sparse vector. This is called *stability* of the reconstruction with respect to sparsity defect. In this section, we will prove that Basis Pursuit is stable. In order to do this, we need a stronger Null Space Property that allow to encompass these defects into the kernel of the measurement matrix.

Definition 3.7. A matrix $A \in \mathbb{C}^{m\times N}$ is said to satisfy the *stable null space property* with constant $0 < \rho < 1$ relative to a set $S \subset [N]$ if

$$||v_S||_1 \leq \rho||v_{\overline{S}}||_1 \qquad \forall v \in \ker A.. \tag{NSP$_\rho$}$$

We say that A satisfies the *stable null space property of order s with constant* $0 < \rho < 1$ if it satisfies the stable null space property with constant $0 < \rho < 1$ relative to any set $S \subset [N]$ with $\#S \leq s$.

We now can generalize Theorem 3.3 to the stable case.

Theorem 3.8. *Suppose that a matrix $A \in \mathbb{C}^{m\times N}$ satisfies the stable null space property of order s with constant $0 < \rho < 1$. Then, for any $x \in \mathbb{C}^N$, a solution $x^\#$ of (P_1) with $y = Ax$ approximates the vector x with ℓ_1-error*

$$||x - x^\#||_1 \leq \frac{2(1+\rho)}{(1-\rho)}\sigma_s(x)_1. \tag{3.7}$$

Remark 15. *It is interesting to note that Theorem 3.8 does not guarantee uniqueness for the ℓ_1-minimization. However, without explaining why, [Rauhut & Foucart '13] argue that nonuniqueness is rather pathological.*

Albeit there is nonuniqueness, the content of Theorem 3.8 is that every solution $x^\#$ of (P_1) with $y = Ax$ satisfies (3.7). We will prove, in fact, a stronger result, namely Theorem 3.9. The key point is to look for any vector $z \in \mathbb{C}^N$ satisfying $Az = Ax$ and establish that, under the stable null space property relative to a set S, the distance between a vector $x \in \mathbb{C}^N$ supported on S and a vector $z \in \mathbb{C}^N$ satisfying $Az = Ax$ can be controlled by the difference of their norms. Theorem 3.8 will follow as a corollary.

Theorem 3.9. *The matrix $A \in \mathbb{C}^{m\times N}$ satisfies the stable null space with constant $0 < \rho < 1$ relative to S if and only if*

$$||z - x||_1 \leq \frac{1+\rho}{1-\rho}(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1), \tag{3.8}$$

for all vectors $x, z \in \mathbb{C}^N$ with $Az = Ax$.

In order to prove this result, we need the following simple observation.

Lemma 3.10. *Given a set $S \subset [N]$ and vectors $x, z \in \mathbb{C}^N$, the following inequality holds*

$$||(x-z)_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + ||(x-z)_S||_1 + 2||x_{\overline{S}}||_1.$$

Proof. We need to separate a vector $x \in \mathbb{C}^N$ into two parts, one relative to the set S and other relative to the complementary set $\overline{S}$. So, given a set $S \subset [N]$ and vectors $x, z \in \mathbb{C}^N$, we have

$$||x||_1 = ||x_{\overline{S}}||_1 + ||x_S||_1 \leq ||x_{\overline{S}}||_1 + ||(x-z)_{\overline{S}}||_1 + ||z_S||_1,$$

and also

$$||(x-z)_{\overline{S}}||_1 \leq ||x_{\overline{S}}||_1 + ||z_{\overline{S}}||_1.$$

We sum these two inequality to conclude

$$||x||_1 + ||(x-z)_{\overline{S}}||_1 \leq 2||x_{\overline{S}}||_1 + ||(x-z)_{\overline{S}}||_1 + ||z||_1.$$

□

Proof. (of Theorem 3.9): We will assume that the matrix A satisfies (3.8) for all vectors $x, z \in \mathbb{C}$ with $Az = Ax$. For $v \in \ker A$, $A(v_S + v_{\overline{S}}) = Av = 0$ and so $A(-v_{\overline{S}}) = Av_S$. From (3.8) applied to $x = -v_S$ and $z = v_{\overline{S}}$ we obtain

$$||v||_1 \leq \frac{1+\rho}{1-\rho}\big(||v_{\overline{S}}||_1 - ||v_S||_1\big),$$

which is equivalent to

$$(1-\rho)\big(||v_S||_1 + ||v_{\overline{S}}||_1\big) \leq (1+\rho)\big(||v_{\overline{S}}||_1 - ||v_S||_1\big).$$

After rearranging, we have $||v_S||_1 \leq \rho||v_{\overline{S}}||_1$ and this is the stable null space property with constant $0 < \rho < 1$ relative to the set S. Conversely, assume that the matrix A satisfies this property. For $x, z \in \mathbb{C}^N$ with $Az = Ax$, since $v = z - x \in \ker A$, the stable null space property yields $||v_S||_1 \leq \rho||v_{\overline{S}}||_1$. Then Lemma 3.10 leads to

$$||v_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + ||v_S||_1 + 2||x_{\overline{S}}||_1. \tag{3.9}$$

Putting the definition of the stable null space property into (3.9), we obtain

$$||v_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + \rho||v_{\overline{S}}||_1 + 2||x_{\overline{S}}||_1,$$

and since $\rho < 1$, this is the same as

$$||v_{\overline{S}}||_1 \leq \frac{1}{1-\rho}\big(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1\big).$$

Using the definition of the stable null space property again, we conclude

$$||v||_1 = ||v_{\overline{S}}||_1 + ||v_S||_1 \leq (1+\rho)||v_{\overline{S}}||_1 \leq \frac{1+\rho}{1-\rho}\big(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1\big).$$

□

In [Rauhut & Foucart '13] it was pointed out that if sparse vectors are exactly recovered then the stability is obtained at no cost. To see how, let us consider for each index set $s \subset [N]$ with $\#S \leq s$, the operator R_s defined on $\ker A$ by $R_s(v) = v_s$. So we have the following equivalence for the definition of null space property,

$$2||v_S||_1 < ||v||_1 \qquad \forall v \in \ker A \backslash \{0\} \iff \mu := \max\{||R_S||_{1\to 1} : S \subset [N], \#S \leq s\} < 1/2.$$

It then follows that A satisfies NSP_ρ with constant $\rho = \mu/(1-\mu) \leq 1$. Then, we have stability. However, note that the constant $2(1+\rho)/(1-\rho)$ in (3.7) can be very large as $\rho \to 1$.

Now, we can deduce Theorem 3.8 as a Corollary.

Proof. (of Theorem 3.8): We just need to prove that $\frac{1+\rho}{1-\rho}(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1)$ can be dominated by $\frac{2(1+\rho)}{(1-\rho)}\sigma_s(x)_1$. In order to do this, we take S as the set of s largest absolute coefficients of x, so $||x_{\overline{S}}||_1 = \sigma_s(x)_1$. Also, if $x^\#$ is a minimizer of (P_1), then $||x^\#||_1 \leq ||x||_1$ and $Ax^\# = Ax$. Therefore we just need to consider $z = x^\#$ and this concludes the proof.

□

Sometimes we can deal with a fixed sparse vector rather than all vectors with a given sparsity. This will be the case in the next two theorems, where we give a geometric characterization for the uniqueness of the ℓ_1-minimization problem.

[Chandrasekaran, Recht, Parrilo & Willsky '12] introduced a general geometric theory for the recovery of "simple" objects, in their own words, from few measurements. They analyzed the connections between the recovery, from limited linear measurements, of objects with some structure and the convex hull of the set of these objects. This was inspired by the fact that the convex hull of (unit Euclidean-norm) one-sparse vectors is the unit ball of the ℓ_1-norm and, similarly, the convex hull of the (unit Euclidean-norm) rank-one matrices is the nuclear norm ball, as we will discuss in Section 3.6.

Definition 3.11. For a vector $x \in \mathbb{R}^N$, the *descent convex cone* is given by

$$T(x) = \text{cone}\{z - x : z \in \mathbb{R}^N, ||z||_1 \leq ||x||_1\},$$

where *cone* means the conic hull of a set.

Theorem 3.12. (Proposition 2.1 of [Chandrasekaran, Recht, Parrilo & Willsky '12]): *For $A \in \mathbb{R}^{m\times N}$, a vector $x \in \mathbb{R}^N$ is the unique minimizer of $||z||_1$ subject to $Az = Ax$ if and only if* $\ker A \cap T(x) = \{0\}$.

Proof. Suppose that $\ker A \cap T(x) = \{0\}$. Let $x^\#$ be an ℓ_1-minimizer. So we have $||x^\#||_1 \leq ||x||_1$ and $Ax^\# = Ax$. And then $v = x^\# - x$ must belong to $T(x) \cap \ker A = \{0\}$. Therefore, $x^\# = x$. Then x is the unique ℓ_1-minimizer.

Now assume that x is the unique ℓ_1-minimizer. Every vector $v \in T(x)\backslash\{0\}$ can be written as $v = \sum \alpha_i(z_i - x)$ with $\alpha_i \geq 0$ (see Lemma 2.6 of [Ruszczynski]) and $||z_i||_1 \leq ||x||_1$. As $v \neq 0$, we have $\sum \alpha_i > 0$ and so we can take the normalized coefficients $\alpha_i' = \alpha_i / \sum \alpha_i$. If $v \in \ker A$, we have $A(\sum \alpha_i' z_i) = Ax$ and also $||\sum \alpha_i' z_i||_1 \leq \sum \alpha_i' ||z_i||_1 \leq ||x||_1$. By uniqueness of the ℓ_1-minimizer, $\sum \alpha_i' z_i = x$ and so $v = 0$, which is a contradiction. Therefore we have $(T(x)\backslash\{0\}) \cap \ker A = \emptyset$ and then $T(x) \cap \ker A = \{0\}$.

□

This theorem shows us that we can rewrite exact recovery from ℓ_1-minimization purely in terms of convex geometry. A proper null space for x will be oriented in such a way that its shift by x will touch $T(x)$ uniquely at x. Also, this geometric characterization can be extended to robust recovery as the following theorem shows.

Theorem 3.13. (Proposition 2.2 of [Chandrasekaran, Recht, Parrilo & Willsky '12]): *For $A \in \mathbb{R}^{m\times N}$, let $x \in \mathbb{R}^N$ and $y = Ax + e \in \mathbb{R}^m$ with $||e||_2 \leq \eta$. If*

$$\inf_{v\in T(x), ||v||_2=1} ||Av||_2 \geq \tau$$

for some $\tau > 0$, then a minimizer $x^\#$ of $||z||_1$, subject to $||Az - y||_2 \leq \eta$ satisfies

$$||x - x^\#||_2 \leq \frac{2\eta}{\tau}.$$

Proof. We may assume that $x^\# - x \neq 0$, since otherwise the result is trivial. As $x^\#$ is a minimizer of $||z||_1$. we have $||x^\#||_1 \leq ||x||_1$ and this leads to fact that $v = (x^\# - x)/||x^\# - x||_2$ belongs to $T(x)$. Since $||v||_2 = 1$ and $v \in T(x)$, our hypothesis says that $||Av||_2 \geq \tau$. Using this and the triangular inequality, we have

$$||x^\# - x||_2 \leq \frac{1}{\tau}||A(x^\# - x)||_2 \leq \frac{1}{\tau}\Big(||Ax^\# - y||_2 + ||Ax - y||_2\Big) \leq \frac{2\eta}{\tau}.$$

□

3.5 Robust Measurements

Due to physical limitations there is no hope to measure a signal $x \in \mathbb{C}^N$ with infinite precision. Hence the measurement vector $y \in \mathbb{C}^m$ will be always an approximation to the vector $Ax \in \mathbb{C}^m$. We want to model our problem in such a way that we can control the error of this approximation by

$$||Ax - y|| \leq \eta,$$

for some $\eta \geq 0$ and some norm $||.||$ on $\mathbb{C}^m$. Typically we will require finite energy of this difference which, in mathematical terms, means that will we work with the ℓ_2-norm. Then, the output of our reconstruction technique cannot be x, but will be $x^* \in \mathbb{C}^N$ whose distance to the original vector $x \in \mathbb{C}^N$ must be controlled by $\eta \geq 0$. This is called the *robustness* of the reconstruction scheme with respect to the measurement error. Let us then substitute the original basis pursuit (P_1) by the following convex problem

$$\min_{z \in \mathbb{C}^N} ||z||_1 \quad \text{subject to } ||Az - y|| \leq \eta, . \tag{$P_{1,\eta}$}$$

where the norm in $||Az - y||$ is a properly chosen norm for the measurement error.

Definition 3.14. The matrix $A \in \mathbb{C}^{m \times N}$ is said to satisfy the *robust null space property* (with respect to the norm $||.||$) with constants $0 < \rho < 1$ and $\tau > 0$ relative to a set $S \subset [N]$ if

$$||v_S||_1 \leq \rho ||v_{\overline{S}}||_1 + \tau ||Av|| \qquad \forall v \in \mathbb{C}^n. \tag{$\text{NSP}_{\rho,\tau}$}$$

Likewise in the definition of the stable null space property, we say that a matrix A satisfies the *robust null space property of order s* with constants $0 < \rho < 1$ and $\tau > 0$ if it satisfies the robust null space property with constants ρ, τ to any set $S \subset [N]$ with $\#S \leq s$.

Remark 16. *In this definition we do not ask for v to be in* $\ker A$. *Note that if $v \in \ker A$, then the term $||Av||$ vanishes and we recover the stable null space property.*

We can now generalize Theorem 3.8 to the case $\eta \neq 0$.

Theorem 3.15. *Suppose that a matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x \in \mathbb{C}^N$, a solution $x^\#$ of $(P_{1,\eta})$ with $y = Ax + e$ and $||e|| \leq \eta$ satisfies*

$$||x - x^\#||_1 \leq \frac{2(1+\rho)}{(1-\rho)} \sigma_s(x)_1 + \frac{4\tau}{1-\rho}\eta.$$

As in Section 3.4, we will prove a stronger statement valid for any index set S and obtain Theorem 3.15 as a particular case.

Theorem 3.16. *The matrix $A \in \mathbb{C}^{m \times N}$ satisfies the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ relative to S if and only if*

$$||z - x|| \leq \frac{(1+\rho)}{(1-\rho)}(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1) + \frac{2\tau}{1-\rho}||A(z-x)||, \tag{3.10}$$

for all vectors $x, z \in \mathbb{C}$.

Proof. The idea of the proof, *mutatis mutandis*, is same from Theorem 3.9. Let us first assume that the matrix A satisfies (3.10) for all $x, z \in \mathbb{C}^N$. Then, for any $v \in \mathbb{C}^N$, we can take $x = -v_S$ and $z = v_{\overline{S}}$. This leads to

$$||v||_1 \leq \frac{1+\rho}{1-\rho}\left(||v_{\overline{S}}||_1 - ||v_S||_1\right) + \frac{2\tau}{1-\rho}||Av||.$$

After rearranging, we obtain $(1-\rho)\left(||v_S||_1 + ||v_{\overline{S}}||_1\right) \leq (1+\rho)\left(||v_{\overline{S}}||_1 - ||v_S||_1\right) + 2\tau||Av||$, and this is the same as $||v_S||_1 \leq \rho||v_{\overline{S}}||_1 + \tau||Av||$. This is exactly the robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ relative to S.

Conversely, assume that the matrix A satisfies the robust null space property with constant $0 < \rho < 1$ and $\tau > 0$ relative to S. For $x, z \in \mathbb{C}^N$, we set $v = z - x$ and then, using $\mathrm{NSP}_{\rho,\tau}$ and Lemma 3.10 yields

$$||v_S||_1 \leq \rho||v_{\overline{S}}||_1 + \tau||Av||,$$
$$||v_{\overline{S}}||_1 \leq ||z||_1 - ||x||_1 + ||v_S||_1 + 2||x_{\overline{S}}||_1.$$

Combining both inequalities we have

$$||v_{\overline{S}}||_1 \leq \frac{1}{1-\rho}\left(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1 + \tau||Av||\right).$$

Using $\mathrm{NSP}_{\rho,\tau}$ once more, we obtain

$$||v||_1 = ||v_{\overline{S}}||_1 + ||v_S||_1 \leq (1+\rho)||v_{\overline{S}}||_1 + \tau||Av|| \leq \frac{1+\rho}{1-\rho}\left(||z||_1 - ||x||_1 + 2||x_{\overline{S}}||_1\right) + \frac{2\tau}{1-\rho}||Av||.$$

□

We will now establish a result where the ℓ_1-error estimate is replaced by a general ℓ_p-error estimate for $p \geq 1$. This means that, from now on, we have other error reconstruction rates available. In order to do this, we need to change and strength the null space property to include other norms than the ℓ_1. Now we will directly define the property of order s instead of define it for a fixed set $S \subset [N]$ first.

Definition 3.17. Given $q \geq 1$, the matrix $A \in \mathbb{C}^{m\times N}$ is said to satisfy the *ℓ_q-robust null space property* of order s (with respect to the norm $||.||$) with constants $0 < \rho < 1$ and $\tau > 0$ if, for any set $S \subset [N]$ with $\#S \leq s$,

$$||v_S||_q \leq \frac{\rho}{s^{1-1/q}}||v_{\overline{S}}||_1 + \tau||Av|| \qquad \forall v \in \mathbb{C}^n. \tag{$\mathrm{NSP}_{q,\rho,\tau}$}$$

Remark 17. *Note that with the aid of the inequality $||v_S||_p \leq s^{1/p-1/q}||v_S||_q$ valid for $1 \leq p \leq q$ we can prove that the ℓ_q-robust null space property with constants $0 < \rho < 1$ and $\tau > 0$ implies that, for any set $S \subset [N]$ with $\#S \leq s$,*

$$||v_S||_p \leq \frac{\rho}{s^{1-1/p}}||v_{\overline{S}}||_1 + \tau s^{1/p-1/q}||Av|| \qquad \forall v \in \mathbb{C}^n.$$

Hence, if we change the norm $||.||$ by the new norm $s^{1/p-1/q}||.||$, the ℓ_q-robust null space property implies the ℓ_q-robust null space property with the same constants for any $1 \leq p \leq q$. Therefore, Definition 3.17 is a natural strengthening of the robust null space property for the ℓ_1 norm.

We will now show that quadratically constrained basis pursuit works for matrices satisfying $\mathrm{NSP}_{2,\rho,\tau}$. The precise statement is as follows.

Theorem 3.18. *Suppose that the matrix $A \in \mathbb{C}^{m\times N}$ satifies the ℓ_2-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x \in \mathbb{C}^N$, a solution $x^\#$ of $(P_{1,\eta})$ with $||.|| = ||.||_2$, $y = Ax + e$, and $||e||_2 \leq \eta$ approximates the vector x with ℓ_p-error*

$$||x - x^\#||_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}\eta, \qquad 1 \leq p \leq 2, \tag{3.11}$$

Consequently, if there is a way to characterize which matrices satisfy this property or, at least, find a useful set of matrices for which this property is valid, we can then ensure the recovery of approximately sparse vectors with controlled error rates. In Section 5.5, in particular Theorem 5.19, proved by [Andersson & Strömberg '14], says that this is the case for matrices with sufficiently small restricted

isometry constants. Nonetheless, it is important to note that the first result of stability and robustness of sparse reconstruction via Basis Pursuit was obtained by [Candès, Romberg & Tao I '06].

Next, we will prove a more general result from which Theorem 3.18 follows by choosing $q = 2$, taking $z = x^\#$ and observing that since $x^\#$ is solution of $\min_{z\in\mathbb{C}^N} ||z||_1$, we necessarily have $||x^\#||_1 - ||x||_1 \leq 0$.

Theorem 3.19. *Given $1 \leq p \leq q$, suppose that the matrix $A \in \mathbb{C}^{m\times N}$ satisfies the ℓ_q-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$. Then, for any $x, z \in \mathbb{C}^N$,*

$$||z - x||_p \leq \frac{C}{s^{1-1/p}}\Big(||z||_1 - ||x||_1 + 2\sigma_s(x)_1\Big) + Ds^{1/p-1/q}||A(z-x)||,$$

where $C = (1+\rho)^2/(1-\rho)$ and $D = (3+\rho)\tau/(1-\rho)$.

Proof. Recall that the ℓ_q-robust null space property implies the ℓ_1-robust null space property and ℓ_p-robust null space property for $p \leq q$. These two statements can be written as follows

$$||v_S||_1 \leq \rho||v_{\overline{S}}||_1 + \tau s^{1-1/q}||Av||. \tag{3.12}$$

$$||v_S||_p \leq \frac{\rho}{s^{1-1/p}}||v_{\overline{S}}||_1 + \tau s^{1/p-1/q}||Av||. \tag{3.13}$$

for all $v \in \mathbb{C}^N$ and all $S \subset [N]$ with $\#S \leq s$. Thus, considering (3.12) and using Theorem 3.16 with S chosen as an index set of s largest (in absolute value) entries of x we obtain

$$||z - x||_1 \leq \frac{1+\rho}{1-\rho}\left(||z||_1 - ||x||_1 + 2\sigma_s(x)_1\right) + \frac{2\tau}{1-\rho}s^{1-1/q}||A(z-x)||. \tag{3.14}$$

Now, considering (3.13) and choosing S as an index set of s largest (in absolute value) entries of $z - x$, we can estimate the ℓ_p-norm by the ℓ_1-norm and obtain

$$||z-x||_p \leq ||(z-x)_{\overline{S}}||_p + ||(z-x)_S||_p \leq \frac{1}{s^{1-1/p}}||z-x||_1 + ||(z-x)_S||_p$$

$$\leq \frac{1}{s^{1-1/p}}||z-x||_1 + \frac{\rho}{s^{1-1/p}}||(z-x)_{\overline{S}}||_1 + \tau s^{1/p-1/q}||A(z-x)|| \leq \frac{1+\rho}{s^{1-1/p}}||(z-x)||_1 + \tau s^{1/p-1/q}||A(z-x)||$$

$$\overset{(3.8)}{\leq} \frac{(1+\rho)^2}{(1-\rho)}\frac{1}{s^{1-1/p}}\Big(||z||_1 - ||x||_1 + 2\sigma_s(x)_1\Big) + \frac{(3+\rho)\tau}{(1-\rho)}s^{1/p-1/q}||A(z-x)||.$$

□

Returning to Theorem 3.18, let us analyze the extremal cases $p = 1$ and $p = 2$. Then , we have the estimates

$$||x - x^\#||_1 \leq C\sigma_s(x)_1 + D\sqrt{s}\eta \quad \text{and} \quad ||x - x^\#||_2 \leq \frac{C}{\sqrt{s}}\sigma_s(x)_1 + D\eta.$$

It is interesting to note that in the first case, the coefficient of $\sigma_s(x)_1$ is constant while in the second one is proportional to $1/\sqrt{s}$. Also, the coefficient of η scales like $\sqrt{s}$ for $p = 1$ while it is constant for $p = 2$. Another curious point to note is that regardless of the chosen norm in which we seek for error estimates, the error $\sigma_s(x)_1$ always appears on the right-hand side. So, for example, in the case $p = 2$, one may inquire why $\sigma_s(x)_2$ does not appear, since we have an ℓ_2-error estimate, but instead $\sigma_s(x)_1/\sqrt{s}$ appears in its place.

In Chapter 8 we will see that such kind of estimate, with $\sigma_s(x)_2$, is impossible in the parameters regime we are looking for. In order to acquire information in a compressible fashion, m must be much smaller than N. After developing some techniques on Geometry of Banach Spaces, we will prove in Theorem 8.21 a curious condition. Namely, if we want estimates with $\sigma_s(x)_2$, then necessarily $m \geq cN$ for some positive constant c, so this ℓ_2 estimates are useless for our purposes.

Lastly, we saw on Chapter 1 that unit ℓ_1-balls with $q < 1$ can model compressible vectors in a satisfactory way. Indeed, by Proposition 1.5, if $||x||_1 \leq 1$ for $q < 1$, then, for $p \geq 1$, we have $\sigma_s(x)_p \leq$

$s^{1/p-1/q}$. On the other hand, taking $\eta = 0$ on Equation (3.11), which is the same as consider measurements without error, we have

$$||x - x^{\#}||_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1 \leq \frac{C}{s^{1-1/p}}s^{1-1/q} = Cs^{1/p-1/q}, \qquad 1 \leq p \leq 2.$$

Hence, we have the *same decay error rate* for the reconstruction error in ℓ_p and the best s-term approximation, provided that $p \in [1, 2]$. Therefore the terms $\sigma_s(x)_1/s^{1-1/p}$ and $\sigma_s(x)_1$ are comparable and the appearance of $\sigma_s(x)_1$ on the right-hand side makes sense, see the comments after Definition 8.19.

3.6 Low-Rank Matrix Recovery

In this dissertation, we will be interested in the problem of sparse *vector* recovery. However, the recovery/reconstruction of higher dimensional structures like matrices and higher order tensors is also possible from few observations. The matrix case is specially important in many practical problems. Consider, for example, a survey data containing the responses of various individuals to specific questions. We can make a table with individuals in the rows and questions in the columns. In any quiz, many questions are left unanswered and the aim is to provide a good estimate for the missing answers. Even more, typically *many* questions will be left with no answer, so we want to recover a matrix from very few measurements.

Matrix-completion problems arise in a natural way into fields where questions about dimensionality or complexity are present. Typically we can model these as a problem about the rank of some appropriate matrix. The main point is that in these cases, the matrix we wish to recover is known to be structured in the sense that it is low-rank or approximately low-rank.

Some applications of the techniques of low-rank matrix recovery are: phase retrieval of diffracted waves in crystallographic and astronomical imaging [Candès, Eldar, Strohmer & Voroninski '13]; low-dimensional embedding of data and the connectivity structure of graphs which arise in the study of social networks [Linial, London & Rabinovich '15]; distance matrices which represents wireless sensor network localization [So & Ye '07] and multi-task learning and recommendation systems in machine learning [Rohde & Tsybakov '11], just to name a few. This last one became very famous after the *Netflix Problem*[3], where some techniques related to low-rank matrix recovery were used.

Solving all of these problems efficiently is particularly important in view of the massive size of actual datasets. Even more, it would be impossible to fully observe the matrices arising in those applications. As [Davenport & Romberg '16] points out, *"it can be prohibitively expensive to fully sample the entire output of a sensor array; we might only be able to measure the strength of a few connections in a graph; and any particular user of a recommendation system will provide only a few ratings"*.

In this section, we briefly explore the connection between the problem of *Low-Rank Matrix Recovery* and Compressive Sensing. More specifically, we will see that one problem can be realized as a particular instance of the other, when we change the ℓ_1-norm minimization by the nuclear norm, defined below. This connection was first explored by [Candès & Recht '09]; important contributions were made by [Recht, Fazel & Parrilo '10], [Candès & Tao '10] and [Keshavan, Montanari & Oh '10].

Suppose that a matrix $A \in \mathbb{C}^{n_1 \times n_2}$ of rank at most r is observed via linear measurements described by

$$\mathcal{A} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m, \quad X \mapsto \sum_{j=1}^{m} \langle X, A_j \rangle e_j = \sum_{j=1}^{m} \operatorname{tr}(XA_j^*)e_j$$

where $A_1, \ldots, A_n$ are suitable $n_1 \times n_2$ matrices and $e_1, \ldots, e_m$ is the standard basis in $\mathbb{R}^m$. If we have noise, then our measurements will be described by $y = \mathcal{A}(X) + e$, where $e \in \mathbb{R}^m$ denotes the additive noise. Like the vectorial case, our problem here is given by

$$\min_{X \in \mathbb{C}^{n_1 \times n_2}} \operatorname{rank}(X) \qquad \text{subject to } \mathcal{A}(X) = y \tag{NSP$_{rank}$}$$

[3] See the websites `http://www.netflixprize.com/` and `https://www.cs.uic.edu/~liub/Netflix-KDD-Cup-2007.html` for the problem description and more information about the solution.

Note that the rank of X is the ℓ_0-norm of the vector $[\sigma_1(X), \ldots, \sigma_n(X)]$ of singular values of X. When the matrix is diagonal, this problem reduces to the (P_0) problem and then it is also NP-hard. Therefore, we need a good relaxation strategy. To solve the problem we will, instead of pursuing rank minimization, try to obtain nuclear norm minimization. This norm, also known by the names Schatten 1-norm and Ky-Fan norm, is defined by $||X||_* = \sum_{j=1}^{n} \sigma_j(X)$ with $n = \min\{n_1, n_2\}$. Let us first recall the following definition.

Definition 3.20. Let $\mathcal{C}$ be a convex set. The convex envelope of a (possible nonconvex) function $f : \mathcal{C} \to \mathbb{R}$ is defined as the largest convex function g such that $g(x) \leq f(x)$ for all $x \in \mathcal{C}$. More precisely,

$$\text{conv } f(x) = \sup\{g(x) \mid g \text{ convex}, g \leq f\}.$$

Remark 18. *From this definition we see that, among all convex functions that bound $f(x)$ from below, g is the best approximation and so instead of minimizing f, we can try to efficiently minimize g.*

Now we state a result about the convex envelope of the matrix rank.

Theorem 3.21. (Theorem 1 of [Fazel, Hindi & Boyd '01]): *The convex envelope of the function $\phi(X) = rank(X)$ on $C = \{X \in \mathbb{R}^{m \times n} \big| ||X||_{2\to 2} \leq 1\}$ is $\phi_{env}(X) = ||X||_*$*

[Fazel, Hindi & Boyd '01] proposed the heuristic of the nuclear norm minimization (three years before the first preprint about Compressive Sensing was released and therefore, without the analogy with sparse recovery and basis pursuit) and this was fully explored in the thesis [Fazel '02].

$$\min_{X \in \mathbb{C}^{n_1 \times n_2}} ||X||_* \qquad \text{subject to } \mathcal{A}(X) = y \tag{$\text{P}_{nuclear}$}$$

This is a convex optimization problem, and therefore, at least in principle, easily solvable. If the matrix variable X is symmetric and positive semidefinite, then its singular values are the same as its eigenvalues and the problem (NSP_{nuc}) reduces to the trace minimization. Moreover, this norm is the ℓ_1-norm of the singular values of X and then we can try to import the definitions and techniques from the vector case to this case. First of all, we have a matrix version of the null space property. Second, the success of the nuclear norm minimization for the recovery of low-rank matrices is also equivalent to matrix NSP as state in the next theorem.

Theorem 3.22. ([Recht, Xu & Hassibi '08]): *Give a linear map $\mathcal{A}$ from $\mathbb{C}^{n_1 \times n_2}$ to $\mathbb{C}^m$, every matrix $X \in \mathbb{C}^{n_1 \times n_2}$ of rank r is the unique solution of $(P_{nuclear})$ with $y = \mathcal{A}(X)$ if and only if, for all $M \in \ker \mathcal{A} \backslash \{0\}$ with singular values $\sigma_1(M) \cdots \geq \sigma_n(M) \geq 0$, $n = \min\{n_1, n_2\}$,*

$$\sum_{j=1}^{r} \sigma_j(M) < \sum_{j=r+1}^{n} \sigma_j(M).$$

In order to prove this result, we need the classical variational characterization of the singular values. Here, we present the proof given by [Rauhut & Foucart '13].

Theorem 3.23. (Courant-Fischer Minimax Theorem): *For a self-adjoint matrix $A \in \mathbb{C}^{m \times n}$, the eigenvalues of A are obtained through*

$$\lambda_k(A) = \max_{\substack{M \subset \mathbb{C}^n \\ \dim M = k}} \min_{\substack{x \in M \\ ||x|_2 = 1}} \langle Ax, x \rangle. \tag{3.15}$$

As a consequence, the singular values $\sigma_1(A) \geq \cdots \geq \sigma_{\min\{m,n\}} \geq 0$ of a matrix $A \in \mathbb{C}^{m \times n}$ can be described by

$$\sigma_k(A) = \max_{\substack{M \subset \mathbb{C}^n \\ \dim M = k}} \min_{\substack{x \in M \\ ||x|_2 = 1}} ||Ax||_2$$

Proof. Let us first prove that $\lambda_k(A)$ is not larger than the right-hand side of Equation (3.15). Taking $\{u_1, \dots, u_n\}$ as an orthonormal basis and $M = \text{span}(u_1, \dots, u_n)$, for $x = \sum_{j=1}^k \alpha_j u_j \in M$ with unit norm we have

$$\langle Ax, x\rangle = \sum_{j=1}^k \lambda_j(A)\alpha_i^2 \geq \lambda_k(A) \sum_{j=1}^k \alpha_i^2 = \lambda_k(A)||x||_2^2 = \lambda_k(A).$$

Now, we need to prove the converse inequality for $\lambda_k(A)$. So, given a k-dimensional subspace M of $\mathbb{C}^n$, we choose a vector $x \in M \cap \text{Span}\{u_k, \dots, u_n\}$ with $||x||_2 = 1$ in such a way that $x = \sum_{j=1}^k \alpha_j u_j$. This leads to

$$\langle Ax, x\rangle = \sum_{j=k}^n \lambda_j(A)\alpha_i^2 \leq \lambda_j(A) \sum_{j=k}^n \alpha_i^2 = \lambda_k(A)||x||_2^2 = \lambda_k(A)$$

□

Let $\boldsymbol{\lambda}(X) = (\lambda_1(X), \dots, \lambda_n(X))$ be the spectrum of any matrix X, considered as a vector, instead of a set. In the short note [Lidskii '50], it was proved that for Hermitian matrices A and B, $\boldsymbol{\lambda}(A+B) - \boldsymbol{\lambda}(A)$ is in the convex hull spanned by $P\boldsymbol{\lambda}(B)$, where P is a permutation matrix. The Birkhoff-von Neumann Theorem states that the set of doubly stochastic matrices, which is a convex polytope, is the convex hull of the set of permutation matrices. So Lidskii's result is equivalent to $\boldsymbol{\lambda}(A+B) - \boldsymbol{\lambda}(A) = O\boldsymbol{\lambda}(B)$ for some doubly stochastic matrix O and this, in turn, is equivalent to the following inequality.

Lemma 3.24. ([Lidskii '50] and [Wielandt '55]): *Let $A, B \in \mathbb{C}^{n\times n}$ be two self-adjoint matrices, and let $(\lambda_j(A))_{j\in[n]}$, $(\lambda_j(B))_{j\in[n]}$, $(\lambda_j(A+B))_{j\in[n]}$ denote the eigenvalues of A, B, and $A+B$ arranged in nonincreasing order. For any $1 \leq i_1 < \cdots < i_k \leq n$,*

$$\sum_{i=1}^k \lambda_{i_j}(A+B) \leq \sum_{i=1}^k \lambda_{i_j}(A) + \sum_{i=1}^k \lambda_i(B).$$

Proof. Note that the inequality is invariant under the change $B \to B - \alpha\text{Id}$, for a given constant α and so, without loss of generality, we may assume that we have translated the spectrum in such a way that $\lambda_{k+1}(B) = 0$. Therefore all the vectors $\lambda_{k+2}(B), \lambda_{k+3}(B), \dots$ are negative. Let us use the spectral decomposition for B and define the positive semidefinite matrix $B^+ \in \mathbb{C}^{n\times n}$ as

$$\begin{aligned} B &= U\text{diag}[\lambda_1(B), \dots, \lambda_k(B), \lambda_{k+1}(B), \dots, \lambda_n(B)]U^*, \\ B^+ &= U\text{diag}[\lambda_1(B), \dots, \lambda_k(B), 0, \dots, 0]U^*. \end{aligned} \tag{3.16}$$

Using the fact that the eigenvalues $\lambda_{k+2}(B), \lambda_{k+3}(B), \dots$ are negative, we have that $B^+ - B \succcurlyeq 0$, i.e., $B^+ - B$ is positive semidefinite. Then, we have that $A + B^+ \succcurlyeq A + B$ and also that $A + B^+ \succcurlyeq A$. Using the variational characterization (3.15), we have $\lambda_i(A+B) \leq \lambda_i(A+B^+)$ and $\lambda_i(A) \leq \lambda_i(A+B^+)$. This leads to

$$\sum_{j=1}^k \big(\lambda_{i_j}(A+B) - \lambda_{i_j}(A)\big) \leq \sum_{j=1}^k \big(\lambda_{i_j}(A+B^+) - \lambda_{i_j}(A)\big) \leq \sum_{j=1}^n \big(\lambda_i(A+B^+) - \lambda_{i_j}(A)\big)$$

$$= \text{tr}(A+B^+) - \text{tr}(A) = \text{tr}(B^+) = \sum_{i=1}^k \lambda_i(B).$$

□

This result is more than an upper bound for the k smallest eigenvalues of $A+B$, because we can take any k eigenvalues of $A+B$, arranged in nonincreasing order and bound them from above by taking the sum eigenvalues of A and B with the same indexes. For more information about inequalities for eigenvalues of Hermitian matrices, see [Tao's Blog - 01/12/2010]. Lemma 3.24 can be used to derive an inequality about singular values as stated in the next Lemma.

Lemma 3.25. *If the matrices $X \in \mathbb{C}^{m\times n}$ and $Y \in \mathbb{C}^{m\times n}$ have singular values $\sigma_1(X) \geq \cdots \geq \sigma_l(X) \geq 0$ and $\sigma_1(Y) \geq \cdots \geq \sigma_l(Y) \geq 0$, where $l = \min\{m,n\}$ the, for any $k \in [l]$,*

$$\sum_{j=1}^{k} |\sigma_j(X) - \sigma_j(Y)| \leq \sum_{j=1}^{k} \sigma_j(X-Y).$$

Proof. Consider the self-adjoint dilatation matrices $S(X), S(Y) \in \mathbb{C}^{(m+n)\times(m+n)}$ defined by

$$S(X) = \begin{bmatrix} 0 & X \\ X^* & 0 \end{bmatrix} \quad \text{and} \quad S(Y) = \begin{bmatrix} 0 & Y \\ Y^* & 0 \end{bmatrix}.$$

Their eigenvalues obey the inequalities

$$\begin{aligned} \sigma_1(X) \geq \cdots \geq \sigma_l(X) \geq 0 = \cdots = 0 \geq -\sigma_l(X) \geq \cdots \geq -\sigma_1(X), \\ \sigma_1(Y) \geq \cdots \geq \sigma_l(Y) \geq 0 = \cdots = 0 \geq -\sigma_l(Y) \geq \cdots \geq -\sigma_1(Y). \end{aligned} \tag{3.17}$$

In order to see this, let us compute $S^2(X)$.

$$S^2(X) = S(X)S(X) = \begin{bmatrix} XX^* & 0 \\ 0 & X^*X \end{bmatrix}.$$

Let σ be any singular values of X. Then σ^2 is an eigenvalue of XX^* (and X^*X). Hence σ^2 will be an eigenvalue for $S^2(X)$ for a given eigenvector v, that is,

$$S^2(X)v = \begin{bmatrix} XX^* & 0 \\ 0 & X^*X \end{bmatrix} v = \sigma^2 v.$$

Therefore, we can finally conclude that $\pm\sigma$ is an eigenvalue for $S(X)$. Then, for $j \in [l]$, there exists a subset I_k of $[m+n]$ with size k such that

$$\sum_{j=1}^{k} |\sigma_j(X) - \sigma_j(Y)| = \sum_{j\in I_k} \big(\lambda_j(S(X)) - \lambda_j(S(Y))\big).$$

So, using Lemma 3.24, with $A = S(Y)$, $B = S(X-Y)$ and $A+B = S(X)$ leads to

$$\sum_{j=1}^{k} |\sigma_j(X) - \sigma_j(Y)| = \sum_{j=1}^{k} \lambda_j(S(X-Y)) = \sum_{j=1}^{k} \sigma_j(X-Y).$$

□

We can now finally prove that if a matrix satisfies the nullspace property it can recovered through the minimization of the nuclear norm.

Proof. (of Theorem 3.22): Assume that every matrix $X \in \mathbb{C}^{n_1\times n_2}$ of rank r is the unique solution of (P_{nuclear}) with $y = \mathcal{A}(X)$. Consider the singular value decomposition of a matrix $M \in \ker\mathcal{A}\backslash\{0\}$, which we will denote by $M = U\text{diag}(\sigma_1,\ldots,\sigma_n)V^*$ for $\sigma_1 \geq \ldots \sigma_n \geq 0$ and $U \in \mathbb{C}^{n_1\times n_1}$, $V \in \mathbb{C}^{n_2\times n_2}$ unitary matrices. So we can threshold M and cut off some singular values in order to define two new matrices

$$M_1 = U\text{diag}(\sigma_1,\ldots,\sigma_r,0,\ldots,0)V^* \quad \text{and} \quad M_2 = U\text{diag}(0,\ldots,0,-\sigma_{r+1},\ldots,-\sigma_n)V^*,$$

in such a way that $M = M_1 - M_2$. Then we translate the condition $\mathcal{A}(M) = 0$ into $\mathcal{A}(M_1) = \mathcal{A}(M_2)$. Since the rank of M_1 is at most r, its nuclear norm must be smaller than the nuclear norm of M_2. This is the same as saying that

$$||M_1||_* = \sigma_1(M) + \cdots + \sigma_r(M) < ||M_2||_* = \sigma_{r+1}(M) + \cdots + \sigma_n(M).$$

In order to prove the converse, let us suppose that $\sum_{j=1}^{r} \sigma_j(M) < \sum_{j=r+1}^{n} \sigma_j(M)$ for every $M \in \ker \mathcal{A} \backslash \{0\}$ with singular values $\sigma_1(M) \geq \cdots \geq \sigma_n(M) \geq 0$. Let us take a matrix $X \in \mathbb{C}^{n_1 \times n_2}$ of rank r and a matrix $Y \in \mathbb{C}^{n_1 \times n_2}$, with $X \neq Y$ and satisfying $\mathcal{A}(Y) = \mathcal{A}(X)$. We will prove that $||X||_* > ||Y||_*$. Let us define $M = X - Y \in \ker \mathcal{A} \backslash \{0\}$. Lemma 3.24, with $A = X - M$ and $B = M$, implies that

$$||Y||_* = \sum_{j=1}^{n} \sigma_j(X - M) \geq \sum_{j=1}^{n} |\sigma_j(X) - \sigma_j(M)|$$

For $j \in [r]$, we have $|\sigma_j(X) - \sigma_j(M)| \geq \sigma_j(X) - \sigma_j(M)$ and for the smallest singular values, $r+1 \leq j \leq n$, we clearly have $|\sigma_j(X) - \sigma_j(M)| = \sigma_j(M)$. So, using our hypothesis, we have

$$||Y||_* \geq \sum_{j=1}^{r} \sigma_j(X) - \sum_{j=1}^{r} \sigma_j(M) + \sum_{j=r+1}^{n} \sigma_j(M) \geq \sum_{j=1}^{r} \sigma_j(X) = ||X||_*$$

□

Recently, [Kabanava, Kueng, Rauhut & Terstiege '16] generalized Theorem 3.22 and established the stable and robust reconstruction of the low-rank matrices through the introduction of the robust null space property for the matrix case. They proved the following theorem.

Theorem 3.26. (Theorem 11 of [Kabanava, Kueng, Rauhut & Terstiege '16]): *Let $\mathcal{A} : \mathbb{C}^{n_1 \times n_2} \to \mathbb{C}^m$ be any linear measurement map, let $n = \min\{n_1, n_2\}$ and $||.||$ be any norm on $\mathbb{C}^m$. Assume that for any matrix $X \in \mathbb{C}^{n_1 \times n_2}$, our measurement vector is given by $y = \mathcal{A}(X) + e$ with $||e||_2 \leq \eta$, for some $\eta \geq 0$. Suppose that the measurement $\mathcal{A}$ satisfies the robust rank null space property of order r (with respect to the norm $||.||$) with constants $0 < \rho < 1$ and $\tau > 0$, that is*

$$\sum_{i=1}^{r} \sigma_i(M) \leq \rho \sum_{i=r+1}^{n} \sigma_i(M) + \tau ||\mathcal{A}(M)||, \qquad \forall M \in \mathbb{C}^{n_1 \times n_2}.$$

Then, for any matrices $X, Z \in \mathbb{C}^{n_1 \times n_2}$, we have

$$||X - Z||_* \leq \frac{1+\rho}{1-\rho} \left(||Z||_* - ||X||_* + 2 \sum_{i=r+1}^{n} \sigma_i(M) \right).$$

Also, if $\mathcal{A}$ satisfies the Frobenius robust rank null space property of order r, i.e., for all $M \in \ker \mathcal{A} \backslash \{0\}$,

$$\left(\sum_{i=1}^{r} \sigma_i(M)^2 \right)^{1/2} \leq \frac{\rho}{\sqrt{r}} \sum_{i=r+1}^{n} \sigma_i(M) + \tau ||\mathcal{A}(M)||,$$

then, for $X^\#$, a solution of the quadratically constrained nuclear norm minimization problem,

$$\min_{X \in \mathbb{C}^{n_1 \times n_2}} ||Z||_* \qquad \text{subject to } ||\mathcal{A}(Z) - y||_2 \leq \eta, \tag{3.18}$$

we have the following error estimate in the Frobenius norm,

$$||X - X^\#||_F \leq \frac{2(1+\rho)^2}{1-\rho} \frac{1}{\sqrt{r}} \sum_{i=r+1}^{n} \sigma_i(X) + \frac{2\tau(3+\rho)}{1-\rho} \eta.$$

As a corollary of Theorem 3.26, we obtain that if $X \in \mathbb{C}^{n_1 \times n_2}$ is a matrix of rank at most r and the measurements are noiseless, i.e., $\eta = 0$, then the Frobenius robust null space property implies that X is the unique solution of ($\mathrm{P}_{nuclear}$).

This section showed that the heuristic of replacing the (nonconvex) rank objective function by the sum of the singular values works, provided that the measurement $\mathcal{A}$ satisfies the null space property. But until now, nothing was said about *how* to minimize the nuclear norm in an efficient form. This was done by [Vandenberghe & Boyd '96] and highlighted by [Fazel, Hindi & Boyd '01], where they expressed the problem as a *Semidefinite Programming* (SDP) problem with constrains given by a linear matrix inequality. They argue that for this kind of problem, a lot of solvers are available to efficiently solve it. In particular, they proved the following theorem:

Theorem 3.27. (Lemma 1 of [Fazel, Hindi & Boyd '01]): *The problem (*NSP_{nuc}*) is equivalent to the following semidefinite programming problem:*

$$\min_{\substack{X \in \mathbb{C}^{n_1 \times n_2}, Y \in \mathbb{C}^{n_1 \times n_1} \\ Z \in \mathbb{C}^{n_2 \times n_2}}} \operatorname{tr} Y + \operatorname{tr} Z \qquad \textit{subject to } \mathcal{A}(X) = y \quad \textit{and} \quad M = \begin{bmatrix} Y & X \\ X^* & Z \end{bmatrix} \succcurlyeq 0.$$

To finish this discussion, let us point out that low-rank matrix recovery has a lot of application. One can cite camera calibration, removal of shadows and specularities from face images, reconstruction of urban structures from low-rank textures and so on. These and many other applications can be found at *Low-Rank Matrix Recovery and Completion via Convex Optimization*, a website maintained by Professor Yi Ma, from the University of Illinois. The URL is `http://perception.csl.illinois.edu/matrix-rank/references.html`.

3.7 Complexity Issues of the NSP

In this chapter, we saw the null space property as a necessary and sufficient condition for the equivalence between (P_1) and (P_0). However, in this section we will state an odd result related to the computational complexity of the null space property. More precisely, we will state a recent theorem which says that the computation of the *null space constant* is intractable from a computational point of view. In order to do this, we must remember one of the three equivalent conditions for NSP given in the beginning of Section 3.2, more precisely, the one given by Equation (3.1). This equation tells us that the null space property relative to a set S is

$$2||v_S||_1 < ||v||_1 \iff ||v_S||_1 < \frac{1}{2}||v||_1 \qquad \forall v \in \ker A \backslash \{0\}.$$

So taking S to be the index set S of s largest (in absolute value) entries of v, we can generalize the inequality above and define the *null space constant*.

Definition 3.28. For a given matrix $A \in \mathbb{C}^{m \times N}$, the null space constant α_s is defined as the smallest constant such that the NSP of order s holds with this constant, that is

$$\alpha_s = \min \alpha \quad \text{such that} \quad ||v_S||_1 \leq \alpha ||v||_1 \qquad \forall v \in \ker A \backslash \{0\}.$$

or equivalently,

$$\alpha_s = \max ||v_S||_1 \quad \text{such that } Av = 0,\ ||v||_1 = 1,\ S \subseteq [N],\ \#S \leq s.$$

So, clearly, by Equation (3.1) and Theorem 3.2, sparse recovery is ensured by (P_1) if and only if $\alpha_s < 1/2$. Therefore, one can ask how to compute α_s or to decide whether a given matrix $A \in \mathbb{C}^{m \times N}$ there exists $\alpha_s < 1$. The next theorem states that there is no polynomial time algorithm that computes the null space constant for all matrices A and all sparsity levels s.

Theorem 3.29. (Theorem 5 of [Tillmann & Pfetsch '14]): *Given a matrix $A \in \mathbb{Q}^{m\times N}$ and a positive integer s, the problem to decide whether A satisfies the NSP of order s with some constant $\alpha_s < 1$ is coNP-complete.*

It is sufficient to prove this theorem for a rational matrix because in floating-point arithmetics there is only the representation of rational numbers. Theorem 3.29 provides a justification to investigate sufficient conditions to Basis Pursuit, instead of relying only on necessary and sufficient condition as NSP. In Chapter 4 and Chapter 5 we will analyze the two most important sufficient conditions in Compressive Sensing literature: Coherence Property and Restricted Isometry Property.

Based on Section 3.3, we could ask why the nonconvex approach is not used instead of Basis Pursuit. The main point is that the null space property for the ℓ_q-norm, with $q < 1$, has also theoretical drawbacks, as the next theorem shows.

Theorem 3.30. (Theorem 1 of [Ge, Jiang & Ye '11]): *For any $0 < q < 1$, the problem (P_q) is NP-hard.*

Therefore, the ℓ_1-minimization approach, a convex strategy, remains as the most widely used approach to find sparse (or compressible) solutions of linear systems. We finish this chapter with an open theoretical problem related to NSP which was stated in the last section of the work [Tillmann & Pfetsch '14].

Open Problem: Prove that the computation of the null space constant is **strongly NP-hard**, that is, prove that cannot exist a fully polynomial time approximation scheme (FPTAS), i.e., an algorithm that solves the minimization problem within a factor of $(1+\varepsilon)$ of the optimal value in polynomial time with respect to the input size and to $1/\varepsilon$.

Chapter 4

The Coherence Property

There's no unique criterion for coherence, but you have to be sensitive to incoherence.
And as we've said, the test for incoherence is whether you're getting the results you don't want.
David Bohm (1917-1992)[1]

4.1 Introduction

Until now, we gave a panorama of some strategies to solve the ℓ_0-minimization problem, like thresholding and greedy algorithms. However, our main focus on this dissertation is Basis Pursuit working as a proxy for the initial combinatorial problem. With this in mind, our investigation relies in finding conditions in the matrix A which ensure exact or approximate reconstruction of the original sparse or compressible vector. A fundamental point in the analysis of the algorithms is whether we are able to prove that some specific matrices arising in applications satisfy the conditions that we examine in this chapter. This question is not so well understood, and will return later in this dissertation.

In Theorem 3.3 we proved that the *Null Space Property* is a necessary and sufficient condition for the solution of (P_0) via ℓ_1-minimization. At the same time, the results about the computational complexity of NSP show us that we must look in other directions to find alternative sufficient ways to solve the problem - the question of necessity will be proven much harder - even if these alternatives are not sharp.

The seminal paper [Candès, Romberg & Tao I '06] defined the *Restricted Isometry Property*[2] in 2004, which will be explored in Chapter 5, whereas [Cohen, Dahmen & DeVore '09] defined NSP in 2009. Historically, other sufficient conditions for sparse recovery appeared before RIP and NSP. Along this chapter, we follow the path taken by Donoho and collaborators to find sufficient conditions for Basis Pursuit.

We will also begin to explore the suitability of matrices that are present in Compressive Sensing and applications in order to guarantee uniqueness of the (sparse) solution. This analysis leads to sufficient conditions under the names *Spark* and *Coherence* of the matrix. We begin this investigation with a new point of view for a very general and ubiquitous principle of physics, which is also a metatheorem in mathematics: the uncertainty principle.

4.2 Uncertainty Principle

In the twenties, Quantum Mechanics changed the physical understanding of the world. One of its deepest results is the *Uncertainty Principle*[3], which dates back to [Heisenberg '27]. The papers [Kennard '27]

[1] In *David Bohm, "Thought as a System", Routledge, 1994. pp. 59.*

[2] These authors originally called it the *Uniform Uncertainty Principle*. Also, despite the publication date, the original preprint appeared in ArXiv at 2004.

[3] For the history of quantum mechanics, one can consult the 6 volumes (and more than 5000 pages) of [Mehra & Rechenberg '01], the most comprehensive work undertaken by anyone on one of the vastest and most important development in the history of science.

and [Weyl '28] quantified and proved it[4].

The signal analysis community probably came to know and to reflect on it after the fundamental work of Gabor [Gabor '46] in 1946. Then this principle evolved throughout the century and we can cite the extensions made by Landau, Pollack and Slepian [Slepian & Pollack '61], [Landau & Pollack '61] and later by Donoho and Stark [Donoho & Stark '89]. Furthermore, there are deep connections with other areas such as Partial Differential Equations [Fefferman '83] and Mathematical Analysis in general, see [Havin & Jöricke '94]. A particularly clear description of the philosophy of the Uncertainty Principle in Mathematics is given by Terence Tao in [Tao's Blog - 06/25/2010]:

"A recurring theme in mathematics is that of duality: a mathematical object X can either be described internally (or in physical space, or locally), by describing what X physically consists of (or what kind of maps exist into X), or externally (or in frequency space, or globally), by describing what X globally interacts or resonates with (or what kind of maps exist out of X). These two fundamentally opposed perspectives on the object X are often dual to each other in various ways: performing an operation on X may transform it one way in physical space, but in a dual way in frequency space, with the frequency space description often being a "inversion" of the physical space description. In several important cases, one is fortunate enough to have some sort of fundamental theorem connecting the internal and external perspectives[...]the uncertainty principle, that describes the dual relationship between physical space and frequency space. There are various concrete formalisations of this principle, most famously the Heisenberg uncertainty principle and the Hardy uncertainty principle - but in many situations, it is the heuristic formulation of the principle that is more useful and insightful than any particular rigorous theorem that attempts to capture that principle. Unfortunately, it is a bit tricky to formulate this heuristic in a succinct way that covers all the various applications of that principle; the Heisenberg inequality $\Delta x \cdot \Delta \xi \gtrsim 1$ is a good start, but it only captures a portion of what the principle tells us."

As we are interested in sparsity, we follow [Donoho & Stark '89] and [Donoho & Huo '01] in order to see what the *Uncertainty Principle* represents to us, namely, that a signal cannot be sparsely represented both in time and in frequency. Suppose we have a non-zero vector $v \in \mathbb{R}^n$ and two orthonormal basis, Ψ and Φ. So we can represent v in two ways: as a linear combination of columns of Ψ or as a linear combination of columns of Φ

$$v = \Psi x = \Phi y.$$

Let us suppose, as an important case, that Ψ is the identity matrix and Φ is the matrix of Discrete Fourier Transform. We then have the time-domain representation and the frequency-domain representation. Now, any kind of uncertainty principle for the representation in both basis at the same time must take into account some distance between Ψ and Φ, since if we have $\Psi = \Phi$, we can define v as one of the columns of Ψ and get the smallest possible cardinality (1 in both x and y).

One of the most important notions of distance between basis was defined by [Donoho & Huo '01], although it appeared in a heuristic way in [Mallat & Zhang '93].

Definition 4.1. Let Ψ and Φ be orthonormal (in the ℓ_2-norm) bases for some finite dimensional vector space V with $\dim V = n$. We define the *mutual coherence* between two bases as

$$\mu(\Psi, \Phi) = \sup\{|\langle \psi, \phi \rangle| : \psi \in \Psi, \phi \in \Phi\}.$$

Proposition 4.2. *Let Ψ and Φ be orthonormal bases for some finite dimensional vector space V with $\dim V = n$. Then the mutual coherence satisfies $\frac{1}{\sqrt{n}} \leq \mu(\Psi, \Phi) \leq 1$.*

Proof. The upper bound is just Cauchy-Schwarz inequality. For the lower bound, note that $(\Psi^T\Phi)^T(\Psi^T\Phi)$ $= \Phi^T\Psi\Psi^T\Phi = Id$, so $\Psi^T\Phi$ is also orthonormal and the sum of squares of its entries in each column is equal to 1. If all of these entries are less than $\frac{1}{\sqrt{n}}$, the sum of their squares is less than 1, which is impossible. To finish, note that this lower bound is achieved by the identity matrix and the Discrete Fourier Transform matrix, for example. □

[4] See also the contributions of von Neumann in [Székely & Rizzo '07].

The first version of an *uncertainty principle* in this context was proved in [Donoho & Stark '89] for the particular case where the representation is given by the identity matrix and the DFT matrix. The following more general version was proved in [Donoho & Huo '01].

Theorem 4.3. ([Donoho & Huo '01]): *For an arbitrary pair of orthogonal bases Ψ, Φ, with mutual coherence $\mu(\Psi, \Phi)$, and for a non-zero vector $v \in \mathbb{R}^n$ with representations x and y respectively*

$$||x||_0 + ||y||_0 \geq \left(1 + \frac{1}{\mu(\Psi, \Phi)}\right).$$

In fact we will prove an improved and stronger version, which resembles the original result in [Donoho & Stark '89].

Theorem 4.4. ([Elad & Bruckstein '02]): *For an arbitrary pair of orthogonal bases Ψ, Φ, with mutual coherence $\mu(\Psi, \Phi)$, and for a non-zero vector $v \in \mathbb{R}^n$ with representations x and y respectively,*

$$||x||_0 + ||y||_0 \geq \frac{2}{\mu(\Psi, \Phi)}.$$

Remark 19. *To see the difference between them, note that*

$$\frac{2}{\mu(\Psi, \Phi)} = \frac{1}{\mu(\Psi, \Phi)} + \frac{1}{\mu(\Psi, \Phi)} \geq 1 + \frac{1}{\mu(\Psi, \Phi)},$$

since $\mu(\Psi, \Phi) \leq 1$.

Remark 20. *Here we present two proofs of Theorem 4.4*

Proof. Assume w.l.o.g. that $||v||_2 = 1$. Since $v = \Psi x = \Phi y$, we have

$$\begin{aligned} 1 = ||v||_2^2 = \langle v, v \rangle = \langle \Psi x, \Phi y \rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} x_i y_j \langle \phi_i, \psi_j \rangle \leq \mu(\Psi, \Phi) \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i||y_j| \\ = \mu(\Psi, \Phi) ||x||_1 ||y||_1. \end{aligned} \tag{4.1}$$

Through the arithmetic-geometric mean inequality we obtain

$$||x||_1 ||y||_1 \geq \frac{1}{\mu(\Psi, \Phi)} \implies ||x||_1 + ||y||_1 \geq \frac{2}{\sqrt{\mu(\Psi, \Phi)}}, \tag{4.2}$$

which is a kind of uncertainty principle for the ℓ_1-norm. For the (P_0) problem, the problem of finding a representation for x with the greatest ℓ_1-norm satisfying $||x||_2 = 1$ and have k non-zeros (i.e. $||x||_0 = k$) leads to the optimization problem

$$\max_x ||x||_1 \quad \text{subject to} \quad ||x||_2^2 = 1 \text{ and } ||x||_0 = k. \tag{4.3}$$

Assume we obtain a solution to this problem of the form $f(k) = f(||x||_0)$. Similarly, suppose we have a solution for the analogous problem for y as $f(B) = f(||y||_0)$. From Equation (4.2), we obtain

$$\frac{1}{\mu(\Psi, \Phi)} \leq ||x||_1 ||y||_1 \leq f(||x||_0) f(||y||_0). \tag{4.4}$$

If we find the solution to the optimization problem (4.3), we can replace it in Equation (4.4), and this will give us the inequality we are looking for. Assume w.l.o.g. that the k non-zeros in x are the first entries and that all of these entries are strictly positive (since we are considering only absolute values in this problem). Using Lagrange multipliers, the ℓ_0 constraint vanishes, and we obtain

$$\mathcal{L}(x) = \sum_{i=1}^{k} x_i + \lambda \left(1 - \sum_{i=1}^{k} x_i^2\right).$$

Setting the derivatives equal to zero

$$\frac{\partial \mathcal{L}(x)}{\partial x_i} = 1 - 2\lambda x_i = 0,$$

from which we obtain the optimal value $x_i = \frac{1}{2\lambda}$. Using the ℓ_2 constraint, we have $x_i = \frac{1}{\sqrt{k}}$ and thus $g(k) = \frac{k}{\sqrt{k}} = \sqrt{k}$, the maximal ℓ_1-norm of the vector x. Using an analogous result for y, we have

$$\frac{1}{\mu(\Psi, \Phi)} \leq ||x||_1 ||y||_1 \leq f(||x||_0) f(||y||_0) = \sqrt{||x||_0 ||y||_0}.$$

With the help of the arithmetic-geometric mean inequality, we obtain

$$\frac{1}{\mu(\Psi, \Phi)} \leq \sqrt{||x||_0 ||y||_0} \leq \frac{1}{2}(||x||_0 + ||y||_0) \tag{4.5}$$

□

As [Elad '10] points out, Allan Pinkus found a simpler proof of Theorem 4.4.

Proof. Since Ψ and Φ are unitary matrices, we have that $||v||_2 = ||x||_2 = ||y||_2$. Let us denote the suport of x by I. From $v = \Psi x = \sum_{i \in I} x_i \psi_i$ we have

$$|y_j|^2 = |\langle v, \phi_j \rangle|^2 = \left| \sum_{i \in I} x_i \langle \psi_i, \phi_j \rangle \right|^2 \leq ||x||_2^2 \left| \sum_{i \in I} (\langle \psi_i, \phi_j \rangle)^2 \right| \leq ||v||_2^2 \ |I| \ \mu(\Psi, \Phi).$$

Summing the above for all $j \in J$, the support of y, we obtain

$$\sum_{j \in J} |y_j|^2 = ||v||_2^2 \leq ||v||_2^2 \ |I| \ |J| \mu(\Psi, \Phi) \implies \frac{1}{\mu(\Psi, \Phi)} \leq \sqrt{|I| \ |J|} = \sqrt{||x||_0 ||y||_0}.$$

□

Remark 21. *Regardless of their locations, this theorem gives us a lower bound on the number of nonzeros and tells us that, if the mutual coherence of two basis is small, then it cannot be that representations of the same vector in both basis are simultaneously sparse. This has a resemblance with the classical uncertainty principle of quantum mechanics.*

Example 4.5. Take $\Psi = Id$ and $\Phi = DFT$. Remember that the *Discrete Fourier Transform* is the unitary matrix $F \in \mathbb{C}^{n \times n}$ given by

$$F_{ik} = \frac{1}{\sqrt{n}} e^{2\pi i (i-1)(k-1)/n}, \qquad i, k \in [n].$$

We have $\mu(\Psi, \Phi) = \frac{1}{\sqrt{n}}$, by the definition of the Discrete Fourier Transform. It follows that a signal cannot have fewer than $2\sqrt{n}$ nonzeros entries in both time and frequency domains. This is a tight relationship. To see it, suppose that $n = N^2$ is a square and consider the signal[5]

$$\text{Ш}_k = \begin{cases} 1 & k = 1 \mod \sqrt{n}, \\ 0 & \text{otherwise.} \end{cases}$$

This signal has, of course, $\sqrt{n}$ coordinates different from zero. The same happens with its Fourier transform, as the next computation shows. By the definition of Fourier transform, we have

[5] This function, sometimes called picked fence and sometimes called *chá*, has this symbol and the second name as a homage to the 26th letter of Russian alphabet.

$$\widehat{\mathrm{III}}_j = \frac{1}{N}\sum_{\ell=1}^{N^2} \mathrm{III}_\ell e^{2\pi i(\ell-1)(j-1)/N^2} = \frac{1}{N}\sum_{k=1}^{N} e^{2\pi i(\ell-1)(j-1)/N} = \begin{cases} 1 & j = 1 \mod N, \\ 0 & \text{otherwise.} \end{cases}$$

This shows that $\widehat{\widehat{\mathrm{III}}} = \mathrm{III}$ and that $||\mathrm{III}||_0 = ||\widehat{\mathrm{III}}||_0 = N = \sqrt{n}$. This implies the sharpness of Theorem 4.4 with the right-hand side given by $2\sqrt{N}$ nonzeros if the bases are given by the identity and the DFT.

4.3 A Case Study: Two Orthogonal Basis

Our main goal in this dissertation is to understand when linear systems $Ax = b$ have sparse solutions and how to find them. The discussion in this section indicates that we can analyze a toy model first. This model will have the matrix A formed by the concatenation of two orthogonal basis placed side by side, i.e. $A = [\Psi, \Phi]$. In this case, a connection with the uncertainty principle arises. Donoho and collaborators found that the study of the solutions of this kind of linear system can be performed by just considering representation in bases Ψ, Φ simultaneously. After showing this, we pass to the general case.

In this particular representation problem, an overcomplete set of vectors (a *dictionary*) could lead to multiple representations of the same signal. This has some advantages. We can, for example, design better compression schemes for the signals we are working with or find natural sparse representations among multiple choices. Due to the idea that natural signals have parsimonious representation, in many situations it is desirable to work with multiple (but sparse) representations instead of a unique but dense representation of a signal.

Therefore, we can state the first consequence from the overcomplete representation idea and from Theorem 4.4. It is the *second uncertainty principle* for ℓ_0-norm, which says that just *one solution* of our linear system, given by a concatenation of basis, can be sufficiently sparse.

Theorem 4.6. *Let x_1 and x_2 be two different solutions of the linear system $Ax = [\Psi, \Phi]x = b$. Then*

$$||x_1||_0 + ||x_2||_0 \geq \frac{2}{\mu(\Psi, \Phi)},$$

i.e. both solutions cannot be very sparse simultaneously.

Proof. Let $v = x_1 - x_2$ be the difference between the two solutions, then $v \in \ker(A)$. We now partition v into first n entries and last n entries, respectively v_ψ and v_ϕ. Then $Av = 0$ implies $\Psi v_\psi + \Phi v_\phi = 0$, so

$$\Psi v_\psi = \Phi v_\phi = y \neq 0, \tag{4.6}$$

where $y \neq 0$ because Ψ and Φ are nonsingular. From (4.6) we see that v_ψ is the representation of v in the basis Ψ and that v_ϕ is the representation of v in the basis Φ. From Theorem 4.4, we have

$$||v||_0 = ||v_\psi||_0 + ||v_\phi||_0 \geq \frac{2}{\mu(\Psi, \Phi)}.$$

Since $v = x_1 - x_2$, this turns into

$$||x_1||_0 + ||x_2||_0 \geq ||x_1 - x_2||_0 = ||v||_0 \geq \frac{2}{\mu(\Psi, \Phi)},$$

where the last step is just triangular inequality for the ℓ_0-norm.

□

An immediate consequence of Theorem 4.6 is the following uniqueness result for linear systems. Henceforward we will denote the coherence of a concatenation of basis by $\mu(A)$ instead of $\mu(\Psi, \Phi)$.

Corollary 4.7. *Let $\overline{x}$ be one solution of $Ax = [\Psi, \Phi]x = b$. Suppose that it satisfies*

$$||\overline{x}||_0 \leq 1/\mu(A).$$

Then it is necessarily the sparsest one and any other solution y must have $||y||_0 > 1/\mu(A)$, that is, we have the uniqueness of sparsest solutions of the linear system.

The importance of Corollary 4.7 relies on the fact that, while usually in non-convex optimization problem we can only verify local optimality, for systems of the form $Ax = [\Psi, \Phi]x = b$ we have not only uniqueness but also global optimality, provided the the solution is sufficiently sparse.

4.4 Uniqueness Analysis for the General Case

The question now is how to pass from the concatenation of basis to the general case. Many important ideas related to it were already presented in the work [Rao & Gorodnistsky '97], including, in a hidden form, the next definition. However, it was just after [Donoho & Elad ' 03] that the connection between the ℓ_0-norm and the kernel of the matrix A appeared. This is given through the following definition.

Definition 4.8. The *spark*[6] of a general matrix A is the smallest number of columns from A that are linearly-dependent. In other words, the spark of A is the infimum of the set $\{||x||_0 \mid Ax = 0\}$.

The resemblance and connection between *spark* and *rank* seems obvious. The latter measures the maximal number of columns thar are linearly independent while the former measures the minimal number of columns that are linearly dependent. However it is much more difficult to compute the spark of a matrix than its rank, as it needs a combinatorial search over all possible subsets of columns of the matrix.

Example 4.9. [Tillmann & Pfetsch '14] Consider the matrix

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Clearly $\text{spark}(A) = 2$. But it is very important to point out that generally, $\text{spark}(A) \leq k$ does not guarantee the existence of a vector with k nonzeros entries in the nullspace of A. Just take $k = 3$ in this example. On the other hand, a nullspace vector with support size k does not yield $\text{spark}(A) = k$ but only $\text{spark}(A) \leq k$.

We list now some simple properties of the spark.

Proposition 4.10. *For a matrix $A \in \mathbb{C}^{m \times n}$ we have:*

I. $spark(A) \in \{1, \ldots, n\} \cup \{\infty\}$;

II. $spark(A) = \infty$ *if and only if* $rank(A) = n$;

III. $spark(A) = 1$ *if and only if A has a zero column;*

IV. If $spark(A) \neq \infty$, *then* $spark(A) \leq rank(A) + 1$.

With this new definition, we can prove a simple but general result.

Theorem 4.11. ([Rao & Gorodnistsky '97]): *If a system of linear equations $Ax = b$ has a solution which satisfies*

$$||x||_0 < spark(A)/2,$$

then this solution is necessarily the sparsest one.

[6] Fusion of the words *sparse* and *rank*.

Proof. Let y be another solution of this linear system, $Ay = b$. This implies $A(x - y) = 0$. Now, any vector $\gamma \in \ker(A)$ must satisfy $||\gamma||_0 \geq \text{spark}(A)$, since $A\gamma = 0$ is a linear combination of the columns of A and by the definition of spark we must have at least as many columns as $spark(A)$. By the triangular inequality for the ℓ_0-norm

$$||x||_0 + ||y||_0 \geq ||x - y||_0 \geq \text{spark}(A).$$

Since by hypothesis $||x||_0 < \text{spark}(A)/2$, any other solution y must have $||y||_0 \geq \text{spark}(A)/2$. □

Theorem 4.12. ([Donoho & Elad ' 03]): *Conversely, if x, the sparsest solution of the linear system $Ay = b$, is unique then*

$$||x||_0 < spark(A)/2.$$

Proof. Suppose, by contradiction, that $spark(A) \leq 2||x||_0$. This means that there exists a set of at most $2k$ columns that are linearly dependent, which in turn implies that there exists an $h \in \ker A$ such that $||h||_0 \leq 2k$. Since we have this limitation for the sparsity of h, we can write $h = x - \tilde{x}$ with $||x||_0 \leq k$ and $||\tilde{x}||_0 \leq k$ and $x \neq \tilde{x}$. Then we have $Ah = A(x - \tilde{x}) = 0$ so that $Ax = A\tilde{x}$. But this is a contradiction with the assumption that there exists a unique sparsest solution of $Ax = b$.

□

Although we have a nice uniqueness result for the general case, the *spark* actually requires calculations with all possible subsets of columns of size up to $spark(A) + 1$. This heuristic indicates that the its computational complexity can make the use of *spark* an inadequate practical choice for finding sparsest solutions of linear systems. Therefore we must try to find another, possibly non-sharp, guarantee of uniqueness.

However, it is important to point out that in some situations, using random matrices, it can be easy to estimate the *spark*. If we consider a matrix $A \in \mathbb{R}^{m \times n}$ with $m \leq n$ and entries given by i.i.d. random variables (random matrices will be explored in Chapter 7), then with probability one, any submatrix of size $m \times m$ has maximal rank m and hence $\text{spark}(A) = m+1$ with probability 1. See [Feng & Zhang '07].

The question about computational tractability of the *spark* was addressed by [Tillmann & Pfetsch '14]. They confirm the heuristics mentioned above. This work considers the *spark* as a particular case of a *circuit* and proves the NP-completeness for the decision problem of the existence of a circuit in a matrix by reducing it to the *k-Clique Problem*. See Section 3.1.3 of [Garey & Johnson '79] for the *k-Clique Problem*.

Definition 4.13. For a given matrix $A \in \mathbb{R}^{m \times n}$, a *circuit* is a set $C \subseteq \{1, \ldots, n\}$ of column indices such that $A_C x = 0$ has a nonzero solution but every proper subset of C does not have this property, i.e., $\text{rank}(A_C) = |C| - 1 = \text{rank}(A_{C \setminus \{j\}})$, $\forall j \in C$. The *spark* of A is the size of its smallest circuit.

The most important theorem from [Tillmann & Pfetsch '14] is the following.

Theorem 4.14. *Given a matrix $A \in \mathbb{Q}^{m \times n}$ and a positive integer k, the problem of deciding whether there exists a circuit of A of size at most k is NP-complete.*

To end this section, just note that there exists a circuit of size at most k if and only if the *spark* is at most k. This proves that computing $spark(A)$ is NP-hard.

4.5 Coherence for General Matrices

A clever way of overcoming the difficulty in computing the *spark* was proposed by [Donoho & Elad ' 03]. They look at the mutual coherence, defined in Section 4.2 for the two-orthogonal case, in a different way. This definition can be seen as the maximal off-diagonal entry (in absolute value) of the Gram matrix:

$$A^T A = \begin{bmatrix} \text{Id} & \Psi^T \Phi \\ \Phi^T \Psi & \text{Id} \end{bmatrix}.$$

Inspired by it, we are able to define the concept of coherence for a general matrix. We stress, again, that, without loss of generality, the columns of the matrix are always implicitly understood to be ℓ_2 normalized.

Definition 4.15. Let $A \in \mathbb{C}^{m\times N}$ be a matrix with ℓ_2-normalized columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$ for all $i \in [N]$. The *coherence* $\mu = \mu(A)$ of a matrix A is defined as

$$\mu(A) := \max_{1\leq i\neq j\leq N} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle|. \tag{4.7}$$

Remark 22. *Many authors call the quantity in the previous definition* mutual coherence. *Here we will emphasize that* coherence *refers to a matrix A and* mutual coherence *to a pair of basis, like the two-orthogonal case.*

The coherence is easier to compute than the spark. Also, it provides a lower bound for the spark, which, as we pointed out above, is in general hard to compute. In order to prove this statement, we first state and prove a well-known theorem in Linear Algebra. This result will reappear later when discussing the relation between the number of measurements and the ambient space dimension of the signals of interest.

Theorem 4.16. ([Gershgorin '31])*: Let λ be an eigenvalue of a matrix $A \in \mathbb{C}^{n\times n}$. There exists an index $j \in [n]$ such that*

$$|\lambda - A_{j,j}| \leq \sum_{l\in[n]\setminus\{j\}} |A_{j,l}|.$$

Proof. Let $\mathbf{v} \in \mathbb{C}^n\setminus\{0\}$ be an eigenvector associated with λ, and let $j \in [n]$ such that $|v_j|$ is maximal, i.e., $|v_j| = ||v||_\infty$. Then $\sum_{l\in[n]} A_{j,l}v_l = \lambda v_j$, and separating the jth coordinate, we have $\sum_{l\in[n]\setminus\{j\}} A_{j,l}v_l = \lambda v_j - A_{j,j}v_j$. By the triangle inequality

$$|\lambda - A_{j,j}||v_j| \leq \sum_{l\in[n]\setminus\{j\}} |A_{j,l}||v_l| \leq ||v||_\infty \sum_{l\in[n]\setminus\{j\}} |A_{j,l}| = |v_j| \sum_{l\in[n]\setminus\{j\}} |A_{j,l}|.$$

Dividing by $|v_j| > 0$ yields the desired statement.

□

As a result of Theorem 4.16, every eigenvalue of a matrix lies within at least one of the Gershgorin discs, which are the discs centered at one of the diagonal entries with radius $\sum_{l\in[n]\setminus\{j\}} |A_{j,l}|$. It has many applications, such as matrix preconditioning. When we try to solve $Ax = b$ for A with a large condition number, we instead precondition and solve $PAx = Pb$ where $P \approx A^{-1}$. Since $PA \approx \mathrm{Id}$, the eigenvalues of PA should all be close to 1 and we can use Theorem 4.16 to estimate how good is our choice of P. For more details on Gershgorin's Theorem and its applications, see [Varga '04]. Now we can state the relation between $spark(A)$ and the coherence.

Lemma 4.17. (Theorem 5 of [Donoho & Elad ' 03]): *For any matrix $A \in \mathbb{R}^{n\times m}$,*

$$spark(A) \geq 1 + \frac{1}{\mu(A)}.$$

Proof. The Spark and the coherence of a matrix do not change when we ℓ_2-normalize its columns, therefore without loss of generality, we will consider a column normalized matrix. Let $\mathrm{spark}(A) = k$ be the smallest number of dependent columns of A and let K be the set of such columns. Consider the matrix A_K given by the restriction of A to its columns in K. The Gram matrix $G = A_K^T A_K$ is singular, since the original A_k is singular too. Therefore, the spectrum of G contains 0. Applying Theorem 4.16 to the eigenvalue $\lambda = 0$, we obtain

$$|\lambda - g_{ii}| = |1 - 0| \leq \sum_{i\neq j} |g_{ij}| = \sum_{i\neq j} \langle a_i, a_j\rangle \leq (k-1)\mu(A) = (\mathrm{spark}(A) - 1)\mu(A).$$

□

Now we have another uniqueness corollary, this time based on the coherence.

Theorem 4.18. ([Donoho & Huo '01]): *If a system of linear equations $Ax = b$ has a solution which satisfies*

$$||x||_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right),$$

this solution is necessarily the sparsest one.

The difference between Theorem 4.11 and Theorem 4.18 is that while the first one is sharp, the second one just gives a lower bound. Coherence can never be smaller than $1/\sqrt{n}$ and therefore the cardinality bound in the second case in never larger than $\sqrt{n}/2$. However, the spark can be as large as n, so the first theorem gives a bound for the cardinality as large as $n/2$.

Theorem 4.18 provides a sufficient condition for linear systems to have a unique solution. Albeit it is not sharp, it is easier to be estimated. See Section 2.3 of [Elad '10]. Now, it is interesting to explore this property and find out what it can tell us about the solution of linear systems.

4.6 Properties and Generalizations of Coherence

The definition of coherence is a way to represent the dependence between columns of the matrix A. It was generalized in many different ways in order to characterize sharper degrees of dependence. As [Kutyniok '12] pointed out, it is interesting to note that different variations of coherence have appeared in the literature, for example *structured p-Babel function* [Borup, Gribonval & Nielsen '08], *Babel function* [Donoho & Elad ' 03], *cluster coherence* [Donoho & Kutyniok '13], *cumulative coherence function* [Tropp '04], and *fusion coherence* [Boufounos, Kutyniok & Rauhut '11].

Following Tropp, we will introduce a generalization of *coherence* which he called *cumulative coherence function*. However, as in [Rauhut & Foucart '13], we call it *ℓ_1-coherence function*.

Definition 4.19. Let $A \in \mathbb{C}^{m\times N}$ be a matrix with ℓ_2-normalized columns $\mathbf{a}_1, \ldots, \mathbf{a}_N$. The *$\ell_1$-coherence function* μ_1 of a matrix A is defined for $s \in [N-1]$ by

$$\mu_1(s) := \max_{i\in[N]} \max \left\{ \sum_{j\in S} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle|,\ S \subset [N],\ \mathrm{card}(S) = s,\ i \notin S \right\}. \tag{4.8}$$

Remark 23. *Note that for $s = 1$, the ℓ_1-coherence function is the coherence of a matrix.*

Remark 24. *The ℓ_1-coherence function can be generalized in a straightforward way for $p > 0$, to the* ℓ_p-coherence function*, that is*

$$\mu_p(s) := \max_{i\in[N]} \max \left\{ \left(\sum_{j\in S} |\langle \boldsymbol{a}_i, \boldsymbol{a}_j\rangle|^p \right)^{1/p},\ S \subset [N],\ \mathrm{card}(S) = s,\ i \notin S \right\}.$$

Now we will state the first properties that follow from the definition of $\mu(A)$.

Proposition 4.20. *For $1 \le s \le N-1$ we have*

$$\mu \le \mu_1(s) \le s\mu. \tag{4.9}$$

More generally, for $1 \le s, t \le N-1$ with $s + t \le N-1$, we have

$$\max\{\mu_1(s), \mu_1(t)\} \le \mu_1(s+t) \le \mu_1(s) + \mu_1(t). \tag{4.10}$$

A very important property of matrices with small coherence is that column submatrices of moderate size are well conditioned. This is the content of Theorem 4.21, which is the reason behind the motto *"matrices with small coherence are great for sensing"*. Let us recall that A_S denotes the matrix formed by the columns of $A \in \mathbb{C}^{m\times N}$ indexed by a subset S of $[N]$.

Theorem 4.21. *Let $A \in \mathbb{C}^{m\times N}$ be a matrix with ℓ_2-normalized columns and let $s \in [N]$. For all s-sparse vectors $x \in \mathbb{C}^N$*

$$(1-\mu_1(s-1))||x||_2^2 \leq ||Ax||_2^2 \leq (1+\mu_1(s-1))||x||_2^2,$$

or equivalently, for each set $S \subset [N]$ with card$(S) \leq s$, the eigenvalues of the matrix $A_S^*A_S$ lie in the interval $[1-\mu_1(s-1), 1+\mu_1(s-1)]$. In particular, if $\mu_1(s-1) < 1$, then $A_S^*A_S$ is invertible.

Proof. Note that for a fixed set $S \subset [N]$ with card$(S) \leq s$, the matrix $A_S^*A_S$ is positive semidefinite, so it has an orthonormal basis of eigenvectors associated with real positive eigenvalues. Note also that $Ax = A_Sx_S$ for any $x \in \mathbb{C}^n$ supported on S. Due to normalization, i.e. $||a_j||_2 = 1$ for all $j \in [N]$, the diagonal entries of A^*A are all equal to one. By Gershgorin's theorem, the eigenvalues of A^*A are contained in the union of the disks centered at 1 with radii

$$r_j = \sum_{l\in S, l\neq j} |(A^*A)_{j,l}| = \sum_{l\in S, l\neq j} |\langle a_l, a_j\rangle| \leq \mu_1(s-1) \qquad j \in S.$$

Since the eigenvalues are real, they must lie in $[1-\mu_1(s-1), 1+\mu_1(s-1)]$. This is equivalent to the first inequality in the Theorem's statement, which follows from the fact that the Rayleigh-Ritz quotient

$$R_{A^*A}(x) = \frac{\langle x, A^*Ax\rangle}{\langle x, x\rangle}$$

attain its maximum and minimum $\lambda_{\max}$ and $\lambda_{\min}$ at their associated eigenvectors, respectively $x = v_{\max}$ and $y = v_{\min}$.

□

Corollary 4.22. *Given a matrix $A \in \mathbb{C}^{m\times N}$ with ℓ_2-normalized columns and an integer $s \geq 1$, if*

$$\mu_1(s) + \mu_1(s-1) < 1$$

then, for each set $S \subset [N]$ with card$(S) \leq 2s$, the matrix $A_S^*A_S$ is invertible and the matrix A_S is injective. In particular, the conclusion holds if

$$\mu < \frac{1}{2s-1}. \tag{4.11}$$

Proof. Using Equation (4.10), the condition $\mu_1(s) + \mu_1(s-1) < 1$ implies that $\mu_1(2s-1) < 1$. For a set $S \subset [N]$ with card$(S) \leq 2s$, from the previous Theorem we deduce that the smallest eigenvalue of the matrix A^*A satisfies $\lambda_{min} \geq 1 - \mu_1(2s-1) > 0$ which shows that A^*A is invertible. To prove that A is injective, observe that $A_Sz = 0$ leads to $A^*Az = 0$ and so $z = 0$. Now, the second statement follows from (4.9), because $\mu_1(s) + \mu_1(s-1) \leq s\mu + (s-1)\mu = (2s-1)\mu < 1$ if $\mu < 1/(2s-1)$.

□

Corollary 4.22 is a reformulation of Theorem 4.18. This theorem tells us what kind of vectors we can retrieve from the solution of the linear system, that is, with what degree of sparsity we can deal with. Clearly, with any reasonable method, we expect to recover not only 1-sparse or 2-sparse vectors but maybe 10000-sparse vectors. From Equation 4.11, we have that the coherence must be as small as possible. Therefore it is important to know what properties a matrix with small coherence has or, at least, how to construct them.

4.7 Constructing Matrices with Small Coherence

Matrices with small coherence are important because they are *well behaved*, in the sense that $A_S^* A_S$ is invertible. Now we address this question of how to construct this type of matrices $A \in \mathbb{K}^{m\times N}$ with $m \ll N$. We first note that if $m = N$, $\mu(A) = 0$ for an orthogonal matrix A. In the $m \ll N$ framework, we expect to have some lower bound. This is the content of the following theorem.

Theorem 4.23. ([Welch '74]): *The coherence of a matrix $A \in \mathbb{K}^{m\times N}$ with ℓ_2-normalized colums satisfies*

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}}. \tag{4.12}$$

Equality holds if and only if the columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ of the matrix A form an equiangular tight frame.

Remark 25. *Frames will be introduced in the next section. We postpone until then to understand what it means the equality in Theorem 4.23 to hold. Anyway, we exhibit the proof here.*

Proof. Consider the Gram matrix $G = A^*A \in \mathbb{K}^{N\times N}$, with entries given by

$$G_{i,j} = \overline{\langle \mathbf{a}_i, \mathbf{a}_j \rangle} = \langle \mathbf{a}_j, \mathbf{a}_i \rangle, \qquad i, j \in [N]$$

and the matrix $H = AA^* \in \mathbb{K}^{N\times N}$. As the columns $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ are ℓ_2-normalized, we have

$$\mathrm{tr}(G) = \sum_{i=1}^{N} ||\mathbf{a}_i||_2^2 = N. \tag{4.13}$$

From the Cauchy-Schwarz inequality we obtain

$$\mathrm{tr}(H) = \langle H, \mathrm{Id}_m \rangle_F \leq ||H||_F ||\mathrm{Id}_m||_F = \sqrt{m}\sqrt{\mathrm{tr}(HH^*)}. \tag{4.14}$$

Now observe that

$$\begin{aligned} &\mathrm{tr}(HH^*) = \mathrm{tr}(A^*AA^*A) = \mathrm{tr}(GG^*) = \sum_{i,j=1}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 = \\ &\sum_{i=1}^{N} ||\mathbf{a}_i||_2^2 + \sum_{i,j=1, i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 = N + \sum_{i,j=1, i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2. \end{aligned} \tag{4.15}$$

Combining the three equations above with the fact that $\mathrm{tr}(H) = \mathrm{tr}(G)$, yields

$$N^2 \leq m \left(N + \sum_{i,j=1, i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|^2 \right). \tag{4.16}$$

Since

$$|\langle \mathbf{a}_i, \mathbf{a}_j \rangle| \leq \mu \quad \text{for all } i, j \in [N], i \neq j \tag{4.17}$$

we obtain $N^2 \leq m(N + (N^2 - N)\mu^2)$ which leads to $\mu \geq \sqrt{\frac{N-m}{m(N-1)}}$.

Equality holds in Equation (4.12) exactly when it holds in Equations (4.14) and (4.17). Equality in (4.14) implies that $H = \lambda \mathrm{Id}_m$ for some nonnegative constant λ. In the next section we will see that this is exactly the case when the system $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ is a tight frame and that equality in (4.17) means that this system is equiangular.

□

We can extend the Welch bound to the ℓ_1-coherence function for small values of its argument.

Theorem 4.24. *The ℓ_1-coherence of a matrix $A \in \mathbb{K}^{m\times N}$ with ℓ_2-normalized columns satisfies*

$$\mu_1(s) \geq s\sqrt{\frac{N-m}{m(N-1)}} \quad \text{whenever} \quad s < \sqrt{N-1}. \tag{4.18}$$

Equality holds if and only if the colums $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ of the matrix A form an equiangular tight frame.

In order to prove this theorem, we will need a lemma. The reader may note the similarity in the proof of this lemma with the proof of Theorem 1.7 via optimization techniques.

Lemma 4.25. *For $k < \sqrt{n}$, if the finite sequence $(\alpha_1, \alpha_2, \ldots, \alpha_n)$ satisfies*

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_n \geq 0 \quad \text{and} \quad \alpha_1^2 + \alpha_2^2 + \cdots + \alpha_n^2 \geq \frac{n}{k^2}$$

then

$$\alpha_1 + \alpha_2 + \cdots + \alpha_k \geq 1$$

with equality if and only if $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 1/k$.

Proof. The first thing to note is that the lemma is equivalent to the statement

$$\begin{cases} \alpha_1 \geq \alpha_2 \geq \ldots \alpha_n \geq 0 \\ \alpha_1 + \alpha_2 + \ldots \alpha_k \leq 1 \end{cases}$$

which implies $\alpha_1^2 + \alpha_2^2 + \ldots\ \alpha_n^2 \leq \frac{n}{k^2}$, with equality if and only if $\alpha_1 = \alpha_2 = \cdots = \alpha_n = \frac{1}{k}$. To prove this, we can look at the problem of maximizing the convex function $f(\alpha_1, \alpha_2, \ldots, \alpha_n) := \alpha_1^2 + \alpha_2^2 + \ldots\ \alpha_n^2$ over the convex polygon

$$\mathcal{C} := \{(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}^n : \alpha_1 \geq \alpha_2 \geq \ldots \alpha_n \geq 0 \quad \text{and} \quad \alpha_1 + \alpha_2 + \ldots \alpha_k \leq 1\}.$$

As the function f is convex and $\mathcal{C}$ is a convex combination of its vertices, it follows that the the maximum is attained at a vertex of $\mathcal{C}$ (see Theorem B.16 from [Rauhut & Foucart '13] or Theorem 2.65 from [Ruszczynski]). The vertices of $\mathcal{C}$ are obtained as intersections of n hyperplanes arising when we turn n of the $n+1$ inequality constrains into equalities. Thus, these are the possibilities we have:

1. If $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$, then $f(\alpha_1, \alpha_2, \ldots, \alpha_n) = 0$,
2. If $\alpha_1 + \alpha_2 + \cdots + \alpha_k = 1$ and $\alpha_1 = \cdots = \alpha_l > \alpha_{l+1} = \cdots = \alpha_n = 0$ for $1 \leq l \leq k$ then $\alpha_1 = \alpha_2 = \cdots = \alpha_l = 1/l$, and consequently $f(\alpha_1, \alpha_2, \ldots, \alpha_n) = 1/l$,
3. If $\alpha_1 + \alpha_2 + \cdots + \alpha_k = 1$ and $\alpha_1 = \cdots = \alpha_l > \alpha_{l+1} = \cdots = \alpha_n = 0$ for $k \leq l \leq n$ then $\alpha_1 = \alpha_2 = \cdots = \alpha_l = 1/k$, and consequently $f(\alpha_1, \alpha_2, \ldots, \alpha_n) = l/k^2$.

Given that $k < \sqrt{n}$, it follows that

$$\max_{(\alpha_1,\ldots,\alpha_n)\in\mathcal{C}} f(\alpha_1, \alpha_2, \ldots, \alpha_n) = \max\left\{\max_{1\leq l\leq k} \frac{1}{l}, \max_{k\leq l\leq n} \frac{l}{k^2}\right\} = \max\left\{1, \frac{n}{k^2}\right\} = \frac{n}{k^2}$$

with equality only in the case $l = n$ where $\alpha_1 = \alpha_2 = \cdots = \alpha_n = 1/k$

□

Proof. (of Theorem 4.24) From equation (4.16), we have

$$\sum_{i,j=1,i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle|^2 \geq \frac{N^2}{m} - N = \frac{N(N-m)}{m},$$

which yields

$$\max_{i\in[N]} \sum_{j=1,i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle|^2 \geq \frac{1}{N} \sum_{i,j=1,i\neq j}^{N} |\langle \mathbf{a}_i, \mathbf{a}_j\rangle|^2 \geq \frac{N-m}{m}.$$

Let $i^* \in [N]$ be the index at which the maximum is achieved. If we reorder our sequence $\{|\langle \mathbf{a}_i^*, \mathbf{a}_j\rangle|\}_{j=1,j\neq i^*}^{N}$ in such way that $|\langle \mathbf{a}_i^*, \mathbf{a}_1\rangle| \geq |\langle \mathbf{a}_i^*, \mathbf{a}_2\rangle| \geq \cdots \geq |\langle \mathbf{a}_i^*, \mathbf{a}_{N-1}\rangle| \geq 0$, we have

$$|\langle \mathbf{a}_i^*, \mathbf{a}_1\rangle|^2 + |\langle \mathbf{a}_i^*, \mathbf{a}_2\rangle|^2 + \cdots + |\langle \mathbf{a}_i^*, \mathbf{a}_{N-1}\rangle|^2 \geq \frac{N-m}{m}.$$

Lemma 4.25 with $n = N-1$, $k = s$ and $\alpha_l := (\sqrt{m(N-1)/(N-m)/s})|\langle \mathbf{a}_i^*, \mathbf{a}_l\rangle|$ leads to $\alpha_1 + \cdots + \alpha_s \geq 1$. It then follows that

$$\mu_1(s) \geq |\langle \mathbf{a}_i^*, \mathbf{a}_1\rangle| + |\langle \mathbf{a}_i^*, \mathbf{a}_2\rangle| + \cdots + |\langle \mathbf{a}_i^*, \mathbf{a}_s\rangle| \geq s\sqrt{\frac{N-m}{m(N-1)}}.$$

Let us assume now that equality holds in (4.18), which implies that all previous inequalities are in fact equalities. As in the proof of the Welch bound for coherence, equality in (4.16) implies that the system $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ is a tight frame. Besides that, the case of equality in Lemma 4.25 implies that $|\langle \mathbf{a}_i^*, \mathbf{a}_j\rangle| = (\sqrt{(N-1)/m(N-m)})$ for all $j \in [N]$, $j \neq i^*$. Since the index i^* can be arbitrarily chosen from $[N]$, the system $(\mathbf{a}_1, \ldots, \mathbf{a}_N)$ is also equiangular. Conversely, the proof that equiangular tight frames yield equality in (4.18) follows easily from Theorem 4.23 and (4.9).

□

The content of Theorem 4.23 tells us that *"equality holds if and only if the columns $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ of the matrix A form an equiangular tight frame."* Then, in order to achieve the lower bound for coherence and construct good sensing matrices, we need to understand what is an equiangular tight frame.

4.8 A Glimpse of Frame Theory

4.8.1 History and Motivation

Along this dissertation we emphasized the important question of what are the building blocks of our signals of interest and how to represent them in a sparse way. In this context the concept of basis can be a bit restrictive because of the requirements of linear independence and uniqueness of coefficients in the decomposition of a vector. If we relax some of these conditions we can introduce the idea of *frames*, which are intuitively "basis with extra elements".

Fourier Transform is a major tool in Analysis, with connections to many other areas of Mathematics. However, it has some shortcomings when used in signal analysis. For example, from Fourier Analysis alone we cannot know the moment of emission and duration of a signal, because this information is hidden in the phase. Consider for example an orchestra with a very high-pitch piccolo and a low-pitched tuba. The Fourier Transform gives us that the piccolo and the tuba are present in the music but it cannot give us the information about when one of the instruments starts to play or the other one ceases.

The situation about this kind of analysis started changing in 1946, when Dennis Gabor formulated a fundamental approach to signal decomposition in terms of some elementary signals, introducing the concept of localization in the frequency space that quickly became a paradigm for spectral analysis and is, nowadays, central for many areas such as *image and audio processing*. See [Cohen '94].

Frames were introduced in 1952 by [Duffin & Schaeffer '52], who used Gabor's ideas in the study of nonharmonic Fourier series. In their work, they used highly overcomplete families of exponential functions (in their terminology, a "Hilbert space frame") and found a very efficient way to compute the coefficients in these families. Their work is considered seminal in the field of *Time-Frequency Analysis*. See [Gröchenig '01] for a book with mathematical orientation and [Cohen '94] for a book with a more applied approach. Early on, their ideas did not achieve general interest outside of nonharmonic Fourier series

studies (one might look at [Young '80]). After the initial work of Gabor, Frame Theory became popular in the area of Signal Processing just in 1985, due to the work [Daubechies, Grossman & Meyer '85].

The importance of Frames relies on the fact that they provide a great approach to deal with redundant, yet stable, representations of data dealing with noise, erasures and quantization effects. Frames still have the property that they can be constructed to fit a particular problem in a way not possible by a set of linearly independent vectors, so today frames provide an extensive framework for the analysis and decomposition accompanied by various reconstruction procedures of signals. Besides the great scope of applications that go beyond signal processing, data compression and sampling theory we can also cite optics, filterbanks, signal detection just to name a few subjects. Nowadays, it is also a highly prolific area for mathematicians, with connections with Besov Spaces, Geometry of Banach Spaces and Random Matrices.

In this Section we follow [Casazza '00] closely, which is a great introduction to the history and the theory of frames. Other good references on this area are [Christensen '08] and [Casazza & Kutyniok '13].

4.8.2 An Example

The following example, which gives a clear picture of the differences between basis and frames, is taken from [Han et al. '07]. Consider two finite sequences of vectors in the plane $\mathbb{R}^2$:

$$A = \{(1,0),(0,1)\} \quad \text{and} \quad B = \left\{\sqrt{\frac{2}{3}}(1,0), \sqrt{\frac{2}{3}}\Big(-\frac{1}{2},\frac{\sqrt{3}}{2}\Big), \sqrt{\frac{2}{3}}\Big(-\frac{1}{2},-\frac{\sqrt{3}}{2}\Big)\right\}.$$

Let us compare the two sequences:

1. Both A and B are spanning sets for $\mathbb{R}^2$.
2. The vectors in A are linearly independent, so the coefficients in the linear expansion are unique, while the vectors in B are not linearly independent, so a linear expansion is not unique;
3. Vectors in A are normalized. Vectors in B all have length $\sqrt{\frac{2}{3}}$;
4. Vectors in A are orthogonal. Vectors in B are not orthogonal;
5. Both A and B satisfy Parseval's identity for orthonormal bases:

$$||x||^2 = \sum_{i=1}^{2} c_i^2 = \sum_{i=1}^{3} d_i^2$$

   where c_i and d_i are the coefficients in the expansion in the sequence A and B respectively;
6. The coefficients for a linear expansion of some vector x in the sequence A can be computed easily using the dot product, as A is an orthonormal basis. One way to write x as a linear combination of the vectors in B is to use the coefficients formed by taking dot products also, so even without being orthogonal or linearly independent, the set B maintains one of the extremely useful features of an orthonormal basis.

We will later see that both sequences A and B are examples of a particular type of frame, called a *Parseval frame*. Sequence B has many of the properties of an orthonormal basis. This is exactly the motivation behind the study of frames.

In some situations, the properties of uniqueness of coefficients and orthogonality of vectors are not necessary. If you are sending your side of a phone conversation or a photo, what matters is quickly computing a working set of expansion coefficients, not whether those coefficients are unique. What if one of the coefficients representing a vector gets lost in transmission? That piece of information cannot be reconstructed. It is lost. Perhaps we would like our system to have some redundancy, such that if one piece gets lost, the information can be pieced together from what does get through.

4.8.3 Definition and Basic Facts

Definition 4.26. Let V be a vector space. A countable family of elements $\{a_k\}_{k\in I}$ in V is a *frame* for V if there exist constants $A, B > 0$ such that

$$A||x||^2 \leq \sum_{k\in I} |\langle x, a_k\rangle|^2 \leq B||x||^2 \quad \forall x \in V.$$

The numbers A, B are called *frame bounds*. The *optimal lower frame bound* is the supremum over all lower frame bounds, and the *optimal upper frame bound* is the infimum over all upper frame bounds. When $A = B$ we will say that the frame is a *tight frame*. When $A = B = 1$, we call it a *Parseval frame*. We also say that we have a *Uniform Frame* when all vectors in it have equal norm.

A practical way to characterize frames in a equivalent manner is given by the following proposition.

Proposition 4.27. *For a system of vectors $a_1, \ldots, a_N$ in $\mathbb{K}^m$ that is a tight frame, the following three statements are equivalent:*

i.) $||x||_2^2 = A\sum_{j=1}^N |\langle x, a_j\rangle|^2, \qquad \forall x \in \mathbb{K}^m;$

ii.) $x = A\sum_{j=1}^N \langle x, a_j\rangle a_j, \qquad \forall x \in \mathbb{K}^m;$

iii.) $\Phi\Phi^* = \frac{1}{A} Id_m$ *where Φ is the matrix with columns $a_1, \ldots, a_N$.*

Let $\{a_k\}_{k=1}^N$ be a sequence in V. By Cauchy-Schwarz' inequality we have that

$$\sum_{k=1}^N |\langle x, a_k\rangle|^2 \leq \sum_{k=1}^N ||a_k||^2||x||^2 \quad \forall x \in V,$$

so choosing $B = \sum_{k=1}^N ||a_k||^2$ we prove that this sequence is a frame. The quest is to find a smaller upper bound than this one. Now, in order to find lower bounds, it is necessary that $\text{span}\{a_k\}_{k=1}^N = V$. This condition is also sufficient.

Proposition 4.28. *Let $F = \{a_k\}_{k=1}^N$ be a sequence in V. Then F is a frame for V if and only if span$F = V$.*

Proof. We can assume that not all a_k are zero. We know that $B = \sum_{k=1}^N ||a_k||^2$ is an upper bound. Now consider the continuos mapping

$$\phi : V \to \mathbb{R}, \quad \phi(x) = \sum_{k=1}^N |\langle x, a_k\rangle|^2.$$

The unit ball in V is compact, so we can find $y \in V$ with $||y|| = 1$ such that

$$A := \sum_{k=1}^N |\langle y, a_k\rangle|^2 = \inf\left\{\sum_{k=1}^N |\langle x, a_k\rangle|^2 : x \in V, ||x|| = 1\right\}.$$

Clearly $A > 0$, because if $A = 0$, y would be orthogonal to all a_k, thus contradicting the fact that $V = \text{span}\{a_k\}_{k=1}^N$. Now given $x \in V$ with $x \neq 0$, we have

$$\sum_{k=1}^N |\langle x, a_k\rangle|^2 = \sum_{k=1}^N |\langle \frac{x}{||x||}, a_k\rangle|^2||x||^2 \geq A||x||^2,$$

therefore F is a frame. We will now prove the converse. Assume that $\{a_k\}_{k=1}^N$ does not span V. Then, there exists a vector x in the orthogonal complement $W^\perp$ of the subspace $W = \text{span}\{a_k\}_{k=1}^N$. Since x is orthogonal to each a_i, the sum $\sum_{i=1}^k |\langle x, a_i\rangle|^2 = 0$, the lower frame bound would not exist and this collection would not be a frame. □

Remark 26. *From Proposition 4.28, we see that a frame might contain more elements than needed to be a basis. In particular, if $\{a_k\}_{k=1}^N$ is a frame for V and $\{g_k\}_{k=1}^N$ is an arbitrary finite collection of vectors in V, then $\{a_k\}_{k=1}^N \cup \{g_k\}_{k=1}^N$ is also a frame for V. A frame which is not a basis is said to be* ***overcomplete*** *or* ***redundant****.*

We want to use frames to reconstruct vectors with redundance. Proposition 4.27 tells us how to do this with *tight frames*. For general frames, we need the concept of *dual frames*, given by the following proposition:

Proposition 4.29. *Let $\{a_k\}_{k=1}^N$ be a frame for V. Then there exists a frame $\{b_k\}_{k=1}^N$ such that every $x \in V$ can be reconstructed through the formula*

$$x = \sum_{i=1}^{k} \langle x, b_i \rangle a_i = \sum_{i=1}^{k} \langle x, a_i \rangle b_i.$$

Remark 27. *Such frame $\{b_k\}_{k=1}^N$ is called a* dual frame.

Proof. Let $T : V \to \mathbb{C}^k$ be defined by

$$Tx = (\langle x, a_1 \rangle, \langle x, a_2 \rangle, \ldots, \langle x, a_n \rangle).$$

Note that T is linear and also injective, because $Tx = 0$ implies $\langle x, a_i \rangle = 0 \quad \forall i \in \{1, \ldots, k\}$. Since $\{a_k\}_{k=1}^N$ spans V, we conclude that $x = 0$. We also conclude that the operator given by $S = T^*T : V \to V$ is invertible. The operator S is called the *frame operator* for $\{a_k\}_{k=1}^N$. Now, for $\{e_i\}_{i=1}^k$ the canonical basis of $\mathbb{C}^k$, we have $Tx = \sum_{i=1}^k \langle x, a_i \rangle e_i$, and for each $x \in V$,

$$\langle x, T^* e_j \rangle = \langle Tx, e_j \rangle = \langle \sum_{i=1}^{k} \langle x, a_i \rangle e_i, e_j \rangle = \langle x, a_j \rangle,$$

thus $T^* e_j = a_j$. This implies, for every $x \in V$

$$Sx = T^*Tx = T^* \left(\sum_{i=1}^{k} \langle x, a_i \rangle e_i \right) = \sum_{i=1}^{k} \langle x, a_i \rangle T^* e_i = \sum_{i=1}^{k} \langle x, a_i \rangle a_i.$$

Define $b_i = S^{-1} a_i$ for $i = 1, \ldots k$. Then, we have

$$x = S^{-1}Sx = S^{-1} \sum_{i=1}^{k} \langle x, a_i \rangle a_i = \sum_{i=1}^{k} \langle x, a_i \rangle S^{-1} a_i = \sum_{i=1}^{k} \langle x, a_i \rangle b_i.$$

Now note that S (and S^{-1}) are self-adjoint, as

$$x = SS^{-1}x = \sum_{i=1}^{k} \langle S^{-1}x, a_i \rangle a_i = \sum_{i=1}^{k} \langle x, S^{-1} a_i \rangle a_i = \sum_{i=1}^{k} \langle x, b_i \rangle a_i.$$

It remains to be proved that $\{b_k\}_{k=1}^N$ is a frame. Let A and B be the lower and upper frame bounds for the frame $\{a_k\}_{k=1}^N$. Then

$$\frac{A}{||S||^2} ||x||^2 \leq A||S^{-1}x||^2 \leq \sum_{i=1}^{k} |\langle S^{-1}x, a_i \rangle|^2 = \sum_{i=1}^{k} |\langle x, S^{-1} a_i \rangle|^2 \leq B||S^{-1}x||^2 \leq B||S^{-1}||^2 ||x||^2,$$

which implies

$$\tilde{A}||x||^2 \leq \sum_{i=1}^{k} |\langle x, b_i \rangle|^2 \leq \tilde{B}||x||^2$$

with

$$\tilde{A} = \frac{A}{||S||^2} \text{ and } \tilde{B} = B||S^{-1}||^2.$$

□

Frames plays a fundamental role in the reconstruction of *quantized vectors*. In this context, the concept of *Sobolev frames* arises. See Chapter 8 of [Casazza & Kutyniok '13] and references therein. Due to the *Welch Bound*, a set of vectors having the same angle between any two of them is very important. This motivates the next definition.

Definition 4.30. A system of ℓ_2-normalized vectors $\mathbf{a}_1, \ldots, \mathbf{a}_N$ in $\mathbb{K}^{m\times N}$ is called *equiangular* if there is a constant $c \geq 0$ such that

$$|\langle \mathbf{a}_i, \mathbf{a}_j \rangle| = c \qquad \forall i, j \in [N], i \neq j.$$

Example 4.31. The following examples are important in order to understand properties of equiangular tight frames.

1. An orthonormal basis is an equiangular tight frame for $N = m$;
2. A regular simplex is an equiangular tight frame for $N = m + 1$. For a simple construction of this example, take $N-1$ rows from an $N \times N$ Discrete Fourier Transform matrix. The resulting columns, after being scaled to have unit norm, form an equiangular tight frame;
3. When $m = 1$, any unit norm frame amounts to a list of scalars of unit modulus, and such frames are necessarily equiangular and tight.

Remark 28. *It is important to cite that in the literature* equiangular tight frames *also appear under the names Maximum Welch-Bound-Equality Sequences, optimal Grassmannian frames and two-uniform frames.*

4.8.4 Constraints for Equiangular Tight Frames

In the context of Compressive Sensing, we have the trade off between small coherence and the discrepancy between the number of rows and the number of columns, since we expect that, for $m \times N$ matrices, $N \gg m$. This is the same as saying that we expect to have much fewer measurements than the ambient space. Therefore it is impossible to meet the Welch's bound. Indeed, the next result shows that an equiangular set cannot be arbitrarily large.

Theorem 4.32. *The cardinality N of an equiangular set of ℓ_2-normalized vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_N$ in $\mathbb{K}^m$ satisfies*

$$N \leq \frac{m(m+1)}{2} \qquad \textit{when} \qquad \mathbb{K} = \mathbb{R},$$

$$N \leq m^2 \qquad \textit{when} \qquad \mathbb{K} = \mathbb{C}.$$

If there is equality, then this set of vectors is also a tight frame.

Remark 29. *Is important to note that it is a* one way *theorem, i.e.* if *there is equality, then we have a equiangular tight frame, but this is not a necessary condition. We can have equiangular tight frames without having $N = m(m+1)/2$ or $N = m^2$. This will be very significant in the subsequent discussion.*

To prove this Theorem, we need the following Lemma.

Lemma 4.33. *For any $z \in \mathbb{C}$, the $n \times n$ matrix*

$$\begin{bmatrix} 1 & z & z & \dots & z \\ z & 1 & z & \dots & z \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ z & \dots & z & 1 & z \\ z & \dots & z & z & 1 \end{bmatrix}$$

has $1+(n-1)z$ as a single eigenvalue and $1-z$ as a multiple eigenvalue of multiplicity $n-1$.

Proof. Sum the elements in each line and note that the vector $(1, \dots, 1)$ is an eigenvector for the eigenvalue $1+(n-1)z$. Now, subtracting from the first column each one of the others, we see that the $(n-1)$ linearly independent vectors $(1,-1,0,\dots,0),(1,0,-1,0,\dots,0),\dots,(1,0,\dots,0,-1)$ are eigenvectors for the eigenvalue $1-z$. □

Proof. (of Theorem 4.32): The main idea in this proof is to associate, to every $\mathbf{a}_i$ in the equiangular set, a projection operator from $\mathbb{K}^m$ to $\mathbb{K}^m$, which will be symmetric, in the real case, or hermitian, in the complex case. This association is a linear map, which preserves the inner product when using the Frobenius inner product in the space of linear maps.
As a consequence of Lemma 4.33, the Gram matrix of these operators is invertible, hence they are linearly independent. In the real case, symmetric operators form a subspace of dimension $m(m+1)/2$, and the result follows. In the complex case, hermitian operators do not form a subspace and the minimal subspace that contais them is the whole space, which has dimension m^2. Therefore, in the complex case we need to consider the space of *all* operators on $\mathbb{C}^m$.
Let us define the orthogonal projectors $P_1, \dots, P_N$ onto the lines spanned by $a_1, \dots, a_N$. These operators are:

$$P_i : \mathbb{K}^m \to \mathbb{K}^m, \qquad P_i(v) = \langle v, a_i \rangle a_i,$$

which verify $P_i^2 = P_i = P_i^*$. Endowing $\mathcal{B}(\mathbb{K}^m, \mathbb{K}^m)$, the space of operators in $\mathbb{K}^m$, with the Frobenius inner product

$$\langle P, Q \rangle_F = \operatorname{tr}(PQ^*),$$

and using that the vectors $\mathbf{a}_i$ are equiangular, we obtain that

$$\langle P_i, P_i \rangle_F = \operatorname{tr}(P_i P_i^*) = \operatorname{tr}(P_i) = \sum_{k=1}^{m} \langle P_i(e_k), e_k \rangle = \sum_{k=1}^{m} \langle e_k, a_i \rangle \langle a_i, e_k \rangle = \sum_{k=1}^{m} |\langle a_i, e_k \rangle|^2 = ||a_i||^2 = 1,$$

$$\langle P_i, P_j \rangle_F = \operatorname{tr}(P_i P_j^*) = \operatorname{tr}(P_i P_j) = \sum_{k=1}^{m} \langle P_i P_j(e_k), e_k \rangle = \sum_{k=1}^{m} \langle P_j(e_k), P_i(e_k) \rangle = \sum_{k=1}^{m} \langle e_k, a_j \rangle \overline{\langle e_k, a_i \rangle} \langle a_j, a_i \rangle =$$

$$\overline{\langle a_i, a_j \rangle} \left\langle \sum_{k=1}^{m} \langle a_i, e_k \rangle e_k, a_j \right\rangle = \overline{\langle a_i, a_j \rangle} \langle a_i, a_j \rangle = |\langle a_i, a_j \rangle|^2 = c^2,$$

where we used the canonical basis $\{e_i\}$ in $\mathbb{K}^m$. The Gram matrix for these projectors is

$$\begin{bmatrix} 1 & c^2 & c^2 & \dots & c^2 \\ c^2 & 1 & c^2 & \dots & c^2 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c^2 & \dots & c^2 & 1 & c^2 \\ c^2 & \dots & c^2 & c^2 & 1 \end{bmatrix},$$

so as $0 \leq c < 1$, Lemma 4.33 implies that this Gram matrix is invertible (as all of its eigenvalues are positive) and this means that the system $P_1, \ldots, P_N$ is linearly independent. Thus the theorem follows. Now assume that equality holds. If we concatenate the identity operator to this system, this new one $Id_m, P_1, \ldots, P_N$, will be linearly dependent, then the determinant of its Gram matrix vanishes. This translates into

$$\begin{vmatrix} m & 1 & 1 & 1 & \ldots & 1 \\ 1 & 1 & c^2 & c^2 & \ldots & c^2 \\ 1 & c^2 & 1 & c^2 & \ldots & c^2 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 1 & c^2 & \ldots & c^2 & 1 & c^2 \\ 1 & c^2 & \ldots & c^2 & c^2 & 1 \end{vmatrix} = 0.$$

After dividing the first row by m and substracting from all the other rows and then expanding the determinant with respect to the first column, we obtain an $N-1 \times N-1$ matrix such that

$$\begin{vmatrix} 1-m^{-1} & c^2-m^{-1} & c^2-m^{-1} & \ldots & c^2-m^{-1} \\ c^2-m^{-1} & 1-m^{-1} & c^2-m^{-1} & \ldots & c^2-m^{-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c^2-m^{-1} & \ldots & c^2-m^{-1} & 1-m^{-1} & c^2-m^{-1} \\ c^2-m^{-1} & \ldots & c^2-m^{-1} & c^2-m^{-1} & 1-m^{-1} \end{vmatrix} = 0$$

Multiplying each of the entries by $\frac{m}{m-1}$, which does not alter the determinant, since it is zero, we obtain

$$\begin{vmatrix} 1 & z & z & \ldots & z \\ z & 1 & z & \ldots & z \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ z & \ldots & z & 1 & z \\ z & \ldots & z & z & 1 \end{vmatrix} = 0 \qquad \text{where} \quad z := \frac{mc^2-1}{m-1}$$

As the determinant is zero, at least one the eigenvalues of this matrix must be zero. Since $1 - z = m(1-c^2)(m-1) \neq 0$, Lemma 4.33 implies that $1 + (N-1)z = 0$, which leads to

$$c^2 = \frac{N-m}{m(N-1)}.$$

Then, the ℓ_2-normalized system $a_1, \ldots, a_N$ meets the Welch bound and therefore it is an equiangular tight frame.

□

Now we address the question of whether this upper bound is sharp. The answer is yes, as seen from the four examples given below (two in the real case and two in the complex case).

Example 4.34. ($m = 3$ and $m(m+1)/2 = 6$)
Let $c = (\sqrt{5}-1)/2$. The following vectors form a system of 6 equiangular vectors in $\mathbb{R}^3$:

$$(1, c, 0), (0, 1, c), (c, 0, 1), (1, -c, 0), (0, 1, -c), (-c, 0, 1).$$

Example 4.35. ($m = 7$ and $m(m+1)/2 = 28$)
Let the vectors obtained by unit cyclic shifts of the following four vectors:

$$(1, 1, 0, 1, 0, 0, 0), (1, -1, 0, 1, 0, 0, 0), (1, 1, 0, -1, 0, 0, 0), (1, -1, 0, -1, 0, 0, 0).$$

They form an equiangular system of 28 vectors in $\mathbb{R}^7$.

Example 4.36. ($m = 2$ and $m^2 = 4$)
Let $c = e^{\frac{i\pi}{4}}\sqrt{2-\sqrt{3}}$. The following vectors form a system of 4 equiangular vectors in $\mathbb{C}^2$:

$$(1, c), (c, 1), (1, -c), (-c, 1).$$

Example 4.37. ($m = 3$ and $m^2 = 9$)
Let $c = e^{\frac{i2\pi}{3}}\sqrt{2-\sqrt{3}}$. The following vectors form a system of 9 equiangular vectors in $\mathbb{C}^3$:

$$(-2, 1, 1), (1, -2, 1), (1, 1, -2), (-2, c, c^2), (c^2, -2, c), (c, c^2, -2), (-2, c^2, c), (c, -2, c^2), (c^2, c, -2).$$

Remark 30. *From now on we use the abbreviations RETF(m, N) for an equiangular tight frame with N vectors in $\mathbb{R}^m$ and CETF(m, N), for the analogous frame in $\mathbb{C}^m$.*

The question of whether there exist equiangular frames in every dimension or if there is some kind of rigidity property about them, has been addressed in many articles. Zauner [Zauner '99] made important numerical studies in this area. More information about these computational studies can be found at his website `www.gerhardzauner.at/sicfiducialsd.html`. [Holmes & Paulsen ' 04] and [Sustik et al. '07] are also good references on this topic. In particular, the following Theorem is important.

Theorem 4.38. (Theorem A of [Sustik et al. '07]): *Suppose $N \neq 2m$ and $m > 3$. The existence of a RETF(m, N) implies that*

$$\sqrt{\frac{m(N-1)}{N-m}} \quad \text{and} \quad \sqrt{\frac{(N-m)(N-1)}{m}},$$

are both odd integers. In particular, N is an even number and if $N = m(m+1)/2$, then $m+2$ is necessarily the square of an odd integer. Furthermore, if there exists an RETF(m,$2m$), then m is an odd number and $2m-1$ is the sum of two squares.

The proof of this theorem relies on field theory arguments as in Theorem A of [Sustik et al. '07] and on strongly regular graphs arguments as in Corollary 5.8 of [Waldron '09]. We will prove here, following [Rauhut & Foucart '13], that *"if $N = m(m+1)/2$, then $m+2$ is necessarily the square of an odd integer"* and the obvious consequence that *"m is an odd number"*. The assertion *"$2m-1$ is the sum of two squares"* follows from a theorem of Euler which states that a natural number is the sum of two squares if and only if each prime factor having the form $4k+3$ occurs in the prime factorization with an even power. More details can be found in [Sustik et al. '07].

Proof. Let $a_1, \ldots, a_N$ be a system with $N = m(m+1)/2$ equiangular ℓ_2-normalized vectors. From Theorem 4.23 we know that this system is a tight frame and by the equivalences in Proposition 4.27, the matrix A with columns $a_1, \ldots, a_N$ satisfies $AA^* = \lambda \mathrm{Id}_m$ for some $\lambda > 0$. As the matrix A^*A has the same nonzero eigenvalues as AA^*, then zero is an eigenvalue of multiplicity $N - m$ for A^*A. Moreover, since A^*A is the Gram matrix of the vectors $a_1, \ldots, a_N$, its diagonal entries are all equal to one, while its off-diagonal entries have the same absolute value c. Then $B = \frac{1}{c}(A^*A - \mathrm{Id}_N)$ is

$$B = \begin{pmatrix} 0 & b_{1,2} & \cdots & b_{1,N} \\ b_{2,1} & 0 & \ddots & \vdots \\ \vdots & \ddots & \ddots & b_{N-1,N} \\ b_{N,1} & \cdots & b_{N,N-1} & 0 \end{pmatrix},$$

where $b_{i,j} = \pm 1$. This matrix has $(\lambda - 1)/c$ as an eigenvalue of multiplicity m and $-1/c$ as an eigenvalue of multiplicity $N - m$. Since the characteristic polynomial of a integer matrix has integer coefficients, we can write it as

$$P_B(x) := \sum_{k=0}^{N} \beta_k (-x)^k \quad \text{with } \beta_N = 1.$$

Using Welch's bound and $N = m(m+1)/2$,, we have

$$c = \sqrt{\frac{N-m}{m(N-1)}} = \sqrt{\frac{(m+1)/2-1}{m(m+1)/2-1}} = \sqrt{\frac{m-1}{m^2+m-2}} = \frac{1}{\sqrt{m+2}},$$

so $-\frac{1}{c} = -\sqrt{m+2}$. Then $P_B(-1/c) = P_B(-\sqrt{m+2}) = 0$, i.e.,

$$\left(\sum_{0\leq k\leq N/2} \beta_{2k}(m+2)^k\right) + \sqrt{m+2}\left(\sum_{0\leq k\leq (N-1/2)} \beta_{2k+1}(m+2)^k\right) = 0.$$

We notice that these sums, now denoted by Σ_1 and Σ_2, are integers, due to the fact that β_i are integer coefficients of the characteristic polynomial. We then obtain the equality $\Sigma_1^2 = (m+2)\Sigma_2^2$, which implies that $m+2$ is a square, since any prime factor of $m+2$ must necessarily appear an even number of times in its prime factorization.

To finish the proof, it remains to be show that $\gamma = \sqrt{m+2}$ is *odd*. Let $\mathbf{J}_{N\times N}$ be the matrix with all entries equal to one. Its kernel has dimension $N-1$ (since its rank is 1), then it intersects the $(N-m)$-dimensional eigenspace of B corresponding to the eigenvalue $-1/c = -n$, since due to our hypothesis $m \geq 3$ implies that $N = m(m+1)/2 > m+1$, hence $N-1+N-m > N$, resulting in a nontrivial intersection because the sum of the dimensions of two subspaces is greater then the ambient space. So, with an argument analogous to the one in first part of the proof, the matrix $C := (B - \mathrm{Id}_n + \mathbf{J}_{N\times N})/2$ has $-(n+1)/2$ as an eigenvalue.

As the diagonal elements of C are zero, while its off-diagonal entries are all equal to one or zero, its characteristic polynomial can be written as

$$P_{\mathbf{C}}(x) = \sum_{k=0}^{N} c_k(-x)^k \quad \text{with} \quad c_N = 1 \quad \text{and} \quad c_i \in \mathbb{Z}.$$

It vanishes at $x = -(n+1)/2$ and we can rewrite the equality $P_{\mathbf{C}}(-(n+1)/2) = 0$ as

$$(n+1)^N = -\sum_{k=0}^{N-1} 2^{N-k} c_k (n+1)^k.$$

This shows that $(n+1)^N$ is an even integer, hence so is $n+1$ and we conclude that $n = \sqrt{m+2}$ is an odd integer.

□

Now, we address some interesting questions concerning equiangular tight frames. [Fickus & Mixon '15] gives an *"account for recent and future developments in the construction or impossibility of ETFs in various dimensions"*. They remark that despite ETF having a functional nature (as an equality in the Welch's bound, for example), every known way to construct infinite families of ETF is based on some kind of combinatorial techniques such as strongly regular graphs, difference sets and Steiner systems. The conference paper [Casazza, Redmond & Tremain '08] also discusses the problem of frame classification. In the complex case, [Zauner '99] made the following conjecture, that is still open today.

Open Problem: For every $m \geq 2$, there exists CETF(m, m^2).

[Scott & Grassl '10] proved that if $m \leq 17$ or $M \in \{19, 24, 35, 48\}$ then there exist CETF(m, m^2) and provided numerical tests indicating that, up to machine precision, there are CETF(m, m^2) for $m \leq 67$, suggesting that Zauner's conjecture should be true. If this conjecture were true, there would exist$m \times m^2$ matrices with the lowest possible coherence, i.e. $\mu = 1/\sqrt{m+1}$. This would be of great importance not only for sparse recovery but in many questions in Coding Theory. See [Bodmann & Kutyniok '09] and references therein.

We now focus on the existence of matrices with small coherence. For example, [Temlyakov '11] constructed $p \times p^k$ matrices ($p > k$ being a prime number) with coherence bounded above by $(k-1)/\sqrt{p}$. We present here a construction from [Strohmer & Heath '03], based on ideas of [Alltop '80], of an explicit $m \times m^2$ matrix with coherence equal to $1/\sqrt{m}$. Note that this is the limit of the Welch bound when $N \to \infty$.

Proposition 4.39. *For each prime number $m \geq 5$, there is an explicit $m \times m^2$ complex matrix with coherence $\mu = 1/\sqrt{m}$*

Proof. Let us identify the set $[m]$ with $\mathbb{Z}/m\mathbb{Z} = \mathbb{Z}_m$ and introduce, for $k, \ell \in \mathbb{Z}_m$, the *translation* and *modulation* operators $\mathbf{T}_k$ and $\mathbf{M}_l$, defined, for $z \in \mathbb{C}^{\mathbb{Z}_m}$ and $j \in \mathbb{Z}_m$, by

$$(\mathbf{T}_k \mathbf{z})_j = z_{j-k}, \qquad (\mathbf{M}_\ell \mathbf{z})_j = e^{2\pi i \ell j/m} z_j.$$

These operators are isometries of $\ell_2(\mathbb{Z}_m)$. Let $\mathbf{x} \in \mathbb{C}^{\mathbb{Z}_m}$, be the ℓ_2-normalized *Alltop vector*, with components

$$x_j = \frac{1}{\sqrt{m}} e^{2\pi i j^3/m}, \qquad j \in \mathbb{Z}_m.$$

We claim that the $m \times m^2$ matrix constructed with columns $\mathbf{M}_\ell \mathbf{T}_k \mathbf{x}$ for $k, \ell \in \mathbb{Z}_m$, i.e., the matrix

$$(\mathbf{M}_1\mathbf{T}_1\mathbf{x}|\dots|\mathbf{M}_1\mathbf{T}_m\mathbf{x}|\mathbf{M}_2\mathbf{T}_1\mathbf{x}|\dots|\dots|\mathbf{M}_m\mathbf{T}_1\mathbf{x}|\dots|\mathbf{M}_m\mathbf{T}_m\mathbf{x})$$

has coherence $\mu = 1/\sqrt{m}$. To prove this, we need to calculate the inner product between two columns. Taking indices (k, ℓ) and the one indexed by (k', ℓ'), we have

$$\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x} \rangle = \sum_{j \in \mathbb{Z}_m} (\mathbf{M}_\ell \mathbf{T}_k \mathbf{x})_j \overline{(\mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x})_j} =$$

$$\sum_{j \in \mathbb{Z}_m} e^{2\pi i \ell j/m} x_{j-k} e^{-2\pi i \ell' j/m} \overline{x_{j-k'}} = \frac{1}{m} \sum_{j \in \mathbb{Z}_m} e^{2\pi i (\ell - \ell') j/m} e^{2\pi i ((j-k)^3 - (j-k')^3)/m}.$$

Set $a = \ell - \ell$ and $b = k - k'$, so that $(a, b) \neq (0, 0)$. Change the summation index to $h = j - k'$ and we obtain

$$|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x} \rangle| = \frac{1}{m} \left| e^{2\pi i a k'/m} \sum_{h \in \mathbb{Z}_m} e^{2\pi i a h/m} e^{2\pi i ((h-b)^3 - h^3)/m} \right|$$

$$= \frac{1}{m} \left| \sum_{h \in \mathbb{Z}_m} e^{2\pi i a h/m} e^{2\pi i (-3bh^2 + 3b^2 h - b^3)/m} \right| = \frac{1}{m} \left| \sum_{h \in \mathbb{Z}_m} e^{2\pi i (-3bh^2 + (a + 3b^2) h/m)} \right|.$$

Now setting $c = -3b$ and $d = a + 3b^2$, we compute

$$|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x} \rangle|^2 = \frac{1}{m^2} \sum_{h \in \mathbb{Z}_m} e^{2\pi i (ch^2 + dh)/m} \sum_{h' \in \mathbb{Z}_m} e^{2\pi i (ch'^2 + dh')/m} = \frac{1}{m^2} \sum_{h, h' \in \mathbb{Z}_m} e^{2\pi i (h - h')(c(h + h') + d)/m}$$

$$= \frac{1}{m^2} \sum_{h, h'' \in \mathbb{Z}_m} e^{2\pi i h''(c(h'' + 2h') + d)/m} = \frac{1}{m^2} \sum_{h'' \in \mathbb{Z}_m} e^{2\pi i h''(ch'' + d)/m} \left(\sum_{h' \in \mathbb{Z}_m} e^{4\pi i c h'' h'/m} \right).$$

The analysis of the last sum, for each $h'' \in \mathbb{Z}_m$, leads to

$$\sum_{h' \in \mathbb{Z}_m} e^{4\pi i c h'' h'/m} = \begin{cases} m \text{ if } 2ch'' = 0 \mod m \\ 0 \text{ if } 2ch' \neq 0 \mod m \end{cases}.$$

We now study the two cases:

1. $c = 0 \mod m$. From the definition of c, we have $c = -3b$ and since $3 \neq 0 \mod m$, we must have $b = 0 \mod m$. Hence, $d = a + 3b^2 \neq 0 \mod m$ because $a = \ell - \ell' \neq 0 \mod m$, so

$$|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x} \rangle| = \frac{1}{m} \sum_{h'' \in \mathbb{Z}_m} e^{2\pi i d h'/m} = 0.$$

2. $c \neq 0 \mod m$. Since $2 \neq 0 \mod m$, the equality $2ch'' = 0$ can only occur when $h'' = 0 \mod m$, so that

$$|\langle \mathbf{M}_\ell \mathbf{T}_k \mathbf{x}, \mathbf{M}_{\ell'} \mathbf{T}_{k'} \mathbf{x} \rangle| = \frac{1}{m}.$$

Then we conclude that the coherence of this matrix is equal to $\mu = 1/\sqrt{m}$.

□

Remark 31. *A good reference concerning Frames is the homepage of the* Frame Research Center *at the University of Missouri:* `http://www.framerc.org/`*, where up-to-date references and information concerning frames can be found.*

4.9 Analysis of Algorithms

In Chapter 3 we saw that to solve (P_0) is equivalent to ensure the *null space property* for the matrix A. Also, it is difficult to determine if the matrix satisfies this property or not. Therefore, one of our quests is to find sufficient conditions that imply the Null Space Property.

However, it is important to stress that the first guarantees for the working sparse recovery techniques, like Basis Pursuit, were discovered before NSP was invented. Using *mutual coherence*, for the two orthogonal case, and using the *spark*, for the general case, many researchers were able to directly provide these guarantees.

The coherence property, studied in this chapter, is one of the sufficient conditions we can use to guarantee the success of the algorithms described in Chapter 2. It was used before NSP, but we can restate the result in a modern language by using the Null Space Property.

We will now prove that Basis Pursuit works under some conditions on the coherence, i.e., small coherence implies NSP of order s and, consequently, s-sparse recovery by Theorem 3.3. After, we comment similar results for greedy and threshold algorithms.

Theorem 4.40. ([Tropp '04]): *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be a matrix with ℓ_2-normalized columns. If*

$$\mu_1(s) + \mu_1(s-1) < 1,$$

then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ via Basis Pursuit.

Proof. It is necessary and sufficient to prove that the matrix $\mathbf{A}$ satisfies the NSP_s, which is

$$||v_S||_1 < ||v_{\overline{S}}||_1 \qquad \forall\ \mathbf{v} \in \ker\mathbf{A} \backslash \{0\} \quad \text{and} \quad \forall S \subset [N] \quad \text{with} \quad |S| = s.$$

Let $a_1, \ldots, a_n$ denote the columns of $\mathbf{A}$. Then the condition $\mathbf{v} \in \ker\mathbf{A}$ is the same as $\sum_{j=1}^N v_j a_j = 0$. What we have to do is take the inner product with a_i and isolate the term v_i to obtain

$$v_i = v_i \langle a_i, a_i \rangle = -\sum_{j=1, j \neq i}^{N} v_j \langle a_j, a_i \rangle = -\sum_{\ell \in \overline{S}} v_\ell \langle a_\ell, a_i \rangle - \sum_{j \in S, j \neq i} v_j \langle a_j, a_i \rangle.$$

This implies

$$|v_i| \leq \sum_{\ell \in \overline{S}} |v_\ell||\langle a_\ell, a_i\rangle| + \sum_{j \in S, j \neq i} |v_j||\langle a_j, a_i\rangle|.$$

Summing all over i and interchanging the two finite summations yields

$$||v_S||_1 = \sum_{i \in S} |v_i| \leq \sum_{\ell \in \overline{S}} |v_\ell| \sum_{i \in S} |\langle a_\ell, a_i\rangle| + \sum_{j \in S} |v_j| \sum_{i \in S, i \neq j} |\langle a_j, a_i\rangle|$$

$$\leq \sum_{\ell \in \overline{S}} |v_\ell| \mu_1(s) + \sum_{j \in S} |v_j| \mu_1(s-1) = \mu_1(s)||v_{\overline{S}}||_1 + \mu_1(s-1)||v_S||_1.$$

Rearranging this last expression leads to

$$(1 - \mu_1(s-1))||v_S||_1 \leq \mu_1(s)||v_{\overline{S}}||_1.$$

And this implies the NSP$_s$ because $\mu_1(s) < 1 - \mu_1(s-1)$, that is

$$\mu_1(s)||v_S||_1 < (1 - \mu_1(s-1))||v_S||_1 \leq \mu_1(s)||v_{\overline{S}}||_1 \implies ||v_S||_1 < ||v_{\overline{S}}||_1$$

□

We have similar results for the *Orthogonal Matching Pursuit* and for the *Hard Thresholding Pursuit*. Their proofs can be found at [Rauhut & Foucart '13] on pages 123 and 127 respectively:

Theorem 4.41. *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be a matrix with ℓ_2-normalized columns. If*

$$\mu_1(s) + \mu_1(s-1) < 1$$

then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ after at most s iterations of Orthogonal Matching Pursuit.

Theorem 4.42. *Let $\mathbf{A} \in \mathbb{C}^{m \times N}$ be a matrix with ℓ_2-normalized columns. If*

$$2\mu_1(s) + \mu_1(s-1) < 1$$

then every s-sparse vector $\mathbf{x} \in \mathbb{C}^N$ is exactly recovered from the measurement vector $\mathbf{y} = \mathbf{A}\mathbf{x}$ after at most s iterations of Hard Thresholding Pursuit.

We should mention that all the results derived in this chapter are the worst-case scenarios, implying that the kind of guarantees we obtain can be over-pessimistic, as they are supposed to hold for all signals, and for all possible supports of a given cardinality. Average-case analysis is also available and typically outstrips the worst-case behavior by a large margin. These analysis are available at [Tropp '05] and [Schnass & Vandergheynst '07].

4.10 The Quadratic Bottleneck

In Section 4.9 we proved that in linear systems formed by matrices having small coherence, unique sparse recovery can be performed. However, until now, we do not address the question about the relation between N, the ambient space dimension, and m, the number of measurements. Since we are trying to recover s-sparse vectors we ideally expected, in a first moment, to have a close relation between m and s, that is, it is presumed that the number of measurement can scale with the sparsity. Reformulating this reasoning, we expect that $m = Cs$, for some $C > 0$.

Despite the fact that the coherence property is a sufficient condition the main algorithms of sparse recovery to work, we will see in this section that, when using it, we cannot guarantee a linear scale between m and s. On one hand, Theorem 4.41 and Theorem 4.40 state that the recovery of s-sparse

vectors is guaranteed via BP or OMP if $(2s-1)\mu < 1$ is satisfied. On the other hand, we can choose a matrix $A \in \mathbb{C}^{m\times N}$ with small coherence $\mu \approx c/\sqrt{m}$ (for instance, using Proposition 4.39). For matrices with coherence that behaves this way, the condition $(2s-1)\mu < 1$ holds only if

$$m \geq Cs^2,$$

for some $C > 0$. This estimate for the number of measurements is too pessimistic, especially if we are in a context of large s.

It is not possible to overcome this estimate in the context of coherence and we will explain why. Instead of working with the condition $(2s-1)\mu < 1$, let us use the more general condition stated in Theorem 4.41 and Theorem 4.40,

$$\mu_1(s) + \mu_1(s-1) < 1. \tag{4.19}$$

Suppose, by contradiction, that Equation (4.19) holds with $m \leq (2s-1)^2$ and say, $s < \sqrt{N-1}$. This choice of parameters was made only to facilitate calculations. Provided that N is large, say $N \geq 2m$ then we use Welch's bound for the ℓ_1-coherence, Theorem 4.24, and obtain

$$\begin{aligned} 1 > \mu_1(s) + \mu_1(s-1) &\geq s\sqrt{\frac{N-m}{m(N-1)}} + (s-1)\sqrt{\frac{N-m}{m(N-1)}} = (2s-1)\sqrt{\frac{N-m}{m(N-1)}} \\ &\geq s\sqrt{\frac{2(N-m)}{m(N-1)}} \geq \sqrt{\frac{N}{N-1}}. \end{aligned}$$

This is clearly a contradiction. Therefore, it is not possible to combine coherence estimates with a linear scale between m and s. This is known as the *quadratic bottleneck* problem. The real reason behind it is related to Gershgorin's Theorem. All known estimates of coherence require this theorem and therefore this bottleneck will always appear.

In fact, the relation between Gershgorin's Theorem and the estimation of the coherence is through the proof of Theorem 4.21. If one looks carefully, it states that the eigenvalues of the matrix $A_S^* A_S$ lie in the interval $[1-\mu_1(s-1), 1+\mu_1(s-1)]$ and we know from Theorem 1.11 that sparse recovery is related to invertibility of $A_S^* A_S$.

Thus, despite the fact that coherence is a sufficient condition for sparse recovery and is easy to compute compared, e.g., to NSP, it gives us a poor scale between the sparsity and the number of measurements. Particularly, coherence is a local property, because it deals with only two vectors at the same time. In order to circumvent it, new ideas of sufficient conditions for sparse recovery are necessary. Notably, some approach that tries to handle all column vectors of the measurement matrix at the same time. From the development of a new path, we will see in Chapter 7 that it is possible to have an almost linear scale between m and s.

Chapter 5

Restricted Isometry Property

Respect nature. There's no guarantee she will respect you back.
$66°North$[1] *advertising in Reykjavík, Iceland, 2011.*

5.1 Introduction

From a practical point of view, coherence is a useful measure of how suitable the matrix A is for the problem of finding sparse solutions of linear systems. However we saw that the lower bound from Theorem 4.23 implies a quadractic bottleneck, which limits the performance of recovery for rather small sparsity levels, as the number of measurements$m$ is of the order of the square of the sparsity s. There can be no linear scale between sparsity and the number of measurements due to estimates relying in Gershgorin's Theorem, as showed in Section 4.10.

On a closer analysis, coherence only depends on pairs of columns. Maybe some properties which depend on larger sets of columns at the same time can be found to give us a finer criterion for sparse recovery. This was the inspiration of Candès and Tao when they defined the *uniform uncertainty principle*, nowadays known as *restricted isometry property* [Candès & Tao II '06]. It became the widely used tool by the signal processing community to analyze the measurement performance of encoder/decoder pairs [Blanchard, Cartis & Tanner '11].

This new property overcomes the limitation mentioned above and its advantage is twofold. First, up to a logarithmic factor, we can get an *almost linear* scaling between sparsity and the number of measurements. Besides, we will grasp the potential of probabilistic analysis in this framework through the use of powerful results as the *concentration of measure inequality* and *Gordon's Lemma*. These results will help us to find families of good measurement matrices in the optimal regime. This notion of optimality will be made precise along the chapter.

The purpose of this chapter is to study the Restricted Isometry Property. We will prove that is is a sufficient condition for the Null Space Property, and therefore for Basis Pursuit, by Theorem 3.3. Then, we will show how its introduction allows us to bypass the constrains which appear if we're limited by coherence methods. Later, we will explore a deep result of [Bourgain, Dilworth, Ford, Konyagin & Kutzarova '11] about breaking the quadratic bottleneck with deterministic matrices. We will also establish the success of sparse recovery under different reconstruction methods from some conditions on this property. More specifically, after the introduction of restricted isometry constants, we will show how we can use them to prove the suitability of thresholding and greedy methods for sparse recovery. We close the chapter with some limitations on this property and some attempts to fix it.

[1] One of the oldest Icelandic manufacturing companies: `www.66north.is`

5.2 The RIP Constant and its Properties

As we described in the introduction, the idea is to create a finer property for the matrix recovery analysis, involving many columns at the same time. The restricted isometry constant of order s involves all s-tuples of columns simultaneously and measures how close to an isometry, for sparse vectors, the matrix is.

Definition 5.1. The s-th *restricted isometry constant* $\delta_s = \delta_s(A)$ of a matrix $A \in \mathbb{C}^{m\times N}$ is the smallest $\delta \geq 0$ such that

$$(1-\delta)||x||_2^2 \leq ||Ax||_2^2 \leq (1+\delta)||x||_2^2, \tag{5.1}$$

for all s-sparse vectors $x \in \mathbb{C}^N$. Equivalently, it is given by

$$\delta_s = \max_{S\subset[N],\ \#S\leq s} ||A_S^* A_S - Id_S||_{2\to 2}.$$

The idea behind this definition is that $A_S^* A_S$, a projection onto a given subspace of s-sparse vectors, will behave almost as an orthogonal transformation. In the literature, it is common to say "this matrix A satisfies RIP". It means that $\delta_s(A)$ is small for reasonably large s. What is understood by large s and small δ_s will be made precise after some discussion. See discussion after Proposition 5.4.

Proposition 5.2. *The two definitions of the restricted isometry property are indeed equivalent.*

Proof. One can start by noticing the equivalence between

$$(1-\delta)||x||_2^2 \leq ||Ax||_2^2 \leq (1+\delta)||x||_2^2 \quad \forall\ x \in \mathbb{C}^N s\text{-sparse}.$$

and

$$|||A_S x||_2^2 - ||x||_2^2| \leq \delta||x||_2^2 \quad \forall\ S\subset[N],\ \#S\leq s \quad \text{and } \forall x \in \mathbb{C}^S. \tag{5.2}$$

Then, $\forall x \in \mathbb{C}^s$,

$$||A_S x||_2^2 - ||x||_2^2 = \langle A_S x, A_S x\rangle - \langle x, x\rangle = \langle (A_S^* A_S - Id)x, x\rangle.$$

Now, the matrix $A_S^* A_S - Id$ is hermitian. So we have

$$\max_{x\in\mathbb{C}^s\setminus\{0\}} \frac{\langle (A_S^* A_S - Id)x, x\rangle}{||x||_2^2} = ||A_S^* A_S - Id||_{2\to 2}.$$

Due to (5.2), (5.1) is equivalent to

$$\max_{S\subset[N],\ \#S\leq s} ||A_s^* A_s - Id||_{2\to 2} \leq \delta.$$

Since δ_s is the smallest such δ and the 2-norm of a matrix is equal to its largest singular value, we have the equality.

□

This characterization will be useful when comparing the restricted isometry constants of a matrix and its coherence. First of all, note that the sequence of such constants is nondecreasing

$$\delta_1 \leq \delta_2 \leq \cdots \leq \delta_s \leq \delta_{s+1} \leq \cdots \leq \delta_N.$$

This follows immediately from the definition of RIP, since every $(s-1)$-sparse vector is also an s-sparse vector. The definition of RIP, given above, can be exceedingly restrictive. There is some asymmetry between the influence of the lower and the upper inequality in the context of sparse recovery. Therefore, we could define the asymmetrical RIP.

Definition 5.3. The s-th *lower restricted isometry constant* $\delta_s^L = \delta_s^L(A)$ of a matrix $A \in \mathbb{C}^{m\times N}$ is the smallest $\delta \geq 0$ such that

$$(1-\delta)||x||_2^2 \leq ||Ax||_2^2 \qquad \forall x \text{ such that } ||x||_0 \leq s. \tag{5.3}$$

Also, the s-th *upper restricted isometry constant* $\delta_s^U = \delta_s^U(A)$ is the smallest $\delta \geq 0$ such that

$$(1+\delta)||x||_2^2 \geq ||Ax||_2^2 \qquad \forall x \text{ such that } ||x||_0 \leq s \tag{5.4}$$

Both the smallest and largest eigenvalues of the Gram matrix $A_S^*A_S$ affect the stability of the reconstruction algorithms. However, it is the smaller eigenvalue which allows to distinguish between sparse vectors from their measurement by A. Clearly, $\delta_S^L < 1$ implies that A injective and so it is a sufficient condition to ensure that no two s-sparse vectors have the same measurements.

The asymmetric RIP also plays an important role in the context of high-dimensional statistics, for example in the study of LASSO and Dantzig Selector. For the recovery of sparse vectors from noisy measurements, the asymmetric RIP appears as a crucial generalization. It is connected to the so-called restricted eigenvalue conditions, see Section 5 of [Oliveira '13], [Bickel, Ritov & Tsybakov '09] for some discussion. Also, Figure 1 of [van de Geer & Buhlmann '09] is especially instructive in order to understand the zoo of concepts related to RIP.

The asymmetric definition is especially important in the context of Gaussian matrices. In Chapter 7 we will develop the connections between random matrices and Compressive Sensing. For now, we just need the following remark from [Blanchard, Cartis & Tanner '11]:

Remark 32. *If A is a Gaussian matrix (see Definition 7.1) then A^*A follows a Wishart distribution. Thus, in order to understand RIP of Gaussian matrices, it is important to analyze the empirical distribution of the eigenvalues from a Wishart matrix. If the expected value of the largest and the smallest eigenvalues were asymmetric with respect to some value then the investigation of δ_s^U and δ_s^L separately could give better results in RIP estimation. Figure 5.1 shows this asymmetry since $\mathbb{E}\lambda_{\max}(A_S^*A_S) = (1+\sqrt{\rho})^2$ and $\mathbb{E}\lambda_{\min}(A_S^*A_S) = (1-\sqrt{\rho})^2$. See [Silverstein '85] and [Geman '80] for more details on these asymptotic formulas.*

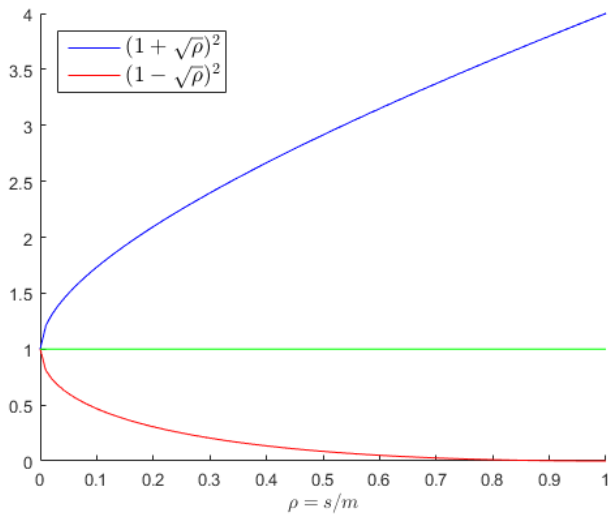


Figure 5.1: For a given $m \times N$ matrix A and $S \subset [N]$ such that $\#S = s$, this figure shows a plot of the expected values of the largest and smallest eigenvalues of a Wishart matrix $A_S^*A_S$. The blue curve represents $\mathbb{E}\lambda_{\max}(A_S^*A_S)$ whereas the red curve represents $\mathbb{E}\lambda_{\min}(A_S^*A_S)$. Note the asymmetry with respect to the constant line equal to 1.

Besides [Blanchard, Cartis & Tanner '11], also the work [Bah & Tanner '10] showed the advantages of the asymmetric definition for sharper results related to RIP for Gaussian matrices.

In this work we are more interested in the signal processing problem than in the statistical problem. Paraphrasing [Oliveira '13], in the former we think about the measurement vector as controlled by the experimenter whereas in the latter, the vectors are generated by a random process. Accordingly, our focus is on the most common definitions historically related to compressive sensing and used by the Signal Processing community: coherence and symmetrical RIP. Now, finally, the proposition which compares both of them.

Proposition 5.4. *If a matrix A has ℓ_2-normalized columns $a_1, \dots, a_N$, then*

$$\delta_1 = 0, \qquad \delta_2 = \mu \qquad \delta_s \leq \mu_1(s-1) \leq (s-1)\mu \qquad \text{for } s \geq 2,$$

where μ and $\mu_1(s-1)$ are the coherence and coherence function from Definition 4.15 and Definition 4.19 respectively.

Proof. By hypothesis, the columns are ℓ_2-normalized, this means that $||Ae_j||_2^2 = ||e_j||_2^2$ for all $j \in [N]$. This is the same as $\delta_1 = 0$ as e_j is 1-sparse. By the equivalence proved above, we have

$$\delta_2 = \max_{1 \leq i \neq j \leq N} ||A_{i,j}^* A_{i,j} - Id||_{2 \to 2} \qquad \text{with} \quad A_{i,j}^* A_{i,j} = \begin{bmatrix} 1 & \langle a_j, a_i \rangle \\ \langle a_i, a_j \rangle & 1 \end{bmatrix}.$$

The eigenvalues of the matrix $A_{i,j}^* A_{i,j} - Id$ are $|\langle a_i, a_j \rangle|$ and $-|\langle a_i, a_j \rangle|$, so its $(2 \to 2)$-norm is $|\langle a_i, a_j \rangle|$, and taking the maximum over $1 \leq i \neq j \leq N$ yields the equality $\delta_2 = \mu$. For $s > 2$, we know from Theorem 4.21 that $(1 - \mu_1(s-1))||x||_2^2 \leq ||Ax||_2^2 \leq (1 + \mu_1(s-1))||x||_2^2$. Since δ_s is the smallest constant to attain this kind of inequality, the theorem follows.

□

This theorem allows us to prove the existence of $m \times m^2$ matrices with $\delta_s < 1$ for $s \leq \sqrt{m}$, since we know that matrices with coherence equal $1/\sqrt{m}$ exists. This is just a simples calculation:

$$\delta_s \leq \mu_1(s-1) \leq (s-1)\mu < s\mu = \frac{s}{\sqrt{m}} \leq 1 \qquad \text{for all } s \leq \sqrt{m}.$$

The main point about using *random matrices* is that we can establish a much better relation between s and m. In fact, we will show that, given $\delta < 1$, there exist $m \times N$ matrices with $\delta_s \leq \delta$ for $s \leq cm/\ln(eN/m)$ with c depending only on δ. Even more surprisingly, this relation cannot be improved, see Chapter 8.

Now we can finally say "How large s and how small δ_s should be". Matrices that have the small restricted isometry constant of this optimal order are said to satisfy the *restricted isometry property* or just RIP. From now we will make no distinction between the words "property" and "constant".

The next simple proposition, taken from [Candès & Tao II '06], is of great importance.

Proposition 5.5. *Let $u, v \in \mathbb{C}^N$ be vectors with $||u||_0 \leq s$ and $||v||_0 \leq t$. If their supports do not intersect, i.e., $supp(u) \cap supp(v) = \emptyset$, then*

$$|\langle Au, Av \rangle| \leq \delta_{s+t} ||u||_2 ||v||_2$$

Proof. Let $S = \text{supp}(u) \cup \text{supp}(v)$ and let u_S and v_S be the restrictions of u and v to S, respectively, so that $Au = A_S u_S$ and $Av = A_S v_S$. Since $\text{supp}(u) \cap \text{supp}(v) = \emptyset$, $\langle u_S, v_S \rangle = 0$, and this yields

$$|\langle Au, Av \rangle| = |\langle A_S u_s, A_S v_s \rangle - \langle u_S, v_S \rangle| = |\langle (A_S^* A_S - Id) u_S, v_S \rangle| \leq ||(A_S^* A_S - Id) u_S||_2 ||v_S||_2$$

$$\leq ||(A_S^* A_S - Id)||_2 ||u_S||_2 ||v_S||_2.$$

Now, observe that $||u_S||_2 = ||u||_2$ and $||v_S||_2 = ||v||_2$. Using the alternative definition for δ_{s+t}, the conclusion follows.

□

We define now a quantity that allows us to estimate RIP and prove some important inequalities.

Definition 5.6. The (s,t)-*restricted orthogonality constant* (or just ROC) $\theta_{s,t} = \theta_{s,t}(A)$ of a matrix $A \in \mathbb{C}^{m\times N}$ is the smallest $\theta \geq 0$ such that

$$|\langle Au, Av\rangle| \leq \theta ||u||_2 ||v||_2$$

for all disjointly supported s-sparse and t-sparse vectors $u, v \in \mathbb{C}^N$.

Similarly to the case of the coherence and RIP constants, we can give an equivalent definition (which has the same proof of equivalence as above):

$$\theta_{s,t} = \max\{||A_T^* A_S||_{2\to 2},\ S \cap T = \emptyset,\ \text{card}(S) \leq s,\ \text{card}(T) \leq t\}$$

Finally we can relate the two constants by the following Proposition.

Proposition 5.7. *Restricted isometry contants and restricted orthogonality constants are related by the inequalities.*

$$\theta_{s,t} \leq \delta_{s+t} \leq \frac{1}{s+t}(s\delta_s + t\delta_t + 2\sqrt{st}\theta_{s,t}).$$

The special case $s = t$ gives

$$\theta_{s,s} \leq \delta_{2s} \leq \delta_s + \theta_{s,s}.$$

Proof. The first inequality follows from Proposition 5.5 and the definition of ROC. For the second one, we need to show that given an $(s+t)$-sparse vector $x \in \mathbb{C}^N$ with $||x||_2 = 1$, we have

$$\Big|||Ax||_2^2 - ||x||_2^2\Big| \leq \frac{1}{s+t}(s\delta_s + t\delta_t + 2\sqrt{st}\theta_{s,t}).$$

In order to do this, let us separate the vector x into two disjointly supported vectors u and v such that u is s-sparse and v is t-sparse. Then

$$||Ax||_2^2 = \langle A(u+v), A(u+v))\rangle = ||Av||_2^2 + ||Au||_2^2 + 2\text{Re}\langle Au, Av\rangle.$$

As u and v are disjointly supported, we have $||x||_2^2 = ||u||_2^2 + ||v||_2^2$ and so

$$\begin{aligned}\Big|||Ax||_2^2 - ||x||_2^2\Big| &\leq \Big|||Au||_2^2 - ||u||_2^2\Big| + \Big|||Av||_2^2 - ||v||_2^2\Big| + 2|\langle Au, Av\rangle| \\ &\leq \delta_s||u||_2^2 + \delta_t||v||_2^2 + 2\theta_{s,t}||u||_2||v||_2 = f(||u||_2^2),\end{aligned}$$

where we have, for $k \in [0,1]$

$$f(k) = \delta_s k + \delta_t(1-k) + 2\theta_{s,t}\sqrt{k(1-k)}.$$

Taking the derivative and equating it to zero, we have

$$f'(k) = \delta_s - \delta_t + \theta_{s,t}(1-2k)\big(k(1-k)\big)^{-1/2} = 0.$$

Therefore, the critical points are the solutions of the equation

$$(\delta_s - \delta_t)\big(k(1-k)\big)^{1/2} + \theta_{s,t}(1-2k) = 0.$$

Making the substitution $1 - 2k = A$, yields

$$\frac{(\delta_s - \delta_t)}{2}\big[(1-A^2)(1+A^2)\big]^{1/2} + \theta_{s,t}A^2 = 0.$$

After solving this equation for A and substituting for the original variable k, we discover that one of the roots is

$$k^* = \frac{1}{2} - \frac{1}{2}(\delta_s - \delta_t)\big(4\theta_{s,t}^2 + (\delta_s - \delta_t)^2\big)^{-1/2}.$$

It can be proved, after simple calculations, that it lies in $[0, 1]$. Besides, the function $f(k)$ is nondecreasing on $[0, k^*]$ and nonincreasing on $[k^*, 1]$. Depending on the location of k^* with respect to $s/(s+t)$, the function f is either nondecreasing on $[0, s/(s+t)]$ or nonincreasing on $[s/(s+t), 1]$. There is freedom to choose u. So, without loss of generality, we will assume that $||u||_2^2$ is in one of these intervals. Taking u having the s smallest absolute entries of x and v having the t largest absolute entries of x, we have $u_i \leq v_j \ \forall i, j$, thus

$$\frac{||u||_2^2}{s} \leq \frac{||v||_2^2}{t} = \frac{1 - ||u||_2^2}{t} \quad \text{and then} \quad ||u||_2^2 \leq \frac{s}{s+t}.$$

In the other case, where u belongs to the other interval and u had the s largest absolute entries of x, then we would likewise have $||u||_2^2 \geq s/(s+t)$. We conclude that

$$\Big|\,||Ax||_2^2 - ||x||_2^2\Big| = f(||u||_2^2) \leq f\left(\frac{s}{s+t}\right) = \delta_s \frac{s}{s+t} + \delta_t \frac{t}{s+t} + 2\theta_{s,t}\frac{\sqrt{st}}{s+t}.$$

□

Sometimes, when proving convergence of the algorithms, some sufficient conditions based on δ_{ks} (for very large k) can be found. For example, in Table 5.5 below, we will see an example that needs $\delta_{8s} < 1$, provided by [Zhou, Kong & Xiu '13]. So it is important to know how to control the restricted isometry constants and restricted orthogonality constant of high order by those of lower order.

Proposition 5.8. *For integers $r, s, t \geq 1$ with $t \geq s$,*

$$\theta_{t,r} \leq \sqrt{\frac{t}{s}}\theta_{s,r} \quad \textit{and} \quad \delta_t \leq \frac{t-d}{s}\delta_{2s} + \frac{d}{s}\delta_s,$$

where $d = \gcd(s, t)$, where gcd *denotes the great common divisor. The special case $t = cs$ leads to*

$$\delta_{cs} \leq c\delta_{2s}$$

Proof. From the definitions of these constants, we need to show that, given a t-sparse vector u and a r-sparse vector v with disjoin support, we have

$$|\langle Au, Av\rangle| \leq \sqrt{\frac{t}{s}}\theta_{s,r}||u||_2||v||_2 \tag{5.5}$$

$$\Big|\,||Au||_2^2 - ||u||_2^2\Big| \leq \left(\frac{t-d}{s}\delta_{2s} + \frac{d}{s}\delta_s\right)||u||_2^2 \tag{5.6}$$

As d is the great common divisor of s and t, we can write $s = kd$ and $t = nd$ for some integers k, n. Denoting the support of u by $T = \{j_1, j_2, \ldots, j_t\}$, we will partition it in n subsets $S_1, S_2, \ldots, S_n \subset T$ of size s defined by

$$S_i = \{j_{(i-1)d+1}, j_{(i-1)d+2}, \ldots, j_{(i-1)d+s}\}$$

with indexes meant modulo t. In this trick partition, each $j \in T$ belongs to exactly $s/d = k$ sets S_i, which gives

$$u = \frac{1}{k}\sum_{i=1}^{n} u_{S_i} \quad \text{and} \quad ||u||_2^2 = \frac{1}{k}\sum_{i=1}^{n}||u_{S_i}||_2^2$$

Let us see through an example why this partitioning makes sense. Suppose that $t = 10 = 2 \times 5$ and $s = 8 = 2 \times 4$, with $d = \gcd(10, 8) = 2$. So, we have $T = \{j_1, j_2, \dots, j_{10}\}$, which we will partition into 5 sets with 8 elements:

$$\begin{aligned}
S_1 &= \{j_1, \dots, j_8\}, \\
S_2 &= \{j_{2+1}, \dots, j_{2+8}\} = \{j_3, \dots, j_{10}\}, \\
S_3 &= \{j_{2\times 2+1}, \dots, j_{2\times 2+8}\} = \{j_5, \dots, j_{10}, j_1, j_2\}, \\
S_4 &= \{j_{3\times 2+1}, \dots, j_{3\times 2+8}\} = \{j_7, \dots, j_{10}, j_1, \dots, j_4\}, \\
S_5 &= \{j_{4\times 2+1}, \dots, j_{4\times 2+8}\} = \{j_9, j_{10}, j_1, \dots, j_6\}.
\end{aligned}$$

Now it is easy to verify that each j_i is contained in only 4 of these 5 sets and that the decomposition of u made above, dividing by k to take repetition into account, holds.

Back to the proof, to prove (5.5) we do the following calculations

$$|\langle Au, Av\rangle| \leq \frac{1}{k}\sum_{i=1}^{n} |\langle Au_{S_i}, Av\rangle| \leq \frac{1}{k}\sum_{i=1}^{n} \theta_{s,r}||u_{S_i}||_2||v||_2$$

$$\leq \theta_{s,r}\frac{\sqrt{n}}{k}\left(\sum_{i=1}^{n}||u_{S_i}||_2^2\right)^{1/2}||v||_2 = \theta_{s,r}\sqrt{\frac{n}{k}}||u||_2||v||_2 = \theta_{s,r}\sqrt{\frac{t/d}{s/d}}||u||_2||v||_2.$$

And inequality (5.6) follows from

$$|||Au||_2^2 - ||u||_2^2| = |\langle (A^*A - Id)u, u\rangle| \leq \frac{1}{k^2}\sum_{i=1}^{n}\sum_{j=1}^{n} |\langle (A^*A - Id)u_{S_i}, u_{S_j}\rangle|$$

$$= \frac{1}{k^2}\left(\sum_{1\leq i\neq j\leq n} |\langle (A^*_{S_i\cup S_j}A_{S_i\cup S_j} - Id)u_{S_i}, u_{S_j}\rangle| + \sum_{i=1}^{n} |\langle (A^*_{S_i}A_{S_i} - Id)u_{S_i}, u_{S_i}\rangle|\right)$$

$$\leq \frac{1}{k^2}\left(\sum_{1\leq i\neq j\leq n} \delta_{2s}||u_{S_i}||_2||u_{S_j}||_2 + \sum_{i=1}^{n}\delta_s||u_{S_i}||_2^2\right) = \frac{\delta_{2s}}{k^2}\left(\sum_{i=1}^{n}||u_{S_i}||_2\right)^2 - \frac{\delta_{2s}-\delta_s}{k^2}\sum_{i=1}^{n}||u_{S_i}||_2^2$$

$$\leq \left(\frac{\delta_{2s}n}{k^2} - \frac{\delta_{2s}-\delta_s}{k^2}\right)\sum_{i=1}^{n}||u_{S_i}||_2^2 = \left(\frac{n}{k}\delta_{2s} - \frac{1}{k}(\delta_{2s}-\delta_s)\right)||u||_2^2$$

$$= \left(\frac{t}{s}\delta_{2s} - \frac{1}{k}(\delta_{2s}-\delta_s)\right)||u||_2^2 = \left(\frac{t-d}{s}\delta_{2s} + \frac{d}{s}\delta_s\right)||u||_2^2.$$

□

As in the coherence case, it is important to know how the scaling between the number of measurements m and the sparsity s affects RIP. We will prove now that $\delta_s \geq c\sqrt{s/m}$. For $s = 2$, this can be interpreted as $\delta_2 = \mu \geq \tilde{c}/\sqrt{m}$, which reminds us of the Welch bound, Theorem 4.23.

Theorem 5.9. *For $A \in \mathbb{C}^{m\times N}$ and $2 \leq s \leq N$, there exists constants c, C and δ_* such that for $N \geq C$, $\delta_s \leq \delta_*$ one has*

$$m \geq c\frac{s}{\delta_s^2}$$

For instance, we could choose $c = 1/162$, $C = 30$ and $\delta_* = 2/3$.

Proof. We start by noticing that the theorem cannot be valid for $s = 1$, as $\delta_1 = 0$ if all the columns of A have ℓ_2-norm equal to 1. If we set $t = \lfloor s/2 \rfloor \geq 1$, then we will decompose the matrix A into blocks of size $m \times t$ (the last block could have less columns)

$$A = [\ A_1 \mid A_2 \mid \ \dots \ \mid A_n\ ], \qquad N \leq nt.$$

From Proposition 5.2, we can use the alternative definition of the restricted isometry constant and also the definition of the restricted orthogonality constant. So we have, for all $i, j \in [n], i \neq j$, that

$$||A_i^* A_i - Id||_{2\to 2} \leq \delta_t \leq \delta_s \qquad \text{and} \qquad ||A_i^* A_j||_{2\to 2} \leq \theta_{t,t} \leq \delta_{2t} \leq \delta_s.$$

Then we conclude that the eingenvalues of $A_i^* A_i$ and the singular values of $A_i^* A_j$ satisfy

$$1 - \delta_s \leq \lambda_k(A_i^* A_i) \leq 1 + \delta_s,$$

Now, we define the matrices $H = AA^* \in \mathbb{C}^{m\times m}$ and $G = A^*A = [A_i^* A_j]_{1\leq i,j\leq n} \in \mathbb{C}^{N\times N}$. First we deduce the estimate

$$\operatorname{tr}(H) = \operatorname{tr}(G) = \sum_{i=1}^{n} \operatorname{tr}(A_i^* A_i) = \sum_{i=1}^{n}\sum_{k=1}^{t} \lambda_k(A_i^* A_i) \geq nt(1-\delta_s). \tag{5.7}$$

Then, using the Frobenius inner product $\langle M_1, M_2\rangle_F = \operatorname{tr}(M_2^* M_1)$, we deduce

$$\begin{aligned}
\operatorname{tr}^2(H) = \langle \mathrm{Id}_m, H\rangle_F^2 \leq ||\mathrm{Id}||_F^2 ||H||_F^2 = m\ \operatorname{tr}(H^*H) = m\ \operatorname{tr}(AA^*AA^*) = m\ \operatorname{tr}(A^*AA^*A) = \\
m\ \operatorname{tr}(GG^*) = m\ \sum_{i=1}^{n} \operatorname{tr}\Big(\sum_{j=1}^{m} A_i^* A_j A_j^* A_i\Big) = m\Bigg[\sum_{1\leq i\neq j\leq n}\sum_{k=1}^{t} \sigma_k(A_i^* A_j)^2 + \sum_{i=1}^{n}\sum_{k=1}^{t} \lambda_k(A_i^* A_i)^2\Bigg] \\
\leq m\big[n(n-1)t\delta_s^2 + nt(1+\delta_s)^2\big] = mnt\big[(n-1)\delta_s^2 + (1+\delta_s)^2\big].
\end{aligned} \tag{5.8}$$

From (5.7) and (5.8), we derive

$$m \geq \frac{nt(1-\delta_s)^2}{(n-1)\delta_s^2 + (1+\delta_s)^2}.$$

If $(n-1)\delta_s^2 < (1+\delta_s)^2/5$, we would obtain, using $\delta_s \leq 2/3$,

$$m > \frac{nt(1-\delta_s)^2}{6(1+\delta_s^2)/5} \geq \frac{5(1-\delta_s)^2}{6(1+\delta_s)^2} N \geq \frac{1}{30} N.$$

This is a contradiction since we assumed that $N \geq 30m$. Therefore we have $(n-1)\delta_s^2 \geq (1+\delta_s)^2/5$. Hence, using again $\delta_s \leq 2/3$ and $s \leq 3t$, we conclude

$$m \geq \frac{nt(1-\delta_s)^2}{6(n-1)\delta_s^2} \geq \frac{1}{54}\frac{t}{\delta_s^2} \geq \frac{1}{162}\frac{s}{\delta_s^2} = c\frac{s}{\delta_s^2}.$$

□

5.3 The Quadratic Bottleneck Reloaded

Theorem 5.9 helps us estimate how small the restricted isometry constant can be. Even more, we can compare the scaling between s, the sparsity of the signals we want to recover, and m, the number of measurements which guarantees the smallness of δ_s.

We will see that there is a great difference between the two sufficient conditions we are working in this dissertation: coherence and RIP. The former provides a quadratic scale between m and s while the latter provides, using probabilistic techniques, an almost linear scale. Also, until recently, the typical technique

to estimate RIP was through coherence and then the phenomenon of a bottleneck appears again. Let us see precisely what all of this means.

Theorem 5.9 provides a linear scaling between s and m

$$m \geq C\delta_s^{-2}\ s$$

On the other hand, if we choose a matrix A with an optimal coherence $\mu = C/\sqrt{m}$ then Theorem 5.4 implies that $\delta_s \leq (s-1)\mu \leq cs/\sqrt{m}$, which leads to

$$C_1 \frac{s}{\sqrt{m}} \geq \delta_s \geq C_2 \sqrt{\frac{s}{m}}, \tag{5.9}$$

for some universal constants C_1 and C_2. Note that the right-hand side is valid for *all* matrices. Meanwhile, for the left-hand side we know how to prove that an explicit family of matrices indeed attain this optimal order of coherence and, consequently, this scale for RIP.

Then, the quadratic scale $m \geq Cs^2$ is a sufficient condition for δ_s to be small. Also, the right side of the inequality (5.9) shows that $m \geq \tilde{C}s$ is a necessary condition. There is a significant gap between both of them and up to this point in this dissertation we are unable to show if this second condition is also sufficient for small RIP constant or not.

In Chapter 7 we will exhibit certain random matrices satisfying $\delta_s \leq \delta$ with high probability provided

$$m \geq C\delta^{-2} s \ln(eN/s). \tag{5.10}$$

This will prove that there exists some (actually, many!) matrices A with small RIP in an asymptotic regime much closer to linear than quadratic. Additionally, in Chapter 8 we will prove that $\delta_s \leq \delta$ *requires* that $m \geq C_\delta s \ln(eN/s)$ for a certain constant C_δ depending on δ. In the literature it is usually said that in order to obtain small RIP constant, *linear scale is optimal up to logarithm factors.*

All the constructions of matrices in the optimal regime involve powerful probabilistic arguments like concentration of measure inequalities. Despite this, it is common in applications to have fixed and deterministic sensing matrices. So we have a problem of constructing (or of guaranteeing that) deterministic matrices satisfying RIP in the optimal regime. On the left hand side of (5.9), setting $\delta_s = \delta_*$, we have $s \geq C_1^{-1}\delta_*\sqrt{m}$ which could be reformulated as $s = \Omega(m^{1/2})$. This is the regime in which *we know how to* guarantee small RIP constant through coherence techniques. However, ideally we would like to improve this and increase the exponent from $1/2$ to $1/2 + \varepsilon$, i.e. break the bottleneck. Then, it will be great to make $\varepsilon \to 1/2$, and provide small RIP in a suitable scale for applications. This will shows the existence of deterministic matrices with RIP property in an *almost linear* regime. Summarizing this, and following [Mixon '15], we provide the following definition.

Definition 5.10. For any $z > 0$, ExRIP[z] denote the following statement:

There exists an explicit family of $M \times N$ matrices with arbitrarily large aspect ratio N/M which have restricted isometry property of order s with constant δ and $s = \Omega(M^{z-\varepsilon})$ for all $\varepsilon > 0$ and $\delta < 1/3$.[2]

The main point is that with the aid of coherence, it is straightforward to prove ExRIP[1/2]. One can use the construction of Proposition 4.39, for example, and the argument above which estimate RIP through coherence. The latter, in its turn, is estimated through Gershgorin's Theorem. *This was the only technique to estimate RIP until 2011.*

The next step is to prove ExRIP[$1/2+\varepsilon_0$] for some $\varepsilon_0 > 0$. That is, we need to construct of an explicit family of matrices A_i such that for $C > 0$, we have $s_i > Cm_i^{1/2-\varepsilon}$ for $\varepsilon > 0$. Here s_i denotes the sparsity level and m_i the number of rows of matrix A_i.

This important problem was solved by [Bourgain, Dilworth, Ford, Konyagin & Kutzarova '11]. The authors created a new technique bypassing Gershgorin's Thereom through additive combinatorics arguments to construct a family of matrices in this regime. This will be the subject of the next section. Despite this major theoretical breakthrough, this is the only known explicit construction which breaks

[2] In section 5.5 the number 1/3 will become clear. Essentially, we can guarantee that with $\delta_s < 1/3$, basis pursuit works to recover sparse vectors. This was proved by [Cai & Zhang '13].

the bottleneck. Besides, for practical purpose, this constant is not useful and does not help im improving computational time, in finding a better measurement scheme or in a suitable scale for practical purposes. In [Mixon's Blog - 12/02/2013], this constant was estimated to be $\varepsilon_0 \approx 5.5169 \times 10^{-28}$. Recently, in four blog posts, Dustin Mixon and Afonso Bandeira[3] optimized the constant to $\varepsilon_0 \approx 4.4466 \times 10^{-24}$.

It is a major problem to go further and prove this in the optimal (linear up to logarithm factor) regime:

Open Problem: Construct deterministic (or constructible in polynomial time) matrices with $\delta_s \leq \delta$ in the optimal regime $m \geq C\delta^{-2} s \ln(eN/s)$

Therefore, until now there is no way to guarantee that deterministic families of matrices coming from applications have small RIP constant in the optimal regime. This makes it impossible to use RIP as a sufficient condition for compressive sensing with deterministic matrices. In spite of the lack of guarantees, applied researchers continue to prefer deterministic matrices as sensing matrices due to well known constructions.

5.4 Breaking the Bottleneck

In this section we briefly describe the problem of estimating RIP via coherence and provide a big picture of the techniques used by Bourgain et al. [Bourgain, Dilworth, Ford, Konyagin & Kutzarova '11], now known as *BDFKK restricted isometry machine* [Mixon '15].

Let S be a set with $\#S = s$. For a matrix A satisfying RIP with δ_s constant, the eigenvalues of $A_S^* A_S$ lie in $[1-\delta_s, 1+\delta_s]$. Therefore we can prove that a matrix has small RIP constant by estimate its eigenvalues. From Theorem 4.21, we known that a matriz A always satisfies RIP with constant $(s-1)\mu$, where μ is it coherence. But the Welch's bound, Theorem 4.23, says that coherence cannot be too small. For $N \geq cM$, $\mu \geq \Omega(M^{-1/2})$ and then, for $\delta < 1/3$ as in Definition 5.10, we have $s = O(M^{1/2})$. This was a variation of what we presented on the last section and it is the standard technique to prove RIP. One should consult [DeVore '07] and [Applebaum, Howard, Searle & Calderbank '09] for examples of this kind of construction.

The problem with this technique is that it relies on Gershgorin Theorem. When estimating the eigenvalues of the Gram matrix $A_S^* A_S$, this theorem only takes into account their *magnitude*. Bourgain et al realized that their *signs* should also be considered and then some more sophisticated combinatorial techniques must be used in order to "cancelate" these signs. So, the initial idea was to convert the RIP statement, about all s-sparse vectors simultaneously, into a statement about finitely many vectors. Toward this, they introduced the following definition.

Definition 5.11. Let $A = [\ a_1 \mid \ldots \mid a_N\ ] \in \mathbb{C}^{m \times N}$ be a matrix with $\{a_i\}_{i \in [N]}$ being its column vectors. We say that A satisfies θ_s-flat restricted isometry property (or flat-RIP of order s and constant θ) if for every disjoint $I, J \subseteq [N]$ with $\#I, \#J \leq s$

$$\left| \left\langle \sum_{i \in I} a_i, \sum_{j \in J} a_j \right\rangle \right| \leq \theta \sqrt{\#I \#J}$$

Note that A has flat-RIP by taking u and v to be the characteristic function χ_I and χ_J into the Definition 5.6. So it is also called the flat-ROC property, after [Bandeira et al. '13]. With this definition in hands we can change the estimation of RIP by the estimation of flat-RIP with the following Theorem:

Theorem 5.12. *If A has the flat restricted isometry property with constant θ_s and has unit-norm columns then A has RIP (for s-sparse vectors) with constant $150\theta \log s$.*

This theorem was proved in [Bourgain, Dilworth, Ford, Konyagin & Kutzarova '11] for $s \geq 2^{10}$ and it was stated there not as here, but in a similar way. The proof for all s as well as the statement like the one the give here appeared in [Bandeira et al. '13]. The former also defined the following:

[3] The progress of this constant optimization was described in the Math Research Wiki [Wiki - Deterministic RIP Matrices].

Definition 5.13. We say that $A = [\ a_1\ |\ldots|\ a_N\ ] \in \mathbb{C}^{m\times N}$ satisfies the θ'_s-weak flat RIP if for every disjoint $I, J \subseteq [N]$ with $\#I, \#J \leq s$,

$$\left|\left\langle \sum_{i\in I} a_i, \sum_{j\in J} a_j \right\rangle\right| \leq \theta' s.$$

And so the following connection between weak flat-RIP and flat-RIP can be proved:

Lemma 5.14. (Essentially Lemma 1 from [Bourgain, Dilworth, Ford, Konyagin & Kutzarova '11]): *If A satisfies the θ'_s-weak flat RIP and has coherence $\mu \leq 1/s$, then A has flat RIP of order s with constant $\sqrt{\theta'}$.*

Proof. By triangle inequality, we have

$$\left|\left\langle \sum_{i\in I} a_i, \sum_{j\in J} a_j \right\rangle\right| \leq \sum_{i\in I}\sum_{j\in J} |\langle a_i, a_j\rangle| \leq \mu \#I\#J \leq \frac{\#I\#J}{s}.$$

Also, A satisfies the weak flat RIP, so

$$\left|\left\langle \sum_{i\in I} a_i, \sum_{j\in J} a_j \right\rangle\right| \leq \min\{\theta' s, \#I\#J/s\} \leq \sqrt{\theta'\#I\#J}.$$

□

The following lemma, presented here without proof, is of fundamental importance in order to break the bottleneck.

Lemma 5.15. (Essentially Lemma 3 from [Bourgain, Dilworth, Ford, Konyagin & Kutzarova '11]): *If A satisfies RIP of order s with constant δ_s then A satisfies RIP of order αs with constant $2\alpha\delta_s$ for all $\alpha \geq 1$.*

These three results together allow us to convert results with modest s and tiny δ_s into large s and modest δ_s. To see how, we will denote a matrix which satisfies RIP of order s with constant δ_s by (s, δ_s)-RIP, one which satisfies flat-RIP of order s with constant δ_s by (s, δ_s)-fRIP and one which satisfies weak flat-RIP by (s, δ_s)-wfRIP. Given this notation, with the aid of Theorem 5.12 and Lemmas 5.14 and 5.15, the following chain of implications can be proved:

$$(2, [\delta_s/(2s150\log s)]^2)\text{-wfRIP} \implies (2, \delta_s/(2s150\log s))\text{-fRIP} \implies (2, \delta_s/s)\text{-RIP} \implies (s, \delta_s)\text{-RIP}.$$

Due to the combinatorial nature of RIP, it is better to get results for 2-sparse vectors then for s-sparse vectors for large s. Instead of constructing matrices with a certain RIP, they construct matrices with appropriate weak flat-RIP. For these, the sign cancellation can be performed. The columns of these matrices where based on complex exponentials knows as *chirps*.

Definition 5.16. Let p be a prime number and $\mathbb{F}_p$ be the field of size p. We define the *chirp* as a complex exponential of the form

$$u_{a,b} = \frac{1}{\sqrt{p}}(e^{2\pi i(ax^2+bx)/p})_{x\in\mathbb{F}_p}.$$

These exponentials were used as entries of the Gram matrix A^*A because their inner product has some interesting properties. If $a_1 = a_2$, then $\langle u_{a_1,b_1}, u_{1_2,b_2}\rangle = 1$ if $b_1 = b_2$ and 0 if $b_1 \neq b_2$. Otherwise, the inner product is

$$\langle u_{a_1,b_1}, u_{1_2,b_2}\rangle = \frac{1}{p}\sum_{x\in\mathbb{F}_p} e^{2\pi i\left((a_1-a_2)x^2+(b_1-b_2)x\right)/p}.$$

This expression can be manipulated with the aid of Legendre symbols[4] and then it is possible to perform the cancellation in the signs and, consequently, improve the estimates of RIP. Since we want these expressions in the Gram matrix A^*A, the columns of A must be $\{u_{a,b}\}_{(a,b)\in\mathcal{A}\times\mathcal{B}}$ for some well-designed sets $\mathcal{A},\mathcal{B}\subseteq\mathbb{F}_p$. Bourgain et at. constructed two set $\mathcal{A}$ and $\mathcal{B}$ with *low additive energy*, in the sense of additive combinatorics [Mixon '15]. This construction is the highly technical part of the paper.

After defining such set, it was proved that for sufficiently large p, the $p\times\#(\mathcal{A})\#(\mathcal{B})$ matrix with columns $u_{a,b}$, for $a\in\mathcal{A}$ and $b\in\mathcal{B}$ satisfies $(p^{1/2+\varepsilon_0-\varepsilon},\delta)$-RIP for any $\varepsilon>0$ and $\delta<\sqrt{2}-1$, thereby implying ExRIP$[1/2+\varepsilon_0]$. This leads to the natural question of how much these techniques could be improved in order to give $1/2+\varepsilon_0\to 1/2+1/2$.

5.5 Analysis of Algorithms

In Chapter 4 we analyzed Basis Pursuit and proved that when coherence is small, the algorithm works and recovers all s-sparse solutions of the linear system. In this section we provide two analogous results for the RIP. We provide two results. The first one is simple and natural while the second is more sophisticated and involved. The philosophy behind both of them is the same: when RIP is smaller than some constant, then the algorithm must work.

Theorem 5.17. *Suppose that the $2s$-th restricted isometry constant of the matrix $A\in\mathbb{C}^{m\times N}$ satisfies*

$$\delta_{2s}<\frac{1}{3}.$$

Then every s-sparse vector $x\in\mathbb{C}^N$ is the unique solution of

$$\min_{z\in\mathbb{C}^N}||z||_1 \qquad \text{subject to } Az=Ax.$$

The following Lemma will be used many times in our proof.

Lemma 5.18. *Given $q>p>0$, if $u\in\mathbb{C}^s$ and $v\in\mathbb{C}^t$ satisfy*

$$\max_{i\in[s]}|u_i|\le\min_{j\in[t]}|v_j|,$$

then

$$||u||_q\le\frac{s^{1/q}}{t^{1/p}}||v||_p.$$

The special case $p=1,q=2$ and $t=s$ leads to

$$||u||_2\le\frac{1}{\sqrt{s}}||v||_1.$$

Proof. Notice that

$$\frac{||u||_q}{s^{1/q}}=\left[\frac{1}{s}\sum_{i=1}^{s}|u_i|^q\right]^{1/q}\le\max_{i\in[s]}|u_i|\le\min_{j\in[t]}|v_j|\le\left[\frac{1}{t}\sum_{j=1}^{t}|v_j|^p\right]^{1/p}\le\frac{||v||_p}{t^{1/p}}.$$

□

[4]The Legendre symbol is a multiplicative function in Number Theory with values 1, −1, 0 that is a quadratic character modulo a prime number p: its value on a (nonzero) quadratic residue mod p is 1 and on a non-quadratic residue (non-residue) is −1. Its value on zero is 0.

Proof. (of Theorem 5.17). By Theorem 3.3, it is enough to prove that the matrix A satisfies the null space property of order s, that is

$$||v_S||_1 < \frac{1}{2}||v||_1 \qquad \forall v \in \ker A \backslash \{0\} \text{ and all } S \subset [N] \text{ with } \#(S) = s.$$

In fact, we will prove a stronger statement

$$||v_S||_2 \leq \frac{\rho}{2\sqrt{s}}||v||_1 \qquad \forall v \in \ker A \text{ and all } S \subset [N] \text{ with } \#(S) = s,$$

where $\rho = \frac{2\delta_{2s}}{1-\delta_{2s}}$ satisfies $\rho < 1$ whenever $\delta_{2s} < 1/3$. It is clear that we need to consider just the index set $S := S_0$ of the s largest absolute entries of v. So let us partition the complement $\overline{S_0}$ of S_0 in $[N]$ as $\overline{S_0} = S_1 \cup S_2 \cup \dots$ where

- S_1 is the index set of the s largest absolute entries of v in $\overline{S_0}$
- S_2 is the index set of the s largest absolute entries of v in $\overline{S_0 \cup S_1}$

and so on. As $v \in \ker A$, then $Av = 0$ and so $A\big(\sum_{i=0} v_{S_i}\big) = 0$ which implies $A(v_{S_0}) = A(-v_{S_1} - v_{S_2} - \dots)$. This leads to

$$\begin{aligned} ||v_{S_0}||_2^2 \leq \frac{1}{1-\delta_{2s}}||A(v_{S_0})||_2^2 &= \frac{1}{1-\delta_{2s}}\langle A(v_{S_0}), A(v_{S_0})\rangle \\ = \frac{1}{1-\delta_{2s}}\langle A(v_{S_0}), A(-v_{S_1}) + A(-v_{S_2}) + \dots\rangle &= \frac{1}{1-\delta_{2s}}\sum_{k\geq 1}\langle A(v_{S_0}), A(-v_{S_k})\rangle. \end{aligned} \tag{5.11}$$

From Proposition 5.5, we obtain

$$\langle A(v_{S_0}), A(-v_{S_k})\rangle \leq \delta_{2s}||v_{S_0}||_2||v_{S_k}||_2.$$

Putting this into (5.11) we obtain

$$||v_{S_0}||_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}}\sum_{k\geq 1}||v_{S_k}||_2 = \sum_{k\geq 1}\frac{\rho}{2}||v_{S_k}||_2.$$

We defined v_{S_k} in a decreasing way, Hence, for $k \geq 1$, the s absolute entries of v_{S_k} do not exceed any of the s absolute entries of $v_{S_{k-1}}$ and Lemma 5.18 yields

$$||v_{S_k}||_2 \leq \frac{1}{\sqrt{s}}||v_{S_{k-1}}||_1,$$

which gives

$$||v_{S_0}||_2 \leq \frac{\rho}{2\sqrt{s}}\sum_{k\geq 1}||v_{S_{k-1}}||_1 \leq \frac{\rho}{2\sqrt{s}}||v||_1.$$

This implies NSP for matrix A and concludes the proof that basis pursuit works under $\delta_{2s} < 1/3$. □

In (5.11), we interpreted the vector v_{S_0} as being $2s$-sparse. In fact, it is an s-sparse vector. So a better bound $||v_{S_0}||_2^2 \leq ||A(v_{S_0})||_2^2/(1-\delta_s)$ can be used. This will turn into a sufficient condition based on δ_s instead of δ_{2s}. There are many such sufficient conditions involving the constants δ_s. One can cite, for example, the works of [Cai, Wang & Xu I '10] and [Cai & Zhang '13]. The first proved that basis pursuit performs sparse recovery if $\delta_s < 0.307$. The latter, that basis pursuit works under $\delta_s < 1/3$. This is the best known result involving δ_s.

In any case, the condition based on δ_{2s} is more natural since it is known, after [Candès & Tao II '06], that an algorithm recovering all s-sparse vectors x from the measurements $y = Ax$ exists if and only if

Estimate	Numerical Approx.	Paper
$\delta_{2s} + \delta_{3s} < 1$	———	[Candès & Tao II '06]
$\delta_{3s} + 3\delta_{4s} < 2$	———	[Candès, Romberg & Tao II '06]
$\delta_{2s} < \sqrt{2} - 1$	0.4142	[Candès '08]
$\delta_{2s} < 2(3 - \sqrt{2})/7$	0.4531	[Foucart & Lai '10]
$\delta_{2s} < 3/(4 + \sqrt{6})$	0.4651	[Foucart '10]
$\delta_{2s} < 1/(1 + \sqrt{1.25})$	0.4721	[Cai, Wang & Xu I '10]
$\delta_{2s} < 4/(6 + \sqrt{6})$	0.4734	[Foucart II '10][5]
$\delta_{2s} < 0.4931$	0.4931	[Mo & Li '11]
$\delta_{2s} < 1/2$	0.5000	[Cai & Zhang '13]
$\delta_{2s} < 3/2 - (1 + \sqrt{41})/8$	0.5746	[Zhou, Kong & Xiu '13][6]
$\delta_{2s} < 4/\sqrt{41}$	0.6246	[Andersson & Strömberg '14]

Table 5.1: Historical Improvements on RIP Bounds

$\delta_{2s} < 1$. In table 5.5 we present the historical improvements in bounding RIP. We will now prove the last result from this list. This is the best known bound on δ_{2s} so far.

Theorem 5.19. ([Andersson & Strömberg '14])[7]: *Suppose that the $2s$-th restricted isometry constant of the matrix $A \in \mathbb{C}^{m \times N}$ satisfies*

$$\delta_{2s} < \frac{4}{\sqrt{41}} \approx 0.6246. \tag{5.12}$$

Then, for any $x \in \mathbb{C}^N$ and $y \in \mathbb{C}^m$ with $||Ax - y||_2 \leq \eta$, a solution $\tilde{x}$ of

$$\min_{z \in \mathbb{C}^N} ||z||_1 \qquad \text{subject to } ||Az - y||_2 \leq \eta,$$

approximates the vector x with errors

$$||x - \tilde{x}||_1 \leq C\sigma_s(x)_1 + D\sqrt{s}\eta \qquad \text{and} \qquad ||x - \tilde{x}||_2 \leq \frac{C}{\sqrt{s}}\sigma_s(x)_1 + D\eta,$$

where the constants $C, D > 0$ depend only on δ_{2s}.

This theorem is not only the best known bound on RIP but it also incorporates stability and robustness for basis pursuit. From Theorem 3.18, we just need to prove that the matrix A satisfies $\text{NSP}_{2,\rho,\tau}$. Precisely, we are going to prove the following.

Theorem 5.20. *If the $2s$-th restricted isometry constant of $A \in \mathbb{C}^{m \times N}$ obeys (5.12), then the matrix A satisfies the ℓ_2-robust null space property of order s with constants $0 < \rho < 1$ and $\tau > 0$ depending only on δ_{2s}.*

For this, we need a lemma called *square root lifting inequality*. It is a kind of a counterpart of the inequality $||v||_1 \leq \sqrt{s}||v||_2$ for $v \in \mathbb{C}^s$. It is important to note that this inequality, together with *shifting inequality* [Cai, Wang & Xu I '10], are the main techniques used in all the papers in Table 5.5 to improve δ_{2s}. The proof of both results can be found in [Cai, Wang & Xu I '10] and [Cai, Wang & Xu II '10].

Lemma 5.21. ([Cai, Wang & Xu I '10]): *For $a_1 \geq a_2 \geq \cdots \geq a_s \geq 0$,*

$$\sqrt{a_1^2 + \cdots + a_s^2} \leq \frac{a_1 + \cdots + a_s}{\sqrt{s}} + \frac{\sqrt{s}}{4}(a_1 - a_s).$$

[5] Valid for large s. It needs some limit arguments when $s \to \infty$.

[6] This result needs also a technical condition $\delta_{8s} < 1$. All the other results do not need supplementary conditions to hold.

[7] The theorem appeared in the literature for the first time on the book [Rauhut & Foucart '13], published on 2013. It was based on a preprint version of the paper [Andersson & Strömberg '14] from 2012.

Proof. Let us start by noting that the lemma is equivalent to the following statement

$$\begin{cases} a_1 \geq a_2 \geq \cdots \geq a_s \geq 0 \\ (a_1 + \cdots + a_s)/\sqrt{s} + \frac{\sqrt{s}}{4} a_1 \leq 1 \end{cases} \implies \sqrt{a_1^2 + \cdots + a_s^2} + \frac{\sqrt{s}}{4} a_s \leq 1.$$

Therefore, we need to maximize the convex function

$$f(a_1, a_2, \ldots, a_s) = \sqrt{a_1^2 + \cdots + a_s^2} + \frac{\sqrt{s}}{4} a_s,$$

over the convex polytope

$$C = \left\{ (a_1, \ldots, a_s) \in \mathbb{R}^s : a_1 \geq \cdots \geq a_s \geq 0 \text{ and } \frac{a_1 + \cdots + a_s}{\sqrt{s}} + \frac{\sqrt{s}}{4} a_s \leq 1 \right\}.$$

As any point of C is a convex combination of its vertices and because the function f is convex, the maximum is attained at a vertex of C. Besides, the vertices are intersections of s hyperplanes when we force s of the $s+1$ inequalities to become equality. From this we obtain following possibilities:

1. If $a_1 = \cdots = a_s = 0$, then $f(a_1, a_2, \ldots, a_s) = 0$.

2. If $(a_1 + \cdots + a_s)/\sqrt{s} + \frac{\sqrt{s}}{4} a_1 = 1$ and $a_1 = \cdots = a_k > a_{k+1} = \cdots = a_s = 0$ for some $1 \leq k \leq s-1$, then one has $a_1 = \cdots = a_k = 4\sqrt{s}/(4k+s)$, and consequently $f(a_1, a_2, \ldots, a_s) = 4\sqrt{ks}/(4k+s) \leq 1$ by the AM-GM inequality with the pair $(4k, s)$.

3. If $(a_1 + \cdots + a_s)/\sqrt{s} + \frac{\sqrt{s}}{4} a_1 = 1$ and $a_1 = \cdots = a_s > 0$ then one has $a_1 = \cdots = a_s = 4/(5\sqrt{s})$, and consequently $f(a_1, a_2, \ldots, a_s) = 4/5 + 1/5 = 1$.

Thus we have obtained

$$\max_{(a_1, \ldots, a_s) \in C} f(a_1, a_2, \ldots, a_s) = 1,$$

and this concludes the proof. □

Proof. (of Theorem 5.19): We need to prove $\mathrm{NSP}_{2,\rho,\tau}$, that is, we need to find constants $0 < \rho < 1$ and $\tau > 0$ such that, for any $v \in \mathbb{C}^N$ and any $S \subset [N]$ with $\#(S) = s$,

$$||v_S||_2 \leq \frac{\rho}{\sqrt{s}} ||v_{\overline{S}}||_1 + \tau ||Av||_2.$$

Again it is enough to consider an index set $S =: S_0$ of s largest absolute entries of v. The same construction of Theorem 5.17 works here, that is we partition the complement $\overline{S_0}$ of S_0 in $[N]$ as $\overline{S_0} = S_1 \cup S_2 \cup \ldots$ where

- S_1 is the index set of the s largest absolute entries of v in $\overline{S_0}$.
- S_2 is the index set of the s largest absolute entries of v in $\overline{S_0 \cup S_1}$.

etc. Now, in contrast to Theorem 5.17 where we interpreted the vector v_{S_0} as being $2s$-sparse, we think of it it as being s-sparse and then we can write

$$||Av_{S_0}||_2^2 = (1+t)||v_{S_0}||_2^2 \qquad \text{with } |t| \leq \delta_s.$$

The first estimate we need to establish is that, for any $k \geq 1$,

$$\left| \langle Av_{S_0}, Av_{S_k} \rangle \right| \leq \sqrt{\delta_{2s}^2 - t^2} ||v_{S_0}||_2 ||v_{S_k}||_2. \tag{5.13}$$

In order to do it, we will first normalize the vectors v_{S_0} and v_{S_k} by defining $u = v_{S_0}/||v_{S_0}||_2$ and $w = e^{i\theta}v_{S_k}/||v_{S_k}||_2$ with θ being chosen to give $|\langle Au, Aw\rangle| = \mathrm{Re}\langle Au, Aw\rangle$. For real numbers $\alpha, \beta \geq 0$ to be chosen later, we have

$$\begin{aligned}
2|\langle Au, Aw\rangle| &= \frac{1}{\alpha+\beta}\Big[||A(\alpha u + w)||_2^2 - ||A(\beta u - w)||_2^2 - (\alpha^2-\beta^2)||Au||_2^2\Big] \\
&\leq \frac{1}{\alpha+\beta}\Big[(1+\delta_{2s})||\alpha u + w||_2^2 - (1-\delta_{2s})||\beta u - w||_2^2 - (\alpha^2-\beta^2)(1+t)||u||_2^2\Big] \\
&= \frac{1}{\alpha+\beta}\Big[(1+\delta_{2s})(\alpha^2+1) - (1-\delta_{2s})(\beta^2+1) - (\alpha^2-\beta^2)(1+t)\Big] \\
&= \frac{1}{\alpha+\beta}\big[\alpha^2(\delta_{2s}-t) + \beta^2(\delta_{2s}+t) + 2\delta_{2s}\big].
\end{aligned}$$

Setting $\alpha = (\delta_{2s}+t)/\sqrt{\delta_{2s}^2-t^2}$ and $\beta = (\delta_{2s}-t)/\sqrt{\delta_{2s}^2-t^2}$ we can derive

$$2|\langle Au, Aw\rangle| \leq \frac{\sqrt{\delta_{2s}^2-t^2}}{2\delta_{2s}}\big[\delta_{2s}+t+\delta_{2s}-t+2\delta_{2s}\big] = 2\sqrt{\delta_{2s}^2-t^2},$$

which is the same as the equation (5.13) we wanted to prove. After this, one observes that

$$\begin{aligned}
||Av_{S_0}||_2^2 = \left\langle Av_{S_0}, A\Big(v - \sum_{k\geq 1} v_{S_k}\Big)\right\rangle &= \langle Av_{S_0}, Av\rangle - \sum_{k\geq 1}\langle Av_{S_0}, Av_{S_k}\rangle \\
&\leq ||Av_{S_0}||_2||Av||_2 + \sum_{k\geq 1}\sqrt{\delta_{2s}^2-t^2}||v_{S_0}||_2||v_{S_k}||_2 \\
&= ||v_{S_0}||_2\Big(\sqrt{1+t}||Av||_2 + \sqrt{\delta_{2s}^2-t^2}\sum_{k\geq 1}||v_{S_k}||_2\Big). \qquad (5.14)
\end{aligned}$$

For each $k \geq 1$, let us denote by v_k^- and v_k^+ the smallest and the largest absolute entries of v on S_k. Using Lemma 5.21, we obtain

$$\sum_{k\geq 1}||v_{S_k}||_2 \leq \sum_{k\geq 1}\left(\frac{1}{\sqrt{s}}||v_{S_k}||_1 + \frac{\sqrt{s}}{4}(v_k^+ - v_k^-)\right) \leq \frac{1}{\sqrt{s}}||v_{\overline{S_0}}||_1 + \frac{\sqrt{s}}{4}v_1^+ \leq \frac{1}{\sqrt{s}}||v_{\overline{S_0}}||_1 + \frac{1}{4}||v_{S_0}||_2.$$

Using this last inequality in the right-hand side of (5.14) while changing $||Av_{S_0}||_2^2$ by $(1+t)||v_{S_0}||_2^2$ on its remote left-hand side, and canceling one $||v_{S_0}||$ factor we conclude that

$$\begin{aligned}
(1+t)||v_{S_0}||_2 &\leq \sqrt{1+t}||Av||_2 + \frac{\sqrt{\delta_{2s}^2-t^2}}{\sqrt{s}}||v_{\overline{S_0}}||_1 + \frac{\sqrt{\delta_{2s}^2-t^2}}{4}||v_{S_0}||_2 \\
&\leq (1+t)\left(\frac{1}{\sqrt{1+t}}||Av||_2 + \frac{\delta_{2s}}{\sqrt{s}\sqrt{1-\delta 2s^2}}||v_{\overline{S_0}}||_1 + \frac{\delta_{2s}}{4\sqrt{1-\delta_{2s}^2}}||v_{S_0}||_2\right),
\end{aligned}$$

where we used $\sqrt{\delta_{2s}^2-t^2}/(1+t) \leq \delta_{2s}/\sqrt{1-\delta_{2s}}$. Then, dividing by $1+t$, using $1/\sqrt{1+t} \leq \sqrt{1-\delta_{2s}}$ and rearranging, the last expression leads to

$$||v_{S_0}||_2 \leq \frac{\delta_{2s}}{\sqrt{1-\delta_{2s}^2}-\delta_{2s}/4}\frac{||v_{\overline{S_0}}||_1}{\sqrt{s}} + \frac{\sqrt{1+\delta_{2s}}}{\sqrt{1-\delta_{2s}^2}-\delta_{2s}/4}||Av||_2.$$

This is exactly the $\mathrm{NSP}_{\rho,\tau}$ if

$$\rho := \frac{\delta_{2s}}{\sqrt{1-\delta_{2s}^2} - \delta_{2s}/4} < 1,$$

namely, if $5\delta_{2s}/4 < \sqrt{1-\delta_{2s}^2}$ or $\delta_{2s} < 4/\sqrt{41}$. Then, if this condition is satisfied, we have $\mathrm{NSP}_{\rho,\tau}$ and therefore stable and robust reconstruction of compressible vectors. □

We have related results for thresholding and greedy algorithms. All the proofs can be found in Sections 6.3 and 6.4 of [Rauhut & Foucart '13]. There are two standard thresholding algorithms: iterative hard threshold and hard threshold pursuit. In the second one, for example, we have a sequence defined inductively by

$$S^{n+1} = L_s(x^n + A^*(y - Ax^n)). \tag{HTP$_1$}$$

$$x^{n+1} = \mathrm{argmin}\{||y - Az||_2,\ \mathrm{supp}(z) \subset S^{n+1}\}, \tag{HTP$_2$}$$

where $L_s(z)$ denotes the index set of s largest absolute entries of a vector $z \in \mathbb{C}^N$. The success of hard thresholding pursuit is guaranteed by the following theorem[8]

Theorem 5.22. (Theorem 3.8 in [Foucart '11]): *Suppose that the $3s$-th restricted isometry constant of the matrix $A \in \mathbb{C}^{m\times N}$ satisfies*

$$\delta_{3s} < \frac{1}{\sqrt{3}} \approx 0.5773.$$

Then, for $x \in \mathbb{C}^N, e \in \mathbb{C}^m$, and $S \subset [N]$ with $\#(S) = s$, the sequence x^n defined by (HTP$_1$) *and* (HTP$_2$) *with $y = Ax + e$ satisfies, for any $n \geq 0$,*

$$||x^n - x_S||_2 \leq \rho^n ||x^0 - x_S||_2 + \tau ||Ax_{\overline{S}} + e||_2. \tag{5.15}$$

where $\rho = \sqrt{\delta_{3s}^2/(1-\delta_{2s}^2)} < 1$ and $\tau \leq 5.15/(1-\rho)$.

It is important to note what occurs with threshold algorithms in general. If we take the limit $n \to \infty$, iteration (5.15) yields $||x^\# - x_S||_2 \leq \tau ||Ax_{\overline{S}} + e||_2$ if $x^\# \in \mathbb{C}^N$ is the limit of the sequence x^n or at least one of its accumulation points. The proof of this Theorem does not guarantee the existence of such limit. However, the boundedness of $||x^n||$ is ensured by (5.15). Then we have, by the triangle inequality, that $||x - x^\#||_2 \leq ||x_{\overline{S}}||_2 + ||x_S - x^\#||_2$, so choosing S as an index set of the s largest absolute entries of x gives

$$||x - x^\#||_2 \leq \sigma_s(x)_2 + \tau ||Ax_{\overline{S}} + e||_2.$$

This is a very interesting estimate which differs from the one available for basis pursuit due to the fact that we have now $\sigma_s(x)_2$ instead of $\sigma_s(x)_1$. This is *exclusive* for threshold algorithms. But the classical error estimates as in the ℓ_1-minimization are also available. We just need to make the replacement $s \to 2s$, and then, instead of a hypothesis based on δ_{3s}, we have one based on δ_{6s}. This is described in the following result

Theorem 5.23. *Suppose that the $6s$-th order restricted isometry constant of the matrix $A \in \mathbb{C}^{m\times N}$ satisfies $\delta_{6s} < 1/\sqrt{3}$. Then, for all $x \in \mathbb{C}^N$ and $e \in \mathbb{C}^m$, the sequence x^n defined by* (HTP$_1$) *and* (HTP$_2$) *with $y = Ax + e$, $x^0 = 0$ and s replaced by $2s$ satisfies, for any $n \geq 0$,*

$$||x - x^n||_1 \leq C\sigma_s(x)_1 + D\sqrt{s}||e||_2 + 2\rho^n\sqrt{s}||x||_2,$$
$$||x - x^n||_2 \leq \frac{C}{\sqrt{s}}\sigma_s(x)_1 + D||e||_2 + 2\rho^n||x||_2,$$

[8] [Foucart '11] established results about the analysis for a family of thresholding algorithms indexed by an integer k. There, iterative hard thresholding and hard thresholding pursuit correspond to the cases $k = 0$ and $k = \infty$, respectively, as described in Section 2.3.

where the constants $C, D > 0$ and $0 < \rho < 1$ depend only on δ_{6s}. In particular, if the sequence x^n clusters around some $x^\# \in \mathbb{C}^N$, then

$$||x - x^\#||_1 \leq C\sigma_s(x)_1 + D\sqrt{s}||e||_2, \quad \textit{and} \quad ||x - x^\#||_2 \leq \frac{C}{\sqrt{s}}\sigma_s(x)_1 + D||e||_2.$$

For greedy algorithms, we will state results for the orthogonal matching pursuit despite the existence of results for other algorithms, such as CoSaMP. See, for example, the original paper [Needell & Tropp '08], where stability and robustness were stated under the condition $\delta_{4s} \leq 0.1$, or the improved Theorem 6.27 of [Rauhut & Foucart '13], where the condition $\delta_{4s} \leq 0.17157$ does the job. For OMP, a curious phenomenon occurs. It was noticed in Theorem 3.2 of [Mo & Shen '12] that standard RIP arguments are not enough to establish the recovery of all s-sparse vectors in *at most* s iterations. Here we present a modified version which appears in [Rauhut & Foucart '13] but follows the same ideas. We take a fixed $1 \leq \eta < \sqrt{s}$ and consider the $(s+1) \times (s+1)$ matrix with ℓ_2-normalized columns defined by

$$A = \left[\begin{array}{ccc|c} & & & \frac{\eta}{s} \\ & \mathrm{Id} & & \vdots \\ & & & \frac{\eta}{s} \\ \hline 0 & \dots & 0 & \sqrt{\frac{s-\eta^2}{s}} \end{array}\right].$$

In order to estimate RIP, we need to compute

$$A^*A - \mathrm{Id} = \left[\begin{array}{ccc|c} & & & \frac{\eta}{s} \\ & \mathbf{0} & & \vdots \\ & & & \frac{\eta}{s} \\ \hline \frac{\eta}{s} & \dots & \frac{\eta}{s} & 0 \end{array}\right].$$

This matrix has eigenvalues $-\eta/\sqrt{s}$, $\eta/\sqrt{s}$ and 0 with multiplicity $s-1$. Then,

$$\delta_{s+1} = ||A^*A - \mathrm{Id}||_{2\to 2} = \eta/\sqrt{s}.$$

However, consider the s-sparse vector $x = [1, \dots, 1, 0]$. It is not recovered from $y = Ax$ after s iterations. To see why, just note that the $s+1$ index is wrongly picked at the first iteration. Indeed

$$A^*(y - Ax^0) = A^*Ax = \left[\begin{array}{ccc|c} & & & \frac{\eta}{s} \\ & \mathrm{Id} & & \vdots \\ & & & \frac{\eta}{s} \\ \hline \frac{\eta}{s} & \dots & \frac{\eta}{s} & 1 \end{array}\right] \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ 0 \end{array}\right] = \left[\begin{array}{c} 1 \\ \vdots \\ 1 \\ \eta \end{array}\right].$$

Since OMP chooses one index in each iteration, we conclude that OMP fails to recover this specific vector in s iterations for this given matrix. This is what the naive greedy algorithm does. As [Rauhut & Foucart '13] points out, there are two ways for circumvent this. Perform more than s iterations or find a way to reject the wrong indices by modifying the OMP. However in both cases the sparse recovery can be established if the restricted isometry constant is sufficiently small.

Now, in order to state the result about OMP convergence, we consider a more general greedy algorithm which starts with an index set S^0, $x^0 = \mathrm{argmin}\{||y - Az||,\ \mathrm{supp}(z) \subset S^0\}$ and iterating scheme

$$S^{n+1} = S^n \cup L_1(A^*(y - Ax^n)) \tag{OMP$_1$}$$

$$x^{n+1} = \mathrm{argmin}\{||y - Az||_2,\ \mathrm{supp}(z) \subset S^{n+1}\}. \tag{OMP$_2$}$$

Note that the usual OMP algorithm corresponds to the choice $S^0 = \emptyset$ and $x^0 = 0$. The following theorem was first established in [Zhang '11] but restated and generalized in [Rauhut & Foucart '13].

Theorem 5.24. (Theorem 6.25 in [Rauhut & Foucart '13]): *Suppose that $A \in \mathbb{C}^{m\times N}$ has restricted isometry constant*

$$\delta_{13s} < \frac{1}{6}.$$

Then there is a constant $C > 0$ depending only on δ_{13s}, such that, for all $x \in \mathbb{C}^N$ and $e \in \mathbb{C}^m$, the sequence x^n defined by (OMP_1) *and* (OMP_2) *with $y = Ax + e$ satisfies*

$$||y - Ax^{12s}||_2 \leq C||Ax_{\overline{S}} + e||_2,$$

for any $S \subset [N]$ with $\#S = s$. Furthermore, if $\delta_{26s} < 1/6$, then there are constants $C, D > 0$ depending only on δ_{26s}, such that, for all $x \in \mathbb{C}^N$ and $e \in \mathbb{C}^m$, the sequence x^n defined by (OMP_1) *and* (OMP_2) *with $y = Ax + e$ satisfies*

$$||y - Ax^{12s}||_2 \leq C||Ax_{\overline{S}} + e||_2$$

for any $S \subset [N]$ with $\#S = s$. Furthermore, if $\delta_{26s} < 1/6$, then there are constants $C, D > 0$ depending only on $\delta_{26s} < 1/6$, such that, for all $x \in \mathbb{C}^N$ and $e \in \mathbb{C}^m$, the sequence x^n defined by (OMP_1) *and* (OMP_2) *with $y = Ax + e$ satisfies, for any $1 \leq p \leq 2$*

$$||x - x^{24s}||_p \leq \frac{c}{s^{1-1/p}}\sigma_s(x)_1 + Ds^{1/p-1/2}||e||_2.$$

Nowadays there are many variations on greedy algorithm for sparse recovery. We could also cite stagewise OMP, regularized OMP and CoSaMP. A comparison between them can be found in the survey [Blanchard & Tanner '15].

5.6 RIP Limitations

We saw that with RIP we could guarantee the convergence of all the main algorithms used for sparse recovery. We just need to ensure the existence of matrices with small RIP constant in the right regime of sparsity. Even more, RIP gives better scale measurements than coherence. However, the restricted isometry property has some limitations.

For example, we could ask if it is possible to increase the bound on δ_s (or δ_{ks} for some k), making $\delta_s \to 1$ and continue to ensure the recovery of sparse vectors for matrices with RIP constant sufficiently close to 1. This would be the ideal situation. If that was true, taking any matrix A, the linear systems $Ax = b$ it would be generically solvable, i.e., would be possible to find sparse solution for this system.

Unfortunately this is not the case. We have two counterexamples, one for δ_s and another for δ_{2s}, that show there is little hope to improve RIP constant indefinitely.

Theorem 5.25. (Theorem 4.1 in [Cai, Wang & Xu II '10]): *Let s be a positive integer. Then there exists a $(2s-1)\times 2s$ matrix A with restricted isometry constant $\delta_s = (s-1)/(2s-1)$ and two nonzero s-sparse vectors β_1 and β_2 with disjoint supports such that $A\beta_1 = A\beta_2$. As a corollary, there exists a matrix with $\delta_s < 1/2$ such that it is impossible to recover sparse vectors.*

While this first counterexample is simple and direct, the second one, which is based on δ_{2s}, is a little more sophisticated. We first need the following definition

Definition 5.26. Given a matrix $\Phi \in \mathbb{R}^{m\times N}$ with unit spectral norm $||\Phi||_{2\to 2} = 1$, we define the asymmetric RIC σ_k^2 as

$$\sigma_k^2(\Phi) = \min_{\substack{x_\Omega \\ \#\Omega \leq k}} \frac{||\Phi x_\Omega||_2^2}{||x_\Omega||_2^2}$$

Remark 33. *As the maximum of a singular value of any squared submatrix is bounded by 1, a matrix Φ with unit spectral norm with given σ_k^2 implies the existence of a rescaled matrix $A_k = (2/(1+\sigma_k^2))^{1/2}\Phi$ with restricted isometry constant satisfying*

$$\delta_k(A_k) \leq \frac{1-\sigma_k^2(\Phi)}{1+\sigma_k^2(\Phi)}. \tag{5.16}$$

Theorem 5.27. (Theorem 3 of [Davies & Gribonval '09]): *Consider $0 < p \leq 1$ and let $0 < \eta_p < 1$ be the unique positive solution to $\eta_p^{2/p} + 1 = \frac{2}{p}(1-\eta_p)$. Then*

1. *If $\Phi \in \mathbb{R}^{m\times N}$ is a unit spectral matrix, $2s \leq m < N$ and*

$$\sigma_{2s}^2(\Phi) > 1 - \frac{2}{2-p}\eta_p, \tag{5.17}$$

   then all s-sparse vectors can be uniquely recovered by basis pursuit

2. *For every $\varepsilon > 0$ there exists integers $s \geq 1$, $N \geq 2s+1$ and a matrix $\Phi \in \mathbb{R}^{(n-1)\times N}$ with*

$$\sigma_{2s}^2(\Phi) > 1 - \frac{2}{2-p}\eta_p - \varepsilon$$

   for which there exists a s-sparse vector which cannot be uniquely recovered via basis pursuit.

For $p = 1$ we have $\eta_1^2 + 2\eta_1 - 1 = 0$, hence $\eta_1 = \sqrt{2} - 1$ and the right-hand side in (5.17) is $3 - 2\sqrt{2}$. In terms of the standard RIC for the rescaled matrix A_k, with $k = 2s$, this means that, using (5.16), that for any $\varepsilon > 0$, there exists a matrix A with $\delta_{2s} < 1/\sqrt{2} + \varepsilon$ where ℓ_1-recovery can fail. Therefore there is no hope in the improvement of RIP constant beyond $1/\sqrt{2} \approx 0.707$. What remains is to close the gap. In the case of δ_s, from 1/3 to 1/2, and in the case of δ_{2s}, from 0.6246 to 0.707.

Up until now we have said nothing about the difficulty in the combinatorial nature of the calculation of RIP. Evaluating RIP, namely, computing the constant δ_s for some matrix A and a level of sparsity s is a difficult problem. The intractability comes from the fact that any brute-force method would have to look at all submatrices with column subsets of size up to s. Despite a lot of computational evidence, this was an *open problem in complexity theory until the end of 2013.* It was solved by [Tillmann & Pfetsch '14].

Theorem 5.28. *Given a matrix $A \in \mathbb{Q}^{m\times N}$ and a positive integer s, the problem to decide whether there exists some rational constant $\delta_s < 1$ such that A satisfies RIP of order s with constant δ_s is coNP-complete.*

Corollary 5.29. *For a given matrix $A \in \mathbb{Q}^{m\times N}$ and a positive integer s, it is NP-hard to compute the restricted isometry constant δ_s.*

It is important to note that the result above is for the case $\delta = 1$, i.e., it does not imply the result for every *fixed* constant $\delta < 1$.

Also, we can ask whether a matrix A satisfies RIP with given order s and *given* constant $\delta_s \in (0,1)$. This is known as the *RIP certification problem.* It was also solved in [Tillmann & Pfetsch '14], although [Bandeira, Dobriban, Mixon & Sawin '13] derived the result independently. The result is a little weaker than the previous one.

Theorem 5.30. *Given a matrix $A \in \mathbb{Q}^{m\times N}$, a positive integer s and some constant $\delta_s \in (0,1)$, it is coNP-hard to decide whether A satisfies the RIP of order s with constant δ_s.*

It remains to be shown that the problem is also in the coNP class of complexity. Remark 5 in [Tillmann & Pfetsch '14] explicitly asks:

Open Problem: Prove that the *RIP certification problem* is coNP-complete.

This kind of complexity results open a branch of investigation on approximation algorithms to compute bounds on δ_s instead of searching for exact polynomial time algorithms.

We saw, after Theorem 3.3, that sparse recovery is guaranteed if we take the measurements matrix and rescale, reshuffle or add some new measurement. This will not change the NSP property, but these operations will corrupt the RIP property.

Reshuffling a measurement matrix is the same as multiplying it by a permutation matrix, but $\delta_s(UA) = \delta_s(A)$ for any unitary matrix U, so in this case the RIP remains unchanged. However, if we rescale the measurements, replacing the measurement matrix $A \in \mathbb{C}^{m\times N}$ by DA, with $D \in \mathbb{C}^{m\times m}$ as a diagonal matrix, we can *increase* the restricted isometry constant. Even worse, we may have trouble with the scalar rescaling, where we replace the matrix A by λA for some $\lambda \in \mathbb{C}$. For a very simple example of this deterioration, suppose that we have a matrix A with $\delta_S(A) < 3/5$. Recalling that $\delta_s = \max_{S\subset[N],\ \#S\leq s} ||A_S^*A_S - \mathrm{Id}||_{2\to 2}$, we can estimate the restricted isometry constant of $2A$ as

$$\delta_s(2A) = \max_{S\subset[N],\ \#S\leq s} ||(2A_S)^*(2A_S) - \mathrm{Id}||_{2\to 2} \geq ||(2A_S)^*(2A_S) - \mathrm{Id}||_{2\to 2} = ||4A_S^*A_S - 4\mathrm{Id} + 3\mathrm{Id}||_{2\to 2}$$

$$= ||4(A_S^*A_S - \mathrm{Id}) + 3\mathrm{Id}||_{2\to 2} = \max_{||x||_2=1} \Big\langle x, \Big(3\mathrm{Id} + 4(A_S^*A_S - \mathrm{Id})\Big)x\Big\rangle = 3 + 4||A_S^*A_S - \mathrm{Id}||_{2\to 2} \geq 3 - 4\delta_s(A)$$

$$> \delta_s(A).$$

In case we add some new measurements, and therefore more information, the RIP could also be perturbed. Consider a matrix with $\delta_s(A) < 1$ and let $\delta > \delta_s(A)$. Let us append a new row $[0\ \ldots\ 0\ \sqrt{1+\delta}]$ and call the new matrix $\tilde{A}$. With the aid of the vector $x = [0\ \ldots\ 0\ 1]$ we conclude that $||Ax||_2^2 \geq 1 + \delta$. This implies $\delta_S(\tilde{A}) > \delta_s(A)$.

Therefore, RIP is a sufficient condition for sparse recovery but can be, sometimes, a poor sufficient condition. It does not capture some operations that can be done with the measurement matrix. Also, some works argue that it does not capture the essence of the structure of sparsity, see [Adcock, Hansen & Roman '15] and references therein.

Chapter 6

Interlude: Non-asymptotic Probability

"There is a tendency in teaching to reduce probability problems to pure analysis as soon as possible and to forget the specific characteristics of probability theory itself. Such treatments are based on a poorly defined notion of random variables usually introduced at the outset."
William Feller in An Introduction to Probability Theory and Its Applications Vol. 1

6.1 Introduction

The purpose of this chapter is to take an interlude on Compressive Sensing theory and introduce some modern concepts related to Probability theory. Probability theory is concerned with obtaining information from the data being generated by some well-known data generating process. For example, one could ask when (and if) the surname "Montgomery" will disappear from families in USA or UK. For this, we can compute the expected number of Montgomery's births if one has a model for birth and death process and a model for the choice of names.

The original purpose of Probability was to calculate odds in games of chance. It was the Theory of Chance. For example, the gambler's ruin was in several circles of intellectual discussions during the Enlightenment. Most of the modern ideas in Probability arose at that time. Some historical account of Probability can be found at [Stigler '90], [Hacking '06], [Fischer '11] and [von Plato '98].

As Feller points out in his legendary book [Feller '68], in the forties *"few mathematicians outside the Soviet Union recognized probability as a legitimate branch of mathematics. Applications were limited in scope, and the treatment of individual problems often led to incredible complications"*. In fact, a tipping point happened when Kolmogorov published his seminal work axiomatizing Probability. He and his collaborators/students Khinchin, Gnedenko, Prokhorov, Dynkin, Shiryaev, etc have had a profound influence on making Probability a prestigious and intense field of research. Also, we should mention Levy, von Mises, Cramer, von Neumann, Jeffreys, Savage, de Finetti, Doob and Feller among many others as being part of this "revolution".

Nowadays probability is a very important branch of mathematics with lots of ramifications and applications to the real world. It pervades all the sciences and thoughts about uncertainty are fundamental in the modeling of any phenomena. Close to the subject of this dissertation, the techniques developed in Machine Learning and Statistical Signal Processing confirm this fact.

Through this chapter we will adopt the non-asymptotic point of view, generalize the Bernoulli and Gaussian distributions through the subgaussian distribution and state some important inequalities such as Gordon's Lemma and the concentration of Gaussian measure for Lipschitz functions. All of this will be done in order to use some probabilistic tools in Compressive Sensing. Then we will prove the suitability of random matrices in the optimal regime, as discussed in the Chapter 5, for sparse vector recovery.

The reader should have in mind that many of the major breakthroughs in Compressive Sensing, and also in many of modern Data Science subareas, relies on probabilistic arguments. Hence we spend some time collecting and demonstrating tools which will help us to understand the rest of this dissertation.

We will not focus on basic concepts of Probability. Instead, the reader should consult [Resnick '13] or [Durrett '10] for essentials from Probability. The book [Feller '68] should always be consulted.

6.2 Subgaussian and Subexponential Random Variables

"Many years ago (in 1893) I called the Laplace-Gaussian curve the normal curve, which name, while it avoids an international question of priority, has the disadvantage of leading people to believe that all other distributions of frequency are in one sense or another "abnormal".
Karl Pearson in [Pearson '20]

The most renowned probability distribution is the Gaussian distribution. The richness of its properties, on one side, and his simplicity on the other side turn it ubiquitous. From the Bean Machine developed by Francis Galton to the Central Limit Theorem, passing by the maximization of entropy and least squares, science changed since its introduction. An account of its ubiquitous use can be seen at [Kim & Shevlyakov '08] and references therein.

Definition 6.1. A *Gaussian random variable* or *normal random variable* has probability density function

$$\psi(t) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(t-\mu)^2/(2\sigma^2)}.$$

It has mean $\mathbb{E}X = \mu$ and variance $\mathbb{E}(X-\mu)^2 = \sigma^2$. A *standard Gaussian variable* g is a Gaussian random variable satisfying $\mathbb{E}g = 0$ and $\mathbb{E}g^2 = 1$.

Proposition 6.2. *Let g be a standard Gaussian random variable. Then, for all $u > 0$,*

$$\mathbb{P}(|g| \geq u) \leq \min\left\{1, \sqrt{\frac{2}{\pi}}\frac{1}{u}\right\} \exp(-u^2/2). \tag{6.1}$$

$$\mathbb{P}(|g| \geq u) \geq \max\left\{\sqrt{\frac{2}{\pi}}\frac{1}{u}\left(1-\frac{1}{u^2}\right), \left(1-\sqrt{\frac{2}{\pi}}u\right)\right\} \exp(-u^2/2). \tag{6.2}$$

Proof. Using the definition of the probability density function of the Gaussian distribution, we have $\mathbb{P}(|g| \geq u) = \sqrt{\frac{2}{2\pi}} \int_u^\infty e^{-t^2/2} dt$. Then, a change of variables yields

$$\int_u^\infty e^{-t^2/2} dt = \int_0^\infty e^{-(t+u)^2/2} dt = e^{-u^2/2} \int_0^\infty e^{-tu} e^{-t^2/2} dt. \tag{6.3}$$

For the upper bound, on the one hand, we can use that $e^{-tu} \leq 1$ for $t, u \geq 0$. Hence

$$\int_u^\infty e^{-t^2/2} dt \leq e^{-u^2/2} \int_0^\infty e^{-t^2/2} dt = \sqrt{\frac{\pi}{2}} e^{-u^2/2}.$$

On the other hand, we can use that $e^{-t^2/2} \leq 1$ and this leads to

$$\int_u^\infty e^{-t^2/2} dt \leq e^{-u^2/2} \int_0^\infty e^{-tu} dt = \frac{1}{u} e^{-u^2/2}.$$

For the lower bound, we use the estimates $e^{-t^2/2} \geq 1 - t^2/2$ and $e^{-tu} \geq 1 - tu$ in Equation (6.3) and this yields, respectively,

$$\int_u^\infty e^{-t^2/2} dt \geq e^{-u^2/2} \int_0^\infty \left(1 - \frac{t^2}{2}\right) e^{-tu} dt = e^{-u^2/2} \left(\frac{1}{u} - \frac{1}{u^3}\right),$$

and

$$\int_u^\infty e^{-t^2/2} dt \geq e^{-u^2/2} \int_0^\infty e^{-t^2/2}(1-ut) dt = e^{-u^2/2} \left(\sqrt{\frac{\pi}{2}} - u\right).$$

□

Besides tail estimates, another useful fact is the expression for the moment-generating function of the Gaussian random variable.

Proposition 6.3. *Let X be a Gaussian random variable. Then, for $\theta \in \mathbb{R}$, we have $\mathbb{E}(e^{\theta X}) = e^{\theta^2/2}$ and, more generally, for $\theta \in \mathbb{R}$ and $a < 1/2$,*

$$\mathbb{E}(e^{aX^2+\theta X}) = \frac{1}{\sqrt{1-2a}} \exp\left(\frac{\theta^2}{2(1-2a)}\right).$$

As a consequence, the moments of the Gaussian distribution are given by $\mathbb{E}X^{2n+1} = 0$ and $\mathbb{E}X^{2n} = (2n)!/(2^n n!)$.

Proof. See Section 5.6 of [DeGroot & Schervish '11]. □

Inspired by the tails and all of the remarkable properties of Gaussian distribution, one might try to find other distributions with similar tails (and, consequently, some similar properties). This leads to the concept of subgaussian random variables.

Definition 6.4. A random variable X is called *subgaussian*[1] if there exists constants $\beta, \kappa > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t^2} \qquad \text{for all } t > 0.$$

By Proposition 6.2, a standard Gaussian random variable is subgaussian with $\beta = 1$ and $\kappa = 1/2$. Other examples of subgaussian random variables include Rademacher random variables, with distribution $\mathbb{P}\{X = -1\} = \mathbb{P}\{X = 1\} = 1/2$, and bounded random variables, which satisfy $|X| < M$ almost surely for some M.

Despite the fact that the class of subgaussian variables is quite wide, it does not encompass some important distributions which have heavier tails than the Gaussian distribution. An example is the exponential distribution, which satisfies $\mathbb{P}(|X| \geq t) \leq e^{-t}$ for all $t > 0$. Because of this we need to define the class of subexponential random variables.

Definition 6.5. A random variable X is called *subexponential* if there exists constants $\beta, \kappa > 0$ such that

$$\mathbb{P}(|X| \geq t) \leq \beta e^{-\kappa t} \qquad \text{for all } t > 0.$$

It follows immediately from this definition that that any subgaussian variable is also subexponential. One can ask if the converse is also true, that is, if any subexponential variable is also subgaussian. After developing some equivalent conditions for a random variable being subgaussian, we will see a counterexample.

This class of subgaussian random variables was introduced by [Kahane '60] in order to establish a sufficient condition for the almost sure uniform convergence of certain random series of functions. Nowadays these variables have been proved to be fundamental not only for the study of series but also in the Geometry of Banach Spaces and for Random Matrices. They shown to be very important in the context of Compressive Sensing, as we will see in Chapter 7. More information and discussion about them can be found at [Buldygin & Kozachenko '98].

Now we proceed to the first equivalence. It relies on a general relations between moments and tails of a random variable. We will see two other ways to characterize them.

Theorem 6.6. *Suppose that X is a random variable satisfying*

$$(\mathbb{E}|X|^p)^{1/p} \leq \alpha \beta^{1/p} p^{1/\gamma} \qquad \textit{for all } p \in [p_0, p_1],$$

for some constants $\alpha, \beta, \gamma, p_1 > 0, p_0 > 0$. Then

$$\mathbb{P}(|X| \geq e^{1/\gamma} \alpha u) \leq \beta e^{-u^\gamma/\gamma},$$

[1] Subgaussian is the free translation of the French *sous-gaussienne* coined in the work [Kahane '60].

for all $u \in [p_0^{1/\gamma}, p_1^{1/\gamma}]$. *Conversely, suppose that a random variable* X *satisfies, for some* $\gamma > 0$,

$$\mathbb{P}(|X| \geq e^{1/\gamma}\alpha u) \leq \beta e^{-u^\gamma/\gamma} \qquad \text{for all } u > 0,$$

then, for $p > 0$,

$$\mathbb{E}|X|^p \leq \beta\alpha^p(e\gamma)^{p/\gamma}\Gamma\left(\frac{p}{\gamma}+1\right). \tag{6.4}$$

As a consequence, for $p \geq 1$,

$$(\mathbb{E}|X|^p)^{1/p} \leq C_1\alpha(C_{2,\gamma}\beta)^{1/p}p^{1/\gamma} \qquad \text{for all } p \geq 1, \tag{6.5}$$

where $C_1 = e^{1/(2e)} \approx 1.2019$ *and* $C_{2,\gamma} = \sqrt{2\pi/\gamma e^{\gamma/12}}$. *In particular,* $C_{2,1} \approx 2.7245$ *and* $C_{2,2} \approx 2.0939$.

Proof. For the first part, for an arbitrary $\kappa > 0$, we use Markov's inequalityfor $|X|^p$ (for some p to be fixed later) and obtain

$$\mathbb{P}(|X| \geq e^\kappa\alpha u) \leq \frac{\mathbb{E}|X|^p}{(e^\kappa\alpha u)^p} \leq \beta\left(\frac{\alpha p^{1/\gamma}}{e^\kappa\alpha u}\right)^p.$$

The choice $p = u^\gamma$ and $\kappa = 1/\gamma$ yields the result. For the converse part, we have

$$\begin{aligned}
\mathbb{E}|X|^p &= \int_\Omega |X|^p d\mathbb{P} = \int_\Omega\int_0^{|X|^p} 1\, dx d\mathbb{P} = \int_\Omega\int_0^\infty I_{\{|X|^p\geq x\}}dxd\mathbb{P} = \int_0^\infty\int_\Omega I_{\{|X|^p\geq x\}}d\mathbb{P}dx \\
&= \int_0^\infty \mathbb{P}(|X|^p \geq x)dx = p\int_0^\infty \mathbb{P}(|X|^p \geq t^p)t^{p-1}dt = p\int_0^\infty \mathbb{P}(|X| \geq t)t^{p-1}dt \\
&= p\alpha^p e^{p/\gamma}\int_0^\infty \mathbb{P}(|X| \geq e^{1/\gamma}\alpha u)u^{p-1}du \leq p\alpha^p e^{p/\gamma}\int_0^\infty \beta e^{-u^\gamma/\gamma}u^{p-1}du \\
&= p\beta\alpha^p e^{p/\gamma}\int_0^\infty e^{-v}(\gamma v)^{p/\gamma-1}dv = \beta\alpha^p(e\gamma)^{p/\gamma}\frac{p}{\gamma}\Gamma\left(\frac{p}{\gamma}\right) = \beta\alpha^p(e\gamma)^{p/\gamma}\Gamma\left(\frac{p}{\gamma}+1\right).
\end{aligned}$$

Now, in order to prove (6.5), we just use Stirling's formula for the Gamma function. It states that, for $x > 0$, $\Gamma(x) = \sqrt{2\pi}x^{x-1/2}e^{-x}\exp(\theta(x)/12x)$ with $0 \leq \theta(x) \leq 1$ (the proof of this fact can be found in many references, e.g., [Jameson '15][2]). Applying this formula to the Gamma function in the equation above yields

$$\mathbb{E}|X|^p \leq \beta\alpha^p(e\gamma)^{p/\gamma}\sqrt{2\pi}\left(\frac{p}{\gamma}\right)^{p/\gamma+1/2}e^{-p/\gamma}e^{\gamma/(12p)} = \sqrt{2\pi}\beta\alpha^p e^{\gamma/(12p)}p^{p/\gamma+1/2}\gamma^{-1/2}.$$

Under the assumption $p \geq 1$, we obtain

$$(\mathbb{E}|X|^p)^{1/p} \leq \left(\frac{\sqrt{2\pi}e^{\gamma/12}}{\sqrt{\gamma}}\beta\right)^{1/p}\alpha p^{1/\gamma}p^{1/(2p)}.$$

Lastly, the maximum value of $p^{1/(2p)}$ is attained for $p = e$ and so $p^{1/(2p)} \leq e^{1/(2e)}$. This concludes the proof of Theorem 6.6. □

[2] For historical references about Stirling's formula, one can look at [Fowler '00] and [Tweddle '84]. Also, sixteen variations on the theme can be found at [Dominici '08].

Particularizing Theorem 6.6 for Subgaussian variables with $\alpha = (2e\kappa)^{-1/2}$ and $\gamma = 2$ shows that its moments satisfy $(\mathbb{E}|X|^p)^{1/p} \leq \tilde{C}\kappa^{-1/2}\beta^{1/p}p^{1/2} = O(\sqrt{p})$ for all $p \geq 1$. Also, for a Subexponential variable X, setting $\alpha = (e\kappa)^{-1}$ and $\gamma = 1$, Theorem 6.6 leads to $(\mathbb{E}|X|^p)^{1/p} \leq \overline{C}\kappa^{-1}\beta^{1/p}p = O(p)$ for all $p \geq 1$.

We have another equivalence for Subgaussian variables. This is sometimes called *super-exponential moment* equivalence.

Proposition 6.7. *A random variable X is Subgaussian, i.e., $\mathbb{P}(|X| \geq t) \leq Ce^{-ct^2}$, if and only if there exist constants $c > 0$ and $C \geq 1$ such that $\mathbb{E}[\exp(cX^2)] \leq C$.*

Proof. First, let us proof the *only if* part. Using Theorem 6.6, (more precisely, the moment estimate (6.4)) with $\kappa = 1/(2e\alpha^2)$, we have $\mathbb{E}X^{2n} \leq \beta\kappa^{-n}n!$. Using the Taylor series of the exponential function and Fubini's theorem leads to

$$\mathbb{E}[\exp(cX^2)] = 1 + \sum_{n=1}^{\infty} \frac{c^n\mathbb{E}[X^{2n}]}{n!} \leq 1 + \beta\sum_{n=1}^{\infty} \frac{c^n\kappa^{-n}n!}{n!} = 1 + \frac{\beta c\kappa^{-1}}{1 - c\kappa{-1}},$$

provided $c < \kappa$. Now, in order to prove the *if* part, we just use Markov's inequality. It leads to

$$\mathbb{P}(|X| \geq t) = \mathbb{P}\big(\exp(cX^2) \geq \exp(ct^2)\big) \leq \mathbb{E}[\exp(cX^2)]e^{-ct^2} \leq Ce^{-ct^2}.$$

□

The study of deviation inequalities for the tail bounds of a random variable can be done through the moment-generating function. This is the content Cramér-Chernoff theorem below. Therefore, an equivalent definition for subgaussian variables will be given via moment-generating function. We begin with a definition.

Definition 6.8. The *cumulant-generating* function of a real-valued random variable is defined as the logarithm of the moment-generating function, that is,

$$C_X(\theta) = \ln \mathbb{E}\exp(\theta X).$$

We now present the general deviation bound.

Theorem 6.9. *Let $X_1, \ldots, X_M$ be a sequence of independent (real-valued) random variables with cumulant-generating functions C_{X_i}, $i \in [M]$. Then, for $t > 0$,*

$$\mathbb{P}\Big(\sum_{i=1}^{M} X_i \geq t\Big) \leq \exp\left(\inf_{\theta>0}\Big\{-\theta t + \sum_{i=1}^{M} C_{X_i}(\theta)\Big\}\right).$$

Proof. For $\theta > 0$, Markov's inequality and independence yield

$$\mathbb{P}\left(\sum_{i=1}^{M} X_i \geq t\right) = \mathbb{P}\left(\exp\left(\theta\sum_{i=1}^{M} X_i\right) \geq \exp(t\theta)\right) \leq e^{-\theta t}\mathbb{E}\left[\exp(\theta\sum_{i=1}^{M} X_i)\right] = e^{-\theta t}\mathbb{E}\left[\prod_{i=1}^{M}\exp(\theta X_i)\right]$$

$$= e^{-\theta t}\prod_{i=1}^{M}\mathbb{E}[\exp(\theta X_i)] = e^{-\theta t}\prod_{i=1}^{M}\exp(C_{X_i}(\theta)) = \exp\left(-\theta t + \sum_{i=1}^{M} C_{X_i}(\theta)\right).$$

Taking the infimum over $\theta > 0$ concludes the proof. □

Thus, another important equivalent definition of a (zero mean) Subgaussian variable is given in terms of the moment-generating function as the following theorem shows.

Theorem 6.10. *Let X be a random variable.*

i.) If X is Subgaussian with $\mathbb{E}X = 0$, then there exists a constant c (depending only on β and κ) such that

$$\mathbb{E}[\exp(\theta X)] \leq \exp(c\theta^2) \qquad \text{for all } \theta \in \mathbb{R}. \tag{6.6}$$

ii.) Conversely, if (6.6) holds, then $\mathbb{E}X = 0$ and X is Subgaussian with parameters $\beta = 2$ and $\kappa = 1/(4c)$.

Proof. First, we will proceed with the proof of ii. Taking $t, \theta > 0$ and applying Markov inequality yields

$$\mathbb{P}(X \geq t) = \mathbb{P}(\exp(\theta X) \geq \exp(\theta t)) \leq \mathbb{E}[\exp(\theta X)]e^{-\theta t} \leq e^{c\theta^2 - \theta t}.$$

Choosing the optimal parameter $\theta = t/(2c)$ on the right-hand side leads to $\mathbb{P}(X \geq t) \leq e^{-t^2/(4c)}$, and the same computation above with $-X$ instead of X shows that $\mathbb{P}(-X \geq t) \leq e^{-t^2/(4c)}$. So, the union bound yields what we want, that is,

$$\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/(4c)}.$$

Now, it remains to deduce that X has zero mean. Using that $1+\theta X \leq \exp(\theta X)$ and taking the expectation on both sides leads, for $|\theta| < 1$, to

$$1 + \theta\mathbb{E}(X) \leq \mathbb{E}[\exp(\theta X)] \leq \exp(c\theta^2) \leq 1 + (c/2)\theta^2 + O(\theta^4).$$

Taking $\theta \to 0$ yields $\mathbb{E}X \leq 0$. Making the same calculations with $-X$ instead of X shows that $\mathbb{E}X \geq 0$. Therefore, $\mathbb{E}X = 0$. Now, to prove the first part, it is enough to prove it for $\theta \geq 0$ because the theorem for $\theta < 0$ follows just by substituting X by $-X$. As in the proof of Proposition 6.7, using the Taylor series for the exponential function and Fubini's theorem yields

$$\mathbb{E}[\exp(\theta X)] = 1 + \theta\mathbb{E}(x) + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}X^n}{n!} = 1 + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}|X|^n}{n!},$$

since $\mathbb{E}(X) = 0$ by hypothesis. Let us consider, for a moment, that $0 \leq \theta \leq \theta_0$ for some θ_0 to be chosen later. Using the moment estimate $(\mathbb{E}|X|^p)^{1/p} \leq \tilde{C}\kappa^{-1/2}\beta^{1/p}p^{1/2}$ from Theorem 6.6, and Stirling's formula in its version $n! \geq \sqrt{2\pi}n^n e^{-n}$ leads to

$$\mathbb{E}[\exp(\theta X)] \leq 1 + \beta\sum_{n=2}^{\infty} \frac{\theta^n \tilde{C}^n \kappa^{-n/2} n^{n/2}}{n!} \leq 1 + \frac{\beta}{\sqrt{2\pi}} \sum_{n=2}^{\infty} \frac{\theta^n \tilde{C}^n \kappa^{-n/2} n^{n/2}}{n^n e^{-n}}$$

$$\leq 1 + \theta^2 \frac{\beta(\tilde{C}e)^2}{\sqrt{2\pi\kappa}} \sum_{n=0}^{\infty} (\tilde{C}e\theta_0\kappa^{-1/2})^n = 1 + \theta^2 \frac{\beta(\tilde{C}e)^2}{\sqrt{2\pi\kappa}} \frac{1}{1 - \tilde{C}e\theta_0\kappa^{-1/2}} = 1 + c_1\theta^2 \leq \exp(c_1\theta^2),$$

provided that $\tilde{C}e\theta_0\kappa^{-1/2} < 1$. Therefore, if we choose $\theta_0 = (2\tilde{C}e)^{-1}\sqrt{\kappa}$, we have the result for $c_1 = \sqrt{2}\beta\kappa^{-1}\big((\tilde{C}e)^2/\sqrt{\pi}\big)$.

It remains to prove the result when $\theta > \theta_0$. Actually, we will reformulate it and our goal will be to prove $\mathbb{E}[\exp(\theta X)] \leq \exp[(c_2\theta^2)]$ for some constant $c_2 > 0$. First, observe that:

$$\theta X - c_2\theta^2 = -\left(\sqrt{c_2}\theta - \frac{X}{2\sqrt{c_2}}\right)^2 + \frac{X^2}{4c_2} \leq \frac{X^2}{4c_2}.$$

Using the constant $c > 0$ and $C \geq 1$ from Proposition 6.7 and choosing $c_2 = 1/(4c)$ yields the following estimate

$$\mathbb{E}[\exp(\theta X - c_2\theta^2)] \leq \mathbb{E}\Big[\exp\big(X^2/(4c_2)\big)\Big] = \mathbb{E}\Big[\exp\big(cX^2\big)\Big] \leq C.$$

Finally, we need again to use θ_0 somehow. Defining $K = \ln(C)\theta_0^{-2}$ leads to

$$\mathbb{E}[\exp(\theta X)] \leq C\exp(c_2\theta^2) = C\exp(-K\theta^2)\exp\left((c_2+K)\theta^2\right) \leq C\exp(-K\theta_0^2)\exp((c_2+K)\theta^2)$$

$$\leq \exp\left((c_2+K)\theta^2\right).$$

We divided the proof in two cases, setting $c_3 = \max\{c_1, c_2+K\}$ completes the proof. □

This result tells us that the moment-generating function of a subgaussian variable exists for all$\theta \in \mathbb{R}$. As [Rudelson '14] points out, here we have a subtle difference from subexponential variables. For the latter, the bound for the moment-generating function only holds in a neighborhood of zero. This comes from the fact the moment-generating function of the exponential random variable with parameter 1 does not exist for $\theta \geq 1$.

Another interesting result is that the sum of zero mean subgaussian variables is also subgaussian as the next theorem shows. It is important to note that there is also a version of this theorem for subexponential variables, i.e, the subexponential property is also preserved under summation in the case of independent subexponential random variables.

Remark 34. *Equation 6.8 below is sometimes called Hoeffding inequality for subgaussian variables.*

Theorem 6.11. *Let $X_1, \ldots, X_M$ be a sequence of independent zero mean subgaussian random variables with subgaussian parameter c in* (6.6) *. For $a = (\alpha_1, \ldots, \alpha_M) \in \mathbb{R}^M$, the random variable $X = \sum_{i=1}^M \alpha_i X_i$ is Subgaussian, i. e.,*

$$\mathbb{E}\exp(\theta X) \leq \exp(c||a||_2^2\theta^2). \tag{6.7}$$

In particular, by Theorem 6.10,

$$\mathbb{P}\left(\left|\sum_{i=1}^M \alpha_i X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{4c||a||_2^2}\right) \qquad \text{for all } t > 0. \tag{6.8}$$

Proof. Using the independence of $X_1, \ldots, X_M$, we have

$$\mathbb{E}\exp\left(\theta\sum_{i=1}^M \alpha_i X_i\right) = \mathbb{E}\prod_{i=1}^M \exp(\theta\alpha_i X_i) = \prod_{i=1}^M \mathbb{E}\exp(\theta\alpha_i X_i) \leq \prod_{i=1}^M \exp(c\theta^2\alpha_i^2) = \exp(c||a||_2^2\theta^2).$$

This proves the first inequality. For the second, we just need to use part ii.) of Theorem 6.10. □

The discussion of Theorem 6.6, Proposition 6.7 and Theorem 6.10 tells us that we have four equivalent ways of defininining a subgaussian random variable. Using the tail of the distribution, the moments' bounds, super-exponential moment equivalence or, in the case of zero mean variables, the moment-generating function. Even more, the class of Subgaussian variables in a given probability space forms a normed space through the following definition.

Definition 6.12. The subgaussian norm of X, denoted by $||X||_{\psi_2}$ is defined by

$$||X||_{\psi_2} = \sup_{p\geq 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}.$$

In the discussion about a Subgaussian variable X, it is not necessary for X to have mean zero. This is done only to simplify the notation in proofs. Anyway, X always can be transformed into a zero-mean variables just by observing that if X is Subgaussian then so is $X - \mathbb{E}X$. Moreover, by the triangle inequality $||X - \mathbb{E}X||_{\psi_2} \leq ||X||_{\psi_2} + ||\mathbb{E}X||_{\psi_2}$ and the simple estimate $||\mathbb{E}X||_{\psi_2} = |\mathbb{E}X| \leq \mathbb{E}|X| \leq ||X||_{\psi_2}$, we have $||X - \mathbb{E}X||_{\psi_2} \leq 2||X||_{\psi_2}$. For the subexponential case, we have:

Definition 6.13. The subexponential norm of X, denoted by $||X||_{\psi_1}$ is defined by

$$||X||_{\psi_1} = \sup_{p\geq 1} p^{-1}(\mathbb{E}|X|^p)^{1/p}.$$

The normed space equipped with any of these two norms can be viewed as a particular cases of a Birnbaum-Orlicz space. Some details of this generalization can be found in Section 4 of [Rivasplata '12] or in the notes of Chapter 8 from [Rauhut & Foucart '13]. Finally, we can exemplify why Rademacher and bounded variables are subgaussian variables and give an example of a subexponential which is not subgaussian.

Example 6.14. (Rademacher variables): Let us compute the moment-generating function of Rademacher variables, that is, variables which satisfies $\mathbb{P}\{X=-1\} = \mathbb{P}\{X=1\} = 1/2$.

$$\mathbb{E}[e^{\theta X}] = \frac{1}{2}\Big(e^{-\theta} + e^{\theta}\Big) = \frac{1}{2}\Big(\sum_{k=0}^{\infty}\frac{(-\theta)^k}{k!} + \sum_{k=0}^{\infty}\frac{\theta^k}{k!}\Big) = \sum_{k=0}^{\infty}\frac{\theta^{2k}}{(2k)!} \leq 1 + \sum_{k=1}^{\infty}\frac{\theta^{2k}}{2^k k!} = e^{\theta^2/2}.$$

This tells us that Rademacher variables are subgaussian.

Example 6.15. (Bounded variables): Let X be mean zero random variable contained in some interval $[a,b]$. Let $\tilde{X}$ be an independent copy of X. Here we will use a technique called *symmetrization*, where after the introduction of this new variable, we use a symmetry argument with the aid of a Rademacher variable. A different argument can be found in Theorem 2.5 of [Rivasplata '12]. Now it will be necessary to denote by $\mathbb{E}_X$ the expected value with respect to the variable X.

$$\mathbb{E}_X[e^{\theta X}] = \mathbb{E}_X\left[\exp\Big(\theta\big(X - \mathbb{E}_{\tilde{X}}[\tilde{X}]\big)\Big)\right] = \int_a^b e^{\theta X}e^{-\theta\mathbb{E}_{\tilde{X}}[\tilde{X}]}p(x)dx \leq \int_a^b e^{\theta X}\mathbb{E}_{\tilde{X}}[e^{-\theta\tilde{X}}]p(x)dx =$$

$$\mathbb{E}_X[e^{\theta X}]\mathbb{E}_{\tilde{X}}[e^{-\theta\tilde{X}}] = \mathbb{E}_{X,\tilde{X}}\big[e^{\theta(X-\tilde{X})}\big].$$

where we used the convexity of the exponential and Jensen's inequality. Now, denoting by Z a Rademacher variable, we have that the distribution of $Z(X-\tilde{X})$ is the same as $X-\tilde{X}$. This yields

$$\mathbb{E}_{X,\tilde{X}}\big[e^{\theta(X-\tilde{X})}\big] = \mathbb{E}_{X,\tilde{X}}\big[\mathbb{E}_Z[e^{\theta Z(X-\tilde{X})}]\big] \leq \mathbb{E}_{X,\tilde{X}}\big[e^{\frac{\theta^2(X-\tilde{X})^2}{2}}\big] \leq e^{\frac{\theta^2(b-a)^2}{2}}.$$

There the penultimate inequality follows from applying the calculations of Example 6.14 conditionally, using that $(X,\tilde{X})$ is fixed, and the last inequality follows from $|X-\tilde{X}| \leq b-a$. This proves that bounded variables are subgaussian.

The main difference between subgaussian and subexponential variables can be described in terms of the moment-generating function. In particular, we can cite the following theorem that can be found in Lemma 5.15 from [Vershynin '12].

Theorem 6.16. *Let X be a mean zero subexponential random variable. Then, for t such that $|t| \leq c/||X||_{\psi_1}$, one has $\mathbb{E}\exp(tX) \leq \exp(Ct^2||X||_{\psi_1}^2)$, where $C, c > 0$ are absolute constants.*

While, by Theorem 6.10, the moment-generating function of subgaussian variables exists *for all* $\theta \in \mathbb{R}$, this is not the case for subexponential variables. Therefore, if we find a subexponential variable with a moment generating function that does not exists for all θ, we will have found an example of a subexponential variable which is not subgaussian. This is shown in the next example.

Example 6.17. Consider the random variable $Y = X^2$, where X has a standard normal distribution. This variables is called χ^2 *random variable*. Since its mean is 1, we calculate the moment-generating function of $Y-1$. Thus, for $\theta < 1/2$ we have

$$\mathbb{E}[e^{\theta(Y-1)}] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{\theta^2(z^2-1)}e^{-z^2/2}dz = \frac{e^{-\theta}}{\sqrt{1-2\theta}}.$$

For $\theta > 1/2$, the moment-generating function of $Y - 1$ does not exist. Therefore, the one for Y also does not exist. This shows that Y cannot be subgaussian. On the other side, we have

$$\frac{e^{-\theta}}{\sqrt{1-2\theta}} \leq e^{2\theta^2} = e^{4\theta^2/2}.$$

for $|\theta| < 1/4$. Thus, this shows that $Y = X^2$ is subexponential.

6.3 Nonasymptotic Inequalities

"Yet, moving beyond this terra firma, one quickly encounters examples where classical methods are brittle".
Joel Tropp in [Tropp '15]

In the first part of the last century, researches of Probability Theory were mainly concerned with the asymptotic behavior of random variables. See [Feller '45]. After many efforts, the two most important theorem of Probability, The Law of Large Numbers and the Central Limit Theorem, got a definitive shape. Here we state them not in the general but instead, in a pleasant form.

Theorem 6.18. (Law of Large Numbers): *Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables with mean μ. Consider the sum $S_n = X_1 + \cdots + X_n$. Then, as $n \to \infty$,*

$$\frac{S_n}{n} \to \mu \text{ almost surely.}$$

Theorem 6.19. (Central Limit Theorem): *Let $X_1, X_2, \ldots$ be a sequence of independent, identically distributed random variables with mean μ and variance σ^2. Consider the sum $S_n = X_1 + \cdots + X_n$ and normalize it to obtain a random variable with zero mean and unit variance as follows*

$$Z_n = \frac{S_n - \mathbb{E}[S_n]}{\sqrt{Var(S_n)}} = \frac{1}{\sigma\sqrt{N}} \sum_{i=1}^{n} (X_i - \mu).$$

We denote the normal distribution by $g \sim N(0,1)$. Then, as $n \to \infty$,

$$\mathbb{P}(Z_n \geq t) \to \mathbb{P}(g \geq t) = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx, \qquad \forall t \in \mathbb{R}.$$

For a historical account, see [Seneta '13] and [Fischer '11]. As one can note there is no mention to the rate of convergence. Thus, the first question is how to give a bound on the maximal error of the approximation between the normal distribution and the true distribution of the scaled sample mean. In the case of Theorem 6.19, we have the following answer, called the Barry-Esseen Theorem. For a proof, see Section XVI.5 of [Feller '72].

Theorem 6.20. (Barry-Esseen Theorem): *In the setting of Theorem* 6.19*, for every n and every $t \in \mathbb{R}$, we have*

$$\left|\mathbb{P}(Z_n \geq t) - \mathbb{P}(g \geq t)\right| \leq \frac{3\rho}{\sqrt{n}}.$$

Here $\rho = \mathbb{E}|X_1 - \mu|^3/\sigma^3$ and $g \sim N(0,1)$.

Despite the fact that this theorem provides a quantitative version for the error, this decays to zero very slowly, even slower than linear in n. Since we are dealing with the normal distribution and we saw in Proposition 6.2 that its distribution has an exponential decay tail, we expect to find better estimates.

This is the purpose of Nonasymptotic Probability: derive quantitative finite versions of limit of sums of random variables with a typically exponential decay. It is a very rich and active research field, specially

important for computational purposes, such as in Monte Carlo methods and modern Data Science in general.

These concentration inequalities will provide bounds on how a random variable deviates from some value. Most often this value will be its expected value. In Section 6.2 we came across an example in Theorem 6.11. This example is a particular case of a very general statement. Also, in Theorem 6.9, we saw that if we know how to compute the cumulant-generating function of a family of random variables, then an exponential decay for the tail appears naturally. Before we move forward, let us look at an example from [Vershynin '15].

Example 6.21. If we toss a coin N times, what is the probability that we get at least $3N/4$ heads? Denoting the Bernoulli variables, i.e., variables which satisfy $\mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = 1/2$, by X_i, we model our problem through the binomial random variable $S_N = \sum_{i=1}^{N} X_i$. It is well known that $\mathbb{E}S_N = N/2$ and $\mathrm{Var}(S_N) = N/4$. Then, Chebyshev's inequality leads to

$$\mathbb{P}\big(S_N \geq 3N/4\big) \leq \mathbb{P}\big(|S_N - N/2| \geq N/4\big) \leq 4/N.$$

This is a linear decay. The main point here is that, for sufficiently large N, we can use a naive estimate from the Central Limit Theorem and then we should expect the following

$$\mathbb{P}\big(S_N \geq 3N/4\big) \leq \mathbb{P}\left(\frac{S_N - N/2}{\sqrt{N/4}} \geq \sqrt{N/4}\right) \approx \mathbb{P}(g \geq \sqrt{N/4}) = \frac{1}{\sqrt{2\pi}} e^{-N/8}.$$

The comparison between the two decay rates is shocking. However it is not possible to make the reasoning above rigorous. New techniques, such as Theorem 6.9, were invented in order to deduce such kind of estimates.

The literature of concentration inequalities is vast and has many results. It is important to cite the contributions of Hoeffding, Bernstein. Cramer, Chernoff, Azuma, McDiarmid [Ledoux '01]. This kind of techniques, where there is solely a finite number of variables, is fundamental for Data-Science and Machine Learning. In this scenario, N typically corresponds to the sample size of the dataset.

We saw in Theorem 6.11 that a concentration inequality is valid for subgaussian variables. For the subexponential case, we could ask if a similar inequality is available. In this case, we must proceed more carefully, since the tails of sub-exponential distributions may not decay fast enough to make the moment-generating function finite everywhere, as Theorem 6.16 shows. The inequality for this case must take into account moments of higher orders. The next theorem, known as Bernstein's inequality, quantifies these informations. Also, if the variance is small, then it can be a large improvement on Theorem 6.11.

Theorem 6.22. *Let $X_1, \ldots, X_M$ be independent zero mean random variables such that, for all integers $n \geq 2$,*

$$\mathbb{E}|X_i|^n \leq n! R^{n-2} \sigma_i^2/2 \qquad \text{for all } i \in [M] \tag{6.9}$$

for some constants $R > 0$ and $\sigma_i > 0$, $i \in [M]$. Then, for all $t > 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{M} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2/2}{\sigma^2 + Rt}\right), \qquad \text{where } \sigma^2 = \sum_{i=1}^{M} \sigma_i^2. \tag{6.10}$$

Proof. Let us estimate the moment-generating function of X_i. This will allows us to use Theorem 6.9. Again, using Taylor series of exponential and Fubini's theorem, we have

$$\mathbb{E}[\exp(\theta X_i)] = 1 + \theta\mathbb{E}[X_i] + \sum_{n=2}^{\infty} \frac{\theta^n \mathbb{E}[X_i^n]}{n!} = 1 + \frac{\theta^2\sigma_i^2}{2} \sum_{n=2}^{\infty} \frac{\theta^{n-2}\mathbb{E}[X_i^n]}{n!\sigma_i^2/2}.$$

If we define $F_i(\theta) = \sum_{n=2}^{\infty} \frac{\theta^{n-2}\mathbb{E}[X_i^n]}{n!\sigma_i^2/2}$, then we have $\mathbb{E}[\exp(\theta X_i)] = 1 + \theta^2\sigma_i^2 F_i(\theta)/2 \leq \exp(\theta^2\sigma_i^2 F_i(\theta)/2)$. Thus, introducing $F(\theta) = \max_{i\in[M]} F_i(\theta)$ and remembering that $\sigma^2 = \sum_{i=1}^{M} \sigma_i^2$, Theorem 6.9 yields

$$\mathbb{P}\Big(\sum_{i=1}^{M} X_i \geq t\Big) \leq \inf_{\theta>0} \exp(\theta^2\sigma^2 F(\theta)/2 - \theta t) \leq \inf_{0<R\theta<1} \exp(\theta^2\sigma^2 F(\theta)/2 - \theta t).$$

Since we have $\mathbb{E}[X_i^n] \leq \mathbb{E}[|X_i|^n]$, our hypothesis 6.9 leads to

$$F_i(\theta) \leq \sum_{n=2}^{\infty} \frac{\theta^{n-2}\mathbb{E}[|X_i|^n]}{n!\sigma_i^2/2} \leq \sum_{n=2}^{\infty} (R\theta)^{n-2} = \frac{1}{1-R\theta},$$

provided that $R\theta < 1$. Thus, we derive that $F(\theta) \leq (1-R\theta)^{-1}$ and obtain that

$$\mathbb{P}\Big(\sum_{i=1}^{M} X_i \geq t\Big) \leq \inf_{0<\theta R<1} \exp\left(\frac{\theta^2\sigma^2}{2(1-R\theta)} - \theta t\right). \tag{6.11}$$

Notice that the choice $\theta = t/(\sigma^2 + Rt)$ satisfies $R\theta < 1$. Plugging into Equation 6.11 allows us to conclude

$$\mathbb{P}\Big(\sum_{i=1}^{M} X_i \geq t\Big) \leq \exp\left(\frac{t^2\sigma^2}{2(\sigma^2+Rt)^2}\frac{1}{1-\frac{Rt}{\sigma^2+Rt}} - \frac{t^2}{\sigma^2+Rt}\right) = \exp\left(-\frac{t^2/2}{\sigma^2+Rt}\right).$$

The same estimate could be derived for $-X_i$ in place of X_i. Applying the union bound concludes the proof. □

From Theorem 6.22 we easily derive a Bernstein inequality type for subexponential variables.

Theorem 6.23. *Let $X_1, \ldots, X_M$ be independent zero mean subexponential random variables, i.e., $\mathbb{P}(|X_i| \geq t) \leq \beta e^{-\kappa t}$ for some constants $\beta, \kappa > 0$ for all $t > 0$, $i \in [M]$. Then*

$$\mathbb{P}\left(\left|\sum_{i=1}^{M} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{-(\kappa t)^2/2}{2\beta M + \kappa t}\right). \tag{6.12}$$

Proof. As in the proof of Theorem 6.6, we will estimate the high moments. So, for $n \geq 2 \in \mathbb{N}$,

$$\mathbb{E}|X_i|^n = n\int_0^\infty \mathbb{P}(|X_i| \geq t)t^{n-1}dt \leq \beta n \int_0^\infty e^{-\kappa t}t^{n-1}dt =$$

$$\beta n\kappa^{-n}\int_0^\infty e^{-x}x^{n-1}dx = \beta n\Gamma(n-1)\kappa^{-n} = n!\kappa^{-(n-2)}\frac{2\beta\kappa^{-2}}{2}.$$

Therefore, condition 6.9 holds with $R = \kappa^{-1}$ and $\sigma_i^2 = 2\beta\kappa^{-2}$. Thus, according to Theorem 6.22, the theorem follows. □

Bernstein's inequality could be generalized for matrices. One can show that there are exponential concentration inequalities for the spectral norm of a sum of independent random matrices. This is fundamental for, among other things, problems of blind deconvolution in Signal Processing and problems of covariance matrix estimation in Statistics. See the fantastic book [Tropp '15]. The next theorem is a large deviation inequality for Rademacher chaos. It tells us that a homogeneous Rademacher chaos can be controlled by a mix of subexponential and subgaussian tails. Such estimates are known by the name of *Hanson-Wright inequalities*.

Definition 6.24. Let $\epsilon = (\epsilon_1, \ldots, \epsilon_M)$ be a Rademacher vector. For a self-adjoint matrix $A \in \mathbb{C}^{M\times M}$ with zero diagonal we define the *homogeneous Rademacher chaos* by

$$X = \boldsymbol{\epsilon}^* A\boldsymbol{\epsilon} = \sum_{j\neq k} \epsilon_j\epsilon_k A_{jk} \tag{6.13}$$

Remark 35. *Since A is self-adjoint, X will be real-valued even when A is a complex matrix. this allows us to reduce the study to real-valued symmetric matrices $A \in \mathbb{R}^{M\times M}$.*

Theorem 6.25. *Let $A \in \mathbb{R}^{M\times M}$ be a symmetric matrix with zero diagonal and let ε be a Rademacher vector. Then the homogeneous Rademacher chaos X defined in Equation (6.13) satisfies, for $t > 0$*

$$\mathbb{P}(|X| \geq t) \leq 2\exp\left(-\min\left\{\frac{3t^2}{128||A||_F^2}, \frac{t}{32||A||_{2\to 2}}\right\}\right) = \begin{cases} 2\exp\left(-\frac{3t^2}{128||A||_F^2}\right) & \text{if } 0 < t \leq \frac{4||A||_F^2}{3||A||_{2\to 2}}, \\ 2\exp\left(-\frac{t}{32||A||_{2\to 2}}\right) & \text{if } t > \frac{4||A||_F^2}{3||A||_{2\to 2}}. \end{cases} \tag{6.14}$$

In order to prove it, we need a powerful technique, called *decoupling*, that reduces stochastic dependencies of variables such as Rademacher chaos. For a proof of this fact, see Theorem 6.1.1 of [Vershynin '15] or Theorem 8.11 of [Rauhut & Foucart '13].

Theorem 6.26. *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a sequence of independet random variables with $\mathbb{E}[X_i] = 0$ for all $i \in [M]$. Let α_{jk} with $j, k \in [n]$, be a double sequence of elements in a finite-dimensional vector space V. If $F : V \to \mathbb{R}$ is a convex functions, then*

$$\mathbb{E}\left[F\Big(\sum_{\substack{j,k=1\\ j\neq k}}^{n} \alpha_{j,k}X_jX_k\Big)\right] \leq \mathbb{E}\left[F\Big(4\sum_{\substack{j,k=1\\ j\neq k}}^{n} \alpha_{j,k}X_jX'_k\Big)\right],$$

where X' denotes an independent copy of X.

Proof. (of Theorem 6.25): Inspired by the ideas of Theorem 6.9, we will estimate the moment-generating function of X. For $\theta > 0$, using the convexity of $f(x) = \exp(\theta x)$ and Theorem 6.26 leads to

$$\mathbb{E}\exp(\theta X) = \mathbb{E}\exp\left(\theta\sum_{j\neq k}\epsilon_j\epsilon_k A_{jk}\right) \leq \mathbb{E}\exp\left(4\theta\sum_{j,k}\epsilon_j\epsilon'_k A_{jk}\right)$$

$$\mathbb{E}_\epsilon\mathbb{E}_{\epsilon'}\exp\left(4\theta\sum_k\epsilon'_k\sum_j\epsilon_j A_{jk}\right) \leq \mathbb{E}\exp\left(8\theta^2\sum_k\Big(\sum_j\epsilon_j A_{jk}\Big)\right), \tag{6.15}$$

where in the last step we used Theorem 6.11 conditionally on $\boldsymbol{\epsilon}$ and the fact that the subgaussian parameter for Rademacher is $c = 1/2$, as calculated in Example 6.14. By the symmetry of A, we have,

$$\sum_k\Big(\sum_j\epsilon_j A_{jk}\Big)^2 = \sum_k\sum_j\epsilon_j A_{jk}\sum_\ell\epsilon_\ell A_{\ell k} = \sum_{j,\ell}\epsilon_j\epsilon_\ell\sum_k A_{jk}A_{k\ell} = \boldsymbol{\epsilon}^* A^2\boldsymbol{\epsilon}.$$

Set $B = A^2$. We can estimate the moment-generatng function of the positive semidefinite chaos $\boldsymbol{\epsilon}^* B\boldsymbol{\epsilon}$ by

$$\mathbb{E}\big[\exp(\lambda\boldsymbol{\epsilon}^* A^2\boldsymbol{\epsilon})\big] = \mathbb{E}\left[\exp\Big(\lambda\sum_j B_j j + \lambda\sum_{j\neq k}\epsilon_j\epsilon_k B_{jk}\Big)\right] \leq e^{\lambda\mathrm{tr}(\boldsymbol{B})}\mathbb{E}\left[\exp\Big(4\lambda\sum_{j,k}\epsilon_j\epsilon'_k B_{jk}\Big)\right]$$

$$\leq e^{\lambda\mathrm{tr}(\boldsymbol{B})}\mathbb{E}\left[\exp\Big(8\lambda^2\sum_k\Big(\sum_j\epsilon_j B_{jk}\Big)^2\Big)\right].$$

Again we have applied Theorem 6.11 conditionally on $\boldsymbol{\epsilon}$ and Theorem 6.26. Since B is positive semidefinite, its square-root exists and then

$$\sum_k\Big(\sum_j\epsilon_j B_{jk}\Big)^2 = \boldsymbol{\epsilon}^* B\boldsymbol{\epsilon} = (B^{1/2}\boldsymbol{\epsilon})^* B(B^{1/2}\boldsymbol{\epsilon}) \leq ||B||_{2\to 2}\boldsymbol{\epsilon}^* B\boldsymbol{\epsilon}.$$

In the case $8\lambda||B||_{2\to 2} < 1$, Jensen's inequality yields

$$\mathbb{E}\exp(\lambda\boldsymbol{\epsilon}^*B\boldsymbol{\epsilon}) \leq \exp(\lambda\mathrm{tr}(B))\mathbb{E}\big[\exp(8\lambda^2||B||_{2\to 2}\boldsymbol{\epsilon}^*B\boldsymbol{\epsilon})\big] \leq \exp(\lambda\mathrm{tr}(B))\Big(\mathbb{E}\big[\exp(\lambda\boldsymbol{\epsilon}^*B\boldsymbol{\epsilon})\big]\Big)^{8\lambda||B||_{2\to 2}}.$$

Rearranging this expression leads to

$$\mathbb{E}\big[\exp(\lambda\boldsymbol{\epsilon}^*B\boldsymbol{\epsilon})\big] \leq \exp\left(\frac{\lambda\mathrm{tr}(B)}{1-8\lambda||B||_{2\to 2}}\right), \qquad 0<\lambda<(8||B||_{2\to 2})^{-1}. \tag{6.16}$$

After setting $\lambda = 8\theta^2$ we have, for $0<\theta<(16||A||_{2\to 2})^{-1}$,

$$\mathbb{E}\exp(\theta X) \leq \exp\left(\frac{8\theta^2\mathrm{tr}(A^2)}{1-64\theta^2||A^2||_{2\to 2}}\right) = \exp\left(\frac{8\theta^2||A||_F^2}{1-64\theta^2||A||_{2\to 2}^2}\right).$$

Using Markov's inequality, for $0<\theta\leq(16||A||_{2\to 2})^{-1}$, we obtain

$$\mathbb{P}(X\geq t) = \mathbb{P}(e^{\theta X}\geq e^{\theta t}) \leq e^{-\theta t}\mathbb{E}\big[e^{\theta X}\big] \leq \exp\left(-\theta t + \frac{8\theta^2||A||_f^2}{1-64\theta^2||A||_{2\to 2}^2}\right)$$

$$\leq \exp\left(-\theta t + \frac{8\theta^2||A||_F^2}{1-1/4}\right) \leq \exp(-\theta t + 32\theta^2||A||_F^2/3).$$

If $t\leq 4||A||_F^2/(3||A||_{2\to 2})$, the optimal choice $\theta = 3t/(64||A||_F^2)$ satisfies $\theta\leq(16||A||_{2\to 2})^{-1}$. For t given by this inequality, we therefore obtain

$$\mathbb{P}(X\geq t) \leq \exp\left(-\frac{3t^2}{128||A||_F^2}\right).$$

Now, we must look to the complementary inequality, namely $t>4||A||_F^2/(3||A||_{2\to 2})$. In this case, setting $\theta=(16||A||_{2\to 2})^{-1}$ leads to

$$\mathbb{P}(X\geq t) \leq \exp(-\theta t + 32\theta^2||A||_F^2/3) \leq \exp(-\theta t+\theta t/2) = \exp(-\theta t/2) = \exp(-t/(32||A||_{2\to 2})),$$

since in this case we have $\theta<3t/(64||A||_F^2)$. To finish, see that X has the same distribution of $-X$ and we can derive the same bounds for $\mathbb{P}(X\leq -t)$. Taking the union bound yields the theorem. □

This kind os estimate can be generalized to quadratic forms involving more general subgaussian random vectors, see [Hanson & Wright '71]. Specifically in the Gaussian case, see [Bechar '09].

6.4 Comparison of Gaussian Processes

"Tout le monde croit que les erreurs suivent une loi normale, me disait un jour M. Lippmann, car les expérimentateurs car ils pensen qu'il s'agit d'un théorème, et les mathématiciens que pensent que c'est un fait expérimental."[3]
Henri Poincaré in Calcul des Probabilités, p.171

In many situations one wishes to compare two families of random variables $(X_t)_{t=1}^n$ and $(Y_t)_{t=1}^n$ in order to extract some information about one of them using the other. There are many reasons for that but typically we choose the latter as having more tractable properties than the former. Thus we wish to provide an upper bound on $(X_t)_{t=1}^n$ just by knowing that $(Y_t)_{t=1}^n$ is larger in some sense.

This is remarkable in the case of Gaussian variables. In the inconspicuous report [Slepian '62], David Slepian realized that the covariance structure can be regarded as an indicator of how Gaussian variables behave jointly. Therefore, using its correlation we can quantify how much the variables in each of the

[3] A free translation of Poincarés' quote is *"Everybody believes that the errors follow a normal law, said one day to me M. Lippmann, because the experimenters think that it is a theorem, and the mathematicians think that it is an experimental fact."*

families maintain similar expected magnitudes. Then, if we want to compare functions of two families of Gaussian random variables, it is a good ideia to compare their covariances. This is confirmed by Theorem 6.29, due to Slepian and Fernique. Before we state it, let us make some definitions.

Definition 6.27. A *stochastic process* is a collection X_t, $t \in T$ of random variables indexed by some set T. It is said to be *centered* if $\mathbb{E}X_t = 0$ for all $t \in T$. A process is called *centered Gaussian* if for each finite collection $t_1, \ldots, t_n \in T$, the random vector $(X_{t_1}, \ldots, X_{t_n})$ is a zero mean Gaussian random vector.

A typical Gaussian Process is given by $X_t = \sum_{i=1}^{M} g_i x_i(t)$, where $\mathbf{g} = (g_1, \ldots, g_M)$ is a standard Gaussian random vector and $x_j : T \to \mathbb{R}$, are arbitrary functions. It is interesting to note that Gaussian process are receiving increasing attention in the Machine Learning community and leading to new solutions for a wide class of regression and classification problems. One can see [Rasmussen & Williams '05] for details of how to deal with them in a computational fashion. As in the finite case, Gaussian processes can be generalized to the concept of subgaussian processes.

Definition 6.28. Given a stochastic process, we define the pseudometric (two distinct points can have zero distance)

$$d(s,t) = (\mathbb{E}|X_s - X_t|^2)^{1/2}, \qquad s,t \in T.$$

A centered stochastic process X_t is called *subgaussian* if

$$\mathbb{E}\exp(\theta(X_s - X_t)) \leq \exp(\theta^2 d(s,t)^2/2), \qquad s,t \in T,\ \theta > 0. \tag{6.17}$$

Now, the comparison lemma.

Theorem 6.29. *Let X, Y be zero mean Gaussian random vectors on $\mathbb{R}^m$. If*

$$\mathbb{E}|X_i - X_j|^2 \leq \mathbb{E}|Y_i - Y_j|^2 \qquad \forall i,j \in [m],$$

then

$$\mathbb{E}\max_{j\in[m]} X_j \leq \mathbb{E}\max_{j\in[m]} Y_j.$$

Since we have $\mathbb{E}|X_i - X_j|^2 = \mathbb{E}X_i^2 - 2\mathbb{E}X_iX_j + \mathbb{E}X_j^2$, if we have the additional assumption $\mathbb{E}X_i^2 = \mathbb{E}Y_i^2$, then the hypothesis of the Slepian-Fernique Comparison Lemma is, actually, $\mathbb{E}X_iX_j \geq \mathbb{E}Y_iY_j$. Therefore, as we said in the beginning of this section, the comparison of covariance implies in the comparison of expected maxima of Gaussian vectors. This theorem has a deep generalization for min-max of Gaussian variables disposed in a retangular array. It was proved in [Gordon '85] and restated in [Gordon '88].

Theorem 6.30. (Theorem A of [Gordon '88]): *Let X_{ij}, Y_{ij}, $i \in [n], j \in [m]$ be two families of zero mean Gaussian random variables. If*

$$\mathbb{E}|X_{ij} - X_{kl}|^2 \leq \mathbb{E}|Y_{ij} - Y_{kl}|^2 \quad \forall i \neq k \text{ and } j,l,$$
$$\mathbb{E}|X_{ij} - X_{il}|^2 \geq \mathbb{E}|Y_{ij} - Y_{il}|^2 \quad \forall i,j,l,$$

then

$$\mathbb{E}\min_{i\in[n]}\max_{j\in[m]} X_{ij} \geq \mathbb{E}\min_{i\in[n]}\max_{j\in[m]} Y_{ij}.$$

This comparison lemma was generalized even more. One should look at [Kahane'86] and [Vitale '00] for some generalizations. Also, [Tong '80] is a whole monograph dedicated to inequalities for multivariate distributions. There, some interesting comments related to comparisons lemmas can be found.

Theorem 6.29 also generalizes to Gaussian processes indexed by infinite sets. If we have that $\mathbf{X} = (X_t)_{t\in T}$ and $\mathbf{Y} = (Y_t)_{t\in T}$ are Gaussian processes and if $\mathbb{E}|X_s - X_t|^2 \leq \mathbb{E}|Y_s - Y_t|^2$ for all $s,t \in T$, then it also holds that

$$\mathbb{E}\sup_{t\in T} X_t \leq \mathbb{E}\sup_{t\in T} Y_t,$$

where $\mathbb{E}\sup_{t\in T} X_t$, in this case, is defined by $\mathbb{E}\sup_{t\in T} X_t = \sup\{\mathbb{E}\sup_{t\in F} X_t, F \subset T, F \text{ finite}\}$. This is called *lattice supremum* and is used in order to avoid problems with nonmensurable sets. Theorem 6.30 generalizes in a similar way, using double index Gaussian processes.

Despite the intrinsic interest on the comparison theorems, in the field of Compressive Sensing we are highly interested in some of their geometric corollaries. Before that, we need a notion about how to measure the width of a set. It is a measure of the size of a set T in the sense that how well, on average, the vectors in the set T can align with a randomly chosen direction.

Definition 6.31. For a set $T \subset \mathbb{R}^N$, we define its *Gaussian width* by

$$\ell(T) = \mathbb{E}\sup_{x\in T}\langle g, x\rangle,$$

where $g \in \mathbb{R}^N$ is a standard Gaussian variable.

With the aid of this definition and denoting, for a Gaussian vector $g \in \mathbb{R}^m$, the mean of its ℓ_2-norm by $\mathbb{E}||g||_2 = E_m$, we have the celebrated *Gordon's escape through the mesh theorem*, consequence of Gordon's Comparison Lemma.

Theorem 6.32. *Let $A \in \mathbb{R}^{m\times N}$ be a Gaussian random variable and T be a subset of the unit sphere $S^{N-1} = \{x \in \mathbb{R}^N, ||x||_2 = 1\}$. Then, for $t > 0$,*

$$\mathbb{P}\left(\inf_{x\in T}||Ax||_2 \leq E_m - \ell(T) - t\right) \leq e^{-t^2/2}. \tag{6.18}$$

Its proof can be found in Theorem 9.21 of [Rauhut & Foucart '13]. It relies on the comparison between two well chosen Gaussian processes and on the use ofTheorem 6.30. The first process is, for $x \in T$ and $y \in S^{m-1}$, $X_{x,y} = \langle Ax, y\rangle$ while the second is $Y_{x,y} = \langle g, x\rangle + \langle h, x\rangle$, for $g \in \mathbb{R}^N$ and $h \in \mathbb{R}^m$ independent standard Gaussian vectors. A very similar argument will be developed in Theorem 7.12.

In this dissertation, we aim to understand when a measurement matrix is good in the sense that we can use it to recover the unique sparsest vector it "captures". Next chapter will deal with random matrices as sensing matrices. In order to determine how good a random matrix is, thinking in terms of the Null Space Property, good matrices will be the ones that avoids certain subsets. This is what Gordon called "escapes a mesh". For this to happen, of course, these subsets must be small in some sense. The precise sense is given exactly by the concept of Gaussian width. A consequence of Theorem 6.32 that really deserves to be called a *escape through the mesh* theorem is the next one.

Its uses for the problem of sparse recovery were first done by [Rudelson & Vershynin '08] and generalized for other convex problems by [Chandrasekaran, Recht, Parrilo & Willsky '12]. We start by defining what is a Grassmanian.

Definition 6.33. Let V be a finite-dimensional vector space over a field $\mathbb{K}$. The *Grassmannian* $\mathrm{G}_k(V)$ is the set of all k-dimensional linear subspaces of V. If V has dimension d, then the Grassmannian is also denoted $\mathrm{G}_k(d)$.

The following useful description is used for the Grassmanian. First, we need some definitions.

Definition 6.34. For $E \in \mathrm{G}_k(d)$, we denote by P_E the orthogonal projection onto E. We denote by $M(d, \mathbb{R})$ the set of all $d \times d$ real matrices endowed with the Frobenius norm. Also, we define the set

$$G(d, k) = \{P \in M(d, \mathbb{R}) \mid P^2 = P,\ P^* = P,\ \mathrm{rank}(P) = k\}.$$

Proposition 6.35. *The function $f : G_k(V) \to G(d, k)$, $E \mapsto P_E$ is a bijection. The set $G(d, k)$ is a compact metric space.*

With a little more work, this identification allows us to represent $\mathrm{G}_k(V)$ as a submanifold of the space of symmetrical matrices. Therefore, we can study all the interesting properties of the Grassmanian through $G(d, k)$ and then transfer the results by using the function $f : \mathrm{G}_k(V) \to G(d, k)$. Finally, we can state the geometrical version of the Gordon's Escape Through the Mesh Theorem.

Theorem 6.36. (Theorem 3.3 and Corollary 3.4 in [Gordon '88]): *Take a closed subset T of the unit sphere $S^{N-1} = \{x \in \mathbb{R}^N, ||x||_2 = 1\}$ and let $B_\varepsilon = \{x \in \mathbb{R}^N : ||x||_2 \leq \varepsilon\}$. If $\ell(T) < (1-\varepsilon)E_m - \varepsilon E_N$, then an $(N-m)$-dimensional subspace Y drawn uniformly from the Grassmannian $G_{N-m}(N)$ satisfies*

$$\mathbb{P}\Big(Y \cap (S + B_\varepsilon) = \emptyset\Big) \geq 1 - \frac{7}{2}\exp\bigg(-\frac{1}{2}\bigg(\frac{(1-\varepsilon)E_m - \varepsilon E_N - \ell(T)}{3 + \varepsilon + \varepsilon E_N/E_m}\bigg)^2\bigg).$$

In particular, with $\varepsilon = 0$, we have that if $\ell(S) < E_m$, then an $(N-m)$-dimensional subspace Y drawn uniformly from the Grassmannian $G_{N-m}(N)$ satisfies

$$\mathbb{P}\Big(Y \cap S = \emptyset\Big) \geq 1 - \frac{7}{2}\exp\bigg(-\frac{1}{18}\Big(E_m - \ell(S)\Big)^2\bigg).$$

The main idea of its proof consists in making the identification of the random subspace Y with the null space of a certain $m \times N$ random matrix A with Gaussian entries. After, we change the original problem of estimating $\mathbb{P}\big(\text{null}(A) \cap (S + B_\epsilon) = \emptyset\big)$ into an estimative on $\mathbb{P}\big(\|A\|_2 \geq (1+\delta)\mathbb{E}\|A\|_2\big)$. This last probability is, in turn, controlled by the concentration of measure inequality that we will develop in the next section. The complete proof (with nice intuitive explanations) can be found at [Mixon's Blog - 02/08/2014].

6.5 Concentration of Measure

"Now only after Fourier and then by Paul Levy
Is rendered easy such a proof;
Without their tools perhaps, dear listener,
You'd demonstrate it as a tour de force;
I've tried without success."
Excerpt of "de Moivre", a poem by Richard Dudley in [Dudley '75].

The roots of concentration of measure phenomena can be traced back to the works of Émile Borel and Paul Lévy. However, it was not before Vitali Milman's work on Geometric Banach Space Theory that modern concepts and theorems related to the phenomena emerged. There, he revisited proofs of theorems and provides deep analysis related to spherical sections of convex sets into high dimension normed spaces. The understanding of this phenomena is a deep achievement, as Michel Talagrand points out: *"The idea of concentration of measure, which was discovered by Vitali Milman, is unarguably one of the greatest ideas of analysis in our times"* [Ledoux - Lecture II].

The quotes' author is another important character in the history of concentration of measure. In the nineties he established connections between isoperimetric inequalities and concentration inequalities. See section 1.2 of [Boucheron, Lugosi & Massart '13] for this connection. After these contributions, he was able to solve many open problems on summation of random variables.

Loosely speaking, if we have a large number of random variables $X_1, \ldots, X_n$, one might ask what happens with their sum assuming that they are bounded in average. This phenomenon essentially tells us that if we have sufficient independence between them, this sum will be highly concentrated in a much narrower range than the one expected by the classical convergence theorems of Probability. We can quantify this by proving *large deviation* inequalities, that is, exponential upper bounds for the probability of the sum (or more general functions) of random variables deviates from its mean.

This hypothesis of high degree of independence is fundamental. The idea is that since the random variables are independent, it is difficult for them to be "synchronized" in order to deviate its sum too far from its mean.

There exist monographs entirely dedicated to this theme such as [Boucheron, Lugosi & Massart '13] and [Ledoux '01]. Once again, Terence Tao's blog entries have elementary and pedagogical insights about the subject [Tao's Blog - 01/03/2010]. Also, the book [Dubhashi & Panconesi '12] shows applications of the theory and starts the discussion about the concentration phenomena from scratch.

Despite the generality of concentration of measure, in this dissertation, it will have a very special meaning, namely, the guarantees that any L-Lipschitz function of a standard Gaussian random vector,

regardless of the dimension, exhibits concentration like a scalar Gaussian variable with variance L^2. It was first proved by [Tsirelson, Ibragimov & Sudakov '76] using arguments based on stochastic calculus.

Theorem 6.37. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function with constant $L > 0$,* i.e.

$$|f(x) - f(y)| \leq L||x - y||_2 \qquad \forall x, y \in \mathbb{R}^n,$$

and g be a standard Gaussian random vector. Then, for all $t > 0$

$$\mathbb{P}\Big(f(g) - \mathbb{E}\big[f(g)\big] \geq t\Big) \leq \exp\bigg(-\frac{t^2}{2L^2}\bigg),$$

and consequently

$$\mathbb{P}\Big(\Big|f(g) - \mathbb{E}\big[f(g)\big]\Big| \geq t\Big) \leq 2\exp\bigg(-\frac{t^2}{2L^2}\bigg).$$

Before we develop all the required tools to prove this theorem, we must say that the constant 2 inside the exponential is optimal. This follows from the case $n = 1$, $f(x) = x$ and the lower tail bound for the standard Gaussian stated in Proposition 6.2.

Remark 36. *It is important to note that without any additional structure for the function f (such as convexity), dimension-free concentration for Lipschitz functions need not hold for an arbitrary subgaussian distribution. One can consult [Ledoux '01] for further results.*

The original proof uses the idea of symmetrization, explored in Example 6.15 and Laplace transform method, that is, to bound $\phi(\theta) = \mathbb{E}\exp(\theta(f(X) - \mathbb{E}[f(X)]))$. It is less laborious than the approach adopted here, but is has the disadvantage that it does not provide optimal constants for concentration. Instead, we will rely on information theory methods, known as *Herbst argument* after [Davies & Simon '84]. First, we need the definition of entropy of a random variable.

Definition 6.38. For a nonnegative random variable X, we define its entropy as

$$\mathcal{E}(X) = \mathbb{E}[\phi(X)] - \phi(\mathbb{E}(X)) = \mathbb{E}[X \ln X] - \mathbb{E}(X)\ln(\mathbb{E}(X)),$$

where ϕ is defined, for $x > 0$, as $\phi(x) = x\ln(x)$ and extends continuously to $x = 0$ by $\phi(0) = 0$. If the first term in infinite, then we set $\mathcal{E}(X) = \infty$.

It follows from the convexity of ϕ and by Jensen inequality that $\mathcal{E}(X) \geq 0$. Also, entropy of a random variable is homogeneous of zeroth order, $\mathcal{E}(tX) = t\mathcal{E}(X)$, by definition. Now, we are able to explain the entropy method for obtaining concentration inequalities.

Remark 37. *(Herbst Argument): We have a method, attributed to Ira Herbst[4] to derive concentration inequalities using information theory ideas. It is also called* entropy method *and it became highly popular after the work of Michel Ledoux on Talagrand's inequalities. See [Tao's Blog - 09/06/2009] and [Ledoux '01]. The idea is to derive a bound on the entropy of $e^{\theta X}$, for $\theta > 0$ of the form $\mathcal{E}(e^{\theta X}) \leq g(\theta)\mathbb{E}[e^{\theta X}]$, for some function g. If we define $F(\theta) = \mathbb{E}[e^{\theta X}]$, this inequality will be equivalent to*

$$\mathcal{E}(e^{\theta X}) \leq \theta F'(\theta) - F(\theta)\ln F(\theta) \leq g(\theta)F(\theta). \tag{6.19}$$

Also, if we set $G(\theta) = \theta^{-1}\ln F(\theta)$, then Equation (6.19) *reduces to $G'(\theta) \leq \theta^{-2}g(\theta)$. Moreover, we can note that $G(0) = \lim_{\theta\to 0}\theta^{-1}\ln F(\theta) = F'(0)/F(0) = \mathbb{E}[X]$. Therefore, integrating on both sides yields*

$$G(\theta) - G(0) = G(\theta) - \mathbb{E}[X] = \int_0^\theta G'(t)dt \leq \int_0^\theta t^{-2}g(t)dt,$$

[4]This argument about how to derive concentration inequalities from Logarithm Sobolev Inequalities appeared for the first time in Appendix A of [Davies & Simon '84].

which implies that

$$\mathbb{E}[e^{\theta(X-\mathbb{E}[X])}] = e^{\theta[G(\theta)-\mathbb{E}(x)]} \leq \exp\left(\theta \int_0^\theta t^{-2} g(t) dt\right), \qquad \theta > 0.$$

To conclude, we simply uses Markov's inequality to derive tail bounds for $X - \mathbb{E}[X]$.

Herbst argument together with logarithmic Sobolev inequality for Gaussian vectors are the main core of our proof of measure concentration. Before we turn to it, we need to develop some information theory results, such as the dual characterization of entropy. Let us begin with a basic convex inequality.

Lemma 6.39. *For a given function* $f : \mathbb{R}^N \to (-\infty, \infty]$ *we have*

$$\langle x, y \rangle \leq f(x) + f^*(y) \qquad \text{for all} \quad x, y \in \mathbb{R}^N,$$

where $f^* : \mathbb{R}^N \to (-\infty, \infty]$ *is the convex conjugate of* f *defined by* $f^*(y) = \sup_{x \in \mathbb{R}^N} \{\langle x, y \rangle - F(x)\}$. *For the particular case* $f(x) = \exp(x)$, *it reads* $xy \leq e^x + y \ln y - y$ *for all* $x \in \mathbb{R}, y > 0$.

Proof. The inequality is obvious from the definition. For the particular case of $f(x) = e^x$, we have that the function $x \mapsto xy - \exp(x)$ takes its maximum at $x = \ln y$ if $y > 0$ and this gives

$$f^*(y) = \sup_{x \in \mathbb{R}} \{xy - e^x\} = \begin{cases} y \ln y - y & \text{if } y > 0, \\ 0 & \text{if } y = 0, \\ \infty & \text{if } y < 0. \end{cases}$$

From this convex conjugate, the inequality follows immediately. □

Using this inequality, we are able to prove a dual characterization for entropy which will be very useful

Lemma 6.40. *Let* X *be a nonnegative random variable satisfying* $\mathbb{E}[X] < \infty$. *Then*

$$\mathcal{E}(X) = \sup\{\mathbb{E}[XY] : \mathbb{E}[\exp(Y)] \leq 1\}.$$

Proof. Let us prove, first, the case of a strictly positive random variable X. We may assume that $\mathbb{E}X = 1$ due to the homogeneity of the entropy. By Lemma 6.39, for a random variable Y satisfying $\mathbb{E}[\exp(Y)] \leq 1$, we have

$$\mathbb{E}[XY] \leq \mathbb{E}[\exp(Y)] + \mathbb{E}[X \ln X] - \mathbb{E}[X] \leq \mathbb{E}[X \ln X] = \mathcal{E}(X).$$

The next step is to prove the converse direction, that is, that $\mathcal{E}(X) \leq \mathbb{E}[XY]$ for some Y. In order to do this, we just simple choose $Y = \ln X - \ln(\mathbb{E}X)$. By the definition of $\mathcal{E}$, this choice satisfies $\mathcal{E}(X) = \mathbb{E}[XY]$ and also $\mathbb{E}\exp(Y) = \mathbb{E}[X]\exp(-\ln(\mathbb{E}X)) = 1$. Therefore we have attained the supremum and this concludes the theorem for positive variables. For the general nonnegative case, we just use an approximation argument and the continuity of $\phi(x) = x \ln x$. □

We have two useful consequences of the above dual characterization. Another characterization using positive random variables and the subadditivity of entropy.

Corollary 6.41. *For a nonnegative variable* X, *we have*

$$\mathcal{E}(X) = \sup\{\mathbb{E}[X \ln(Z)] - \mathbb{E}[X] \ln(\mathbb{E}[Z]) : Z > 0\}.$$

Also, for two nonnegative random variables X *and* Y, *we have* $\mathcal{E}(X + Y) \leq \mathcal{E}(X) + \mathcal{E}(Y)$.

Proof. For the first part, we just make the substitution $Y = \ln(Z/\mathbb{E}(Z))$ for a positive random variable Z. The second part follows from the properties of supremum and expected value. □

It is time to introduce some notation. For a sequence of random variables $\mathbf{X} = (X_1, \ldots, X_n)$ and any index $i \in [n]$, we define $\mathbf{X}^i = (X_1, \ldots, X_{i-1}, \hat{X}_i, X_{i+1}, \ldots, X_n) = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$, that is, we have the exclusion of the variable X_i from our sequence. Also, for a function f of $\mathbf{X}$, we will write

$$\mathbb{E}_{X_i} f(\mathbf{X}) = \mathbb{E}_{X_i}[f(X_1, \ldots, X_i, \ldots, X_n)] = \mathbb{E}[f(\mathbf{X}|\mathbf{X}^i)].$$

for the conditional expectation. As one can note, this is a function of $X_1, \ldots X_{i-1}, X_{i+1}, \ldots, X_n$ but it is constant with respect to the variable X_i. From this, we can define the conditional entropy.

Definition 6.42. For a vector $\mathbf{X} = (X_1, \ldots, X_n)$ and a function $f(\mathbf{X})$ of this vector, the *conditional entropy* is defined by

$$\begin{aligned}\mathcal{E}_{X_i}(f(\mathbf{X})) &= \mathcal{E}\big(f(\mathbf{X}|\mathbf{X}^i)\big) = \mathbb{E}_{X_i}(\phi(f(\mathbf{X}))) - \phi(\mathbb{E}_{X_i}(f(\mathbf{X}))) \\ &= \mathbb{E}_{X_i}[f(\mathbf{X}) \ln f(\mathbf{X})] - \mathbb{E}_{X_i}[f(\mathbf{X})] \ln(\mathbb{E}_{X_i}[f(\mathbf{X})]).\end{aligned}$$

Now we have developed all the basic tools in order to start the laborious series of theorems that will finally lead to Theorem 6.37. The first theorem we will prove is called *tensorization inequality*. It enables us to bound the entropy of a function of several variables by the sum of the entropies of this function in terms of the individual variables.

Theorem 6.43. *Let $\mathbf{X} = (X_1, \ldots, X_n)$ be a vector of independent random variables and let f be a nonnegative function satisfying $\mathbb{E}[f(\mathbf{X})] < \infty$. Then*

$$\mathcal{E}(f(\mathbf{X})) \leq \mathbb{E}\Big[\sum_{i=1}^{n} \mathcal{E}_{X_i}(f(\mathbf{X}))\Big].$$

Proof. We will prove the theorem for strictly positive f. As in Lemma 6.40, for the general case we just use that function ϕ is continuous at 0 and use an approximation argument. Let us define the conditional expectation operator $\mathbb{E}^i$ given by $\mathbb{E}^i[f(\mathbf{X})] = \mathbb{E}_{X_1,\ldots,X_{i-1}}[f(\mathbf{X})] = \mathbb{E}[\mathbf{X}|X_i, \ldots, X_n]$. This operator integrates out the dependence on the first $i-1$ variables. Of course we have $\mathbb{E}^1[f(\mathbf{X})] = f(\mathbf{X})$ and $\mathbb{E}^{n+1}[f(\mathbf{X})] = \mathbb{E}[f(\mathbf{X})]$. Next, a smart telescopic decomposition leads to

$$\ln(f(\mathbf{X})) - \ln(\mathbb{E}[f(\mathbf{X})]) = \sum_{i=1}^{n} (\ln(\mathbb{E}^i[f(\mathbf{X})]) - (\ln(\mathbb{E}^{i+1}[f(\mathbf{X})])). \tag{6.20}$$

Now we use the characterization given by Corollary 6.41 with $X = f(\mathbf{X})$, $Z = \mathbb{E}^i[f(\mathbf{X})] > 0$ and $\mathbb{E} = \mathbb{E}_{X_i}$.

$$\mathbb{E}_{X_i}\Big[f(\mathbf{X})\Big(\ln(\mathbb{E}^i[f(\mathbf{X})]) - \ln\big(\mathbb{E}_{X_i}[\mathbb{E}^i[f(\mathbf{X})]]\big)\Big)\Big] \leq \mathcal{E}_{X_i}(f(\mathbf{X})).$$

Using the independence and Fubini's theorem to interchange the order of integration yields

$$\mathbb{E}_{X_i}[\mathbb{E}[f(\mathbf{X})]] = \mathbb{E}_{X_i}\mathbb{E}_{X_1,\ldots,X_{i-1}}[f(\mathbf{X})] = \mathbb{E}^{i+1}[f(\mathbf{X})].$$

To finish, just take Equation (6.20), multiply it by $f(\mathbf{X})$ and take expectation on both sides. This finally leads to

$$\mathcal{E}(f(\mathbf{X})) = \mathbb{E}\big[f(\mathbf{X})\big(\ln(f(\mathbf{X})) - \ln(\mathbb{E}[f(\mathbf{X})])\big)\big] =$$

$$\sum_{i=1}^{n} \mathbb{E}\Big[\mathbb{E}_{X_i}\Big[f(\mathbf{X})\big(\ln(\mathbb{E}^i[f(\mathbf{X})]) - \ln(\mathbb{E}_{X_i}[\mathbb{E}^i[f(\mathbf{X})]])\big)\Big]\Big] \leq \sum_{i=1}^{n} \mathbb{E}[\mathcal{E}_{X_i}(f(\mathbf{X}))].$$

□

There is a deep connection between concentration of measures and functional inequalities such as the logarithmic Sobolev inequality. These inequalities are ubiquitous in many areas such as Partial Differential Equations, Information Theory and Optimal Transport among others beyond, of course, Probability Theory. In the latter, it has many connections, not only with concentration inequalities but also with mixing times of Markov Chains. It is interesting to mention that in Information Theory, this type of inequality shows a relation between entropy and Fisher information [Kitsos & Tavoularis '09]. They were also used for the celebrated proof of Poincaré's Conjecture by G. Perelman. See [Tao's Blog - 24/04/2008] for further information. The next theorem is an example of a log-Sobolev inequality for the case of Rademacher vectors.

Theorem 6.44. *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a real-valued function and $\boldsymbol{\epsilon}$ be an n-dimensional Rademacher vector. Then*

$$\mathcal{E}(f^2(\boldsymbol{\epsilon})) \leq \frac{1}{2}\mathbb{E}\Big[\sum_{i=1}^{n}(f(\boldsymbol{\epsilon}) - f(\overline{\boldsymbol{\epsilon}}^{(i)})^2\Big], \tag{6.21}$$

where $\overline{\epsilon}^{(i)} = (\epsilon_1, \ldots, \epsilon_{i-1}, -\epsilon_i, \epsilon_{i+1}, \ldots, \epsilon_n)$ is obtained from $\boldsymbol{\epsilon}$ by flipping the ith entry.

Proof. By Theorem 6.43, we have $\mathcal{E}(f^2(\boldsymbol{\epsilon})) \leq \mathbb{E}\big[\sum_{i=1}^{n}\mathcal{E}_{\epsilon_i}(f^2(\boldsymbol{\epsilon}))\big]$. Therefore, we just need to prove that, for each $i \in [n]$,

$$\mathcal{E}_{\epsilon_i}\Big(f^2(\boldsymbol{\epsilon})\Big) \leq \frac{1}{2}\mathbb{E}_{\epsilon_i}\Big[f(\boldsymbol{\epsilon}) - f(\overline{\boldsymbol{\epsilon}}^{(i)})^2\Big]. \tag{6.22}$$

For any realization of $(\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_{i+1}, \ldots, \epsilon_n)$, $f(\boldsymbol{\epsilon})$ and $f(\overline{\boldsymbol{\epsilon}}^{(i)})$ can only have two possible values. These values will be denoted by a and $b \in \mathbb{R}$, respectively. Using the definition of entropy, inequality (6.22) is the same as the following scalar inequality

$$\frac{1}{2}\big(a^2\ln(a^2) + b^2\ln(b^2)\big) - \frac{a^2+b^2}{2}\ln\left(\frac{a^2+b^2}{2}\right) \leq \frac{1}{2}(a-b)^2.$$

Thus, it remains to prove this elementary inequality. In order to do this, let us define, for a fixed b, the function

$$H(a) = \frac{1}{2}\big(a^2\ln(a^2) + b^2\ln(b^2)\big) - \frac{a^2+b^2}{2}\ln\left(\frac{a^2+b^2}{2}\right) - \frac{1}{2}(a-b)^2.$$

A tedious computation shows that the first and the second derivative of $H(a)$ are given by

$$H'(a) = a\ln\left(\frac{2a^2}{a^2+b^2}\right) - (a-b) \qquad \text{and} \qquad H''(a) = \ln\left(\frac{2a^2}{a^2+b^2}\right) + 1 - \frac{2a^2}{a^2+b^2}.$$

The first thing to note is that $H(b) = 0$ and $H'(b) = 0$. Also, since $\ln x \leq x - 1$, we obtain $H''(a) \leq 0$ for all $a \in \mathbb{R}$. This implies that H is concave and then $H(a) \leq 0$ for all $a \in \mathbb{R}$. This establishes the theorem. □

The right-hand side of equation (6.21) can be interpreted as a gradient in a discrete version. If we try to prove a similar inequality for continuous variables, such as Gaussian variables, then an authentic gradient appears. This is the content of the Gaussian logarithm Sobolev inequality, proved by [Gross '75]. As pointed by [Ledoux - Lecture I], nowadays we have more than 15 different proofs of this inequality.

Theorem 6.45. (Essentially Corollary 4.2 of [Gross '75]): *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function satisfying $\mathbb{E}[\phi(f^2(\mathbf{g}))] < \infty$ for a standard Gaussian vector $\mathbf{g}$ on $\mathbb{R}^n$. Then*

$$\mathcal{E}(f^2(\mathbf{g})) \leq 2\mathbb{E}\Big[||\nabla f(\mathbf{g})||^2\Big].$$

Proof. The initial step is to prove the theorem for $n = 1$ and g being a standard Gaussian random variable. After, we will prove a general version. Even more, we will start by taking f with compact support. Since f' is uniformly continuous, its modulus of continuity $\omega(f', \delta) = \sup_{|t-s|\leq\delta} |f'(t) - f'(s)|$ satisfies $\omega(f', \delta) \to 0$ as $\delta \to 0$. Let $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$ be a Rademacher vector and set

$$S_m = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \epsilon_j.$$

The idea is to use the Theorem 6.44, for Rademacher vectors, and the fact that S_m converges in distribution to a standard normal random variable by the Central Limit Theorem. Equation (6.21) yields

$$\mathcal{E}(f^2(S_m)) \leq \frac{1}{2}\mathbb{E}\Big[\sum_{i=1}^{n}(f(\boldsymbol{\epsilon}) - f(\overline{\boldsymbol{\epsilon}}^{(i)})^2\Big] = \frac{1}{2}\mathbb{E}\Big[\sum_{i=1}^{n}\Big(f(S_m) - f\Big(S_m - \frac{2\epsilon_i}{\sqrt{m}}\Big)\Big)^2\Big].$$

Notices that for each $i \in [m]$ we have

$$\begin{aligned}\Big|f(S_m) - f\Big(S_m - \frac{2\epsilon_i}{\sqrt{m}}\Big)\Big| &= \Big|\frac{2\epsilon_i}{\sqrt{m}} f'(S_m) + \int_{S_m - 2\epsilon_i/\sqrt{m}}^{S} (f'(t) - f'(S_m))dt\Big| \\ &\leq \frac{2}{\sqrt{m}}|f'(S_m)| + \frac{2}{\sqrt{m}}\omega\Big(f', \frac{2}{\sqrt{m}}\Big).\end{aligned}$$

It follows that

$$\sum_{i=1}^{m}\Big(f(S_m) - f\Big(S_m - \frac{2\epsilon_i}{\sqrt{m}}\Big)\Big)^2 \leq 4\Big(f'(S_m)^2 + 2|f'(S_m)|\omega\Big(f', \frac{2}{\sqrt{m}}\Big) + \omega\Big(f', \frac{2}{\sqrt{m}}\Big)^2\Big).$$

Using the boundedness of f and f', the central limit theorem implies that $\mathbb{E}[f'(S_m)^2] \to \mathbb{E}[f'(g)^2]$ and $\mathcal{E}[f'(S_m)^2] \to \mathcal{E}[f'(g)^2]$ as $m \to \infty$. Therefore, we conclude the inequality

$$\mathcal{E}(f^2(S_m)) \leq 2\mathbb{E}[f'(g)^2].$$

Now we turn to the general case where f does not have compact support. For a small $\varepsilon > 0$ given, the hypothesis $\mathbb{E}[\phi(f^2(\mathbf{g}))] < \infty$ ensure the existence of $T > 0$ such that for any subset I of $\mathbb{R}\backslash[-T, T]$,

$$\frac{1}{\sqrt{2\pi}}\int_I |\phi(f(t)^2)|e^{-t^2/2}dt \leq \varepsilon \qquad \text{and} \qquad \frac{1}{\sqrt{2\pi}}\int_I e^{-t^2/2}dt \leq \varepsilon.$$

Considering a continuously differentiable function h satisfying $0 \leq h(t) \leq 1$ for all $t \in \mathbb{R}$, $h(t) = 1$ for all $t \in [-T, T]$ and $h(t) = 0$ for $t \notin [-T - 1, T + 1]$. We then set $\tilde{f}(t) = f(t)h(t)$. This is a continuously differentiable function with compact support. The result just proved above applies to this new function and then gives

$$\mathcal{E}(\tilde{f}^2(g)) \leq 2\mathbb{E}[\tilde{f}'(g)^2]. \tag{6.23}$$

The subadditivity of the entropy, Corollary 6.41, provides

$$\mathcal{E}(f^2(g)) = \mathcal{E}\left(\tilde{f}^2(g) + f^2(g)(1 - h^2(g))\right) \leq \mathcal{E}(\tilde{f}^2(g)) + \mathcal{E}(f^2(g)(1 - h^2(g)).$$

Introducing the sets $I_1 = \{t \in \mathbb{R} : |t| \geq T, f(t)^2 < e\}$ and $I_2 = \{t \in \mathbb{R} : |t| \geq T, f(t)^2 \geq e\}$, we can estimate the second term of the right hand side after some calculations

$$\mathbb{E}[f^2(g)(1 - h^2(g))] = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}\backslash[-T,T]} f^2(t)(1 - h^2(t))e^{-t^2}dt \leq \frac{1}{\sqrt{2\pi}}\int_{I_1} f^2(t)e^{-t^2}dt + \frac{1}{\sqrt{2\pi}}\int_{I_2} f^2(t)e^{-t^2}dt$$

$$\leq \frac{e}{\sqrt{2\pi}} \int_{I_1} e^{-t^2} dt + \frac{1}{\sqrt{2\pi}} \int_{I_2} \phi(f^2(t)) e^{-t^2} dt \leq (e+1)\varepsilon. \tag{6.24}$$

For sufficiently small x, e.g., $x < 1/e$, the function $|\phi(x)| = |x \ln x|$ is increasing. Applying $|\phi|$ to both sides of Equation 6.24 leads to $\left|\phi\big(\mathbb{E}\big[f^2(g)\big(1-h^2(g)\big)\big]\big)\right| \leq \left|\phi((1+e)\varepsilon)\right|$ for a sufficiently small ε. With the aid of the auxiliary sets $I_3 = \{t \in \mathbb{R} : |t| \geq T, f^2(t)(1-h^2(t)) < e\}$ and $I_4 = \{t \in \mathbb{R} : |t| \geq T, f^2(t)(1-h^2(t)) \geq e\}$ and denoting $\kappa = \max_{t\in[0,e]} |\phi(t)| = e^{-1}$, we have

$$\left|\mathbb{E}[\phi(f^2(g)(1-h^2(g)))]\right| \leq \frac{\kappa}{\sqrt{2\pi}} \int_{I_3} e^{-t^2/2} dt + \frac{1}{\sqrt{2\pi}} \int_{I_4} \phi(f^2(g)(1-h^2(g))) e^{-t^2/2} dt \leq$$

$$\kappa\varepsilon + \frac{1}{\sqrt{2\pi}} \int_{I_4} \phi(f^2(g)) e^{-t^2/2} dt \leq (\kappa+1)\varepsilon.$$

The triangular inequality in the definition of entropy leads to

$$\mathcal{E}\big(f^2(g)(1-h^2(g))\big) \leq \left|\phi\big(\mathbb{E}\big[f^2(g)(1-h^2(g))\big]\big)\right| + \left|\mathbb{E}\big[\phi\big(f^2(g)(1-h^2(g))\big)\big]\right| \leq |\phi((e+1)\varepsilon)| + (\kappa+1)\varepsilon.$$

And using Equation (6.23), we derive $\mathcal{E}\big(f^2(g)\big) \leq \mathbb{E}[\tilde{f}'(g)^2] + (1+\kappa)\varepsilon + |\phi((1+e)\varepsilon)|$. Besides, applying again the triangular inequality, we obtain

$$\mathbb{E}[\tilde{f}'(g)^2]^{1/2} = \mathbb{E}[(f'h + fh')(g)^2]^{1/2} \leq \mathbb{E}[(f'h)(g)^2]^{1/2} + \mathbb{E}[(fh')(g)^2]^{1/2}.$$

Now it is time to analyze both terms of the right-hand side above. For the first one, we simple note that $\mathbb{E}[(f'h)(g)^2] = \mathbb{E}[f'(g)^2 h(g)^2] \leq \mathbb{E}[f'(g)^2]$. For the second one, using estimate (6.24), it holds that

$$\mathbb{E}[(fh')(g)^2] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(t)^2 h'(t)^2 e^{-t^2/2} dt \leq \frac{||h'||_\infty^2}{\sqrt{2\pi}} \int_{I_1 \cup I_2} f(t)^2 e^{-t^2/2} dt \leq (e+1)||h'||_\infty^2 \varepsilon. \tag{6.25}$$

When we put all these steps together, we conclude that $\mathbb{E}[\tilde{f}'(g)^2] \leq \mathbb{E}[f'(g)^2]^{1/2} + ||h'||_\infty((e+1)\varepsilon)^2$. Hence,

$$\mathcal{E}(f^2(g)) \leq 2\Big(\mathbb{E}[f'(g)^2]^{1/2} + ||h'||_\infty((e+1)\varepsilon)^{1/2}\Big)^2 + |\phi((1+e)\varepsilon)| + (\kappa+1)\varepsilon.$$

This result is valid for any $\varepsilon > 0$ and since $\lim_{t\to 0} \phi(t) = \phi(0) = 0$, we can conclude

$$\mathcal{E}(f^2(g)) \leq \mathbb{E}[f'(g)^2].$$

The next step is to go from the one dimensional case to the general case. This will be done via the tensorization inequality, Theorem 6.43. With its aid, we obtain

$$\mathcal{E}(f^2(g)) \leq \mathbb{E}\left[\sum_{i=1}^n \mathcal{E}_{g_i}(f^2(g))\right] \leq 2\mathbb{E}\left[\sum_{i=1}^n \left(\frac{\partial f}{\partial x_i}(g)\right)^2\right] = 2\mathbb{E}\big[||\nabla f(g)||_2^2\big].$$

□

The main tools are in our hands. After all this effort, we can prove the (optimal) celebrated concentration of measure inequality for Lipschitz functions.

Proof. (of Theorem 6.37): First of all, we will assume that the function f is differentiable. Since f is Lipschitz with constant L. by hypothesis, this implies $||\nabla f(x)||_2 \leq L$ for all $x \in \mathbb{R}^n$. Applying Theorem 6.45 to $e^{\theta f/2}$ leads to

$$\mathcal{E}(e^{\theta f(g)/2}) \leq 2\mathbb{E}\Big[\big|\big|\nabla e^{\theta f(g)/2}/2\big|\big|_2^2\Big] = \frac{\theta^2}{2}\mathbb{E}\Big[e^{\theta f(g)}||\nabla f(g)||_2^2\Big] \leq \frac{\theta^2 L^2}{2}\mathbb{E}\Big[e^{\theta f(g)}\Big]. \tag{6.26}$$

Here we stop for a moment in order to state an important comment. In the hypothesis of Theorem 6.37, we must have $\mathbb{E}[\phi(f^2(\mathbf{g}))] < \infty$. In our case, this translates to $\mathbb{E}[\phi(e^{2\theta f(g)})] < \infty$. This holds by the Lipschitz hypothesis, since $e^{2\theta f(g)} \leq e^{2\theta|f(0)|}e^{L\theta||x||_2}$ for all $x \in \mathbb{R}^n$. If one looks carefully, Inequality (6.26) is exactly what you would like to have to apply Herbst argument. This, in turn, implies

$$\mathbb{E}\big[e^{\theta(f(g)-\mathbb{E}[f(g)])}\big] \leq \exp\left(\theta \int_0^\theta t^{-2} g(t)dt\right) = \exp(\theta^2 L^2/2).$$

Using Markov's inequality, we conclude

$$\mathbb{P}(f(g) - \mathbb{E}[f(g)] \geq t) \leq \inf_{\theta>0} \exp(-\theta t)\mathbb{E}[e^{\theta(f(g)-\mathbb{E}[f(g)])}] \leq \inf_{\theta>0} e^{-\theta t+\theta^2 L^2/2} = e^{-t^2/(2L^2)}.$$

The last step follows because the infimum is attained at $\theta = t/L^2$. If we replace f by $-f$, we have the same bound. Therefore, a union bound argument yields

$$\mathbb{P}\Big(\Big|f(g) - \mathbb{E}\big[f(g)\big]\Big| \geq t\Big) \leq 2\exp\left(-\frac{t^2}{2L^2}\right),$$

for f differentiable. In the general case of not necessarily differentiable f, for each $\varepsilon > 0$, we can find a differentiable Lipschitz function g with the same Lipschitz constant such that $|f(x) - h(x)| \leq \varepsilon$ for all $x \in \mathbb{R}^n$. A proof of this classical analytical fact can be found at Appendix C of [Rauhut & Foucart '13].

$$\mathbb{P}(f(g) - \mathbb{E}[f(g)] \geq t) \leq \mathbb{P}(h(g) - \mathbb{E}[h(g)] \geq t - 2\varepsilon) \leq e^{-(t-2\varepsilon)^2/(4L^2)}.$$

But since $\varepsilon > 0$ can be made arbitrarily small, we have the result for general f. Again, replacing f by $-f$ and using a union bound, the theorem follows. □

This deep theorem is useful to estimate a lot of things. Also, since all norms in $\mathbb{R}^n$ are Lipschitz functions, we can use it to estimate geometrical notions like the Gaussian width. Here we provide a simple example.

Example 6.46. Let us understand the concentration of the χ^2-distribution. If $X_k \sim N(0,1)$ are n independent standard normal random variables, then $Y = \sum_{i=1}^n X_i^2$ will be a χ^2 random variable with n degrees of freedom (see section 8.2 of [DeGroot & Schervish '11]). We define the variable $Z = \sqrt{Y}/\sqrt{n} = ||(X_1, \ldots, X_n)||_2/\sqrt{n}$. Since the ℓ_2-norm is a 1-Lipschitz function, Theorem 6.37 implies that

$$\mathbb{P}\big(Z \geq \mathbb{E}(Z) + \delta\big) \leq e^{-n\delta^2/2}, \qquad \forall \delta \geq 0.$$

Since the square root function is concave, by Jensen's inequality we have

$$\mathbb{E}(Z) \leq \sqrt{\mathbb{E}(Z^2)} = \left(\frac{1}{n}\sum_{i=1}^n \mathbb{E}(X_i^2)\right)^{1/2} = 1.$$

Now we remember that $Z = Y/\sqrt{N}$ to conclude

$$\mathbb{P}\big(Y/n \geq (1+\delta)^2\big) \leq e^{-n\delta^2/2}, \qquad \forall \delta \geq 0.$$

Since $(1+\delta)^2 = 1 + 2\delta + \delta^2 \leq 1 + 3\delta$ for all $\delta \in [0,1]$, making the substitution $t = 3\delta$ leads to

$$\mathbb{P}\big(Y \geq n(1+t)\big) \leq e^{-nt^2/18}, \qquad \forall t \in [0,3].$$

The same kind of concentration inequality holds for suprema of Gaussian processes. Its proofs can be found at Theorem 5.8 of [Boucheron, Lugosi & Massart '13]. To close this section, we state it.

Theorem 6.47. *Let $(X_t)_{t\in T}$ be an almost surely continuous centered Gaussian process, that is, with probability 1, X_t is a continuous function of t, indexed by a totally bounded set T. If $\sigma^2 = \sup_{t\in T}\mathbb{E}[X_t^2]$, then $Z = \sup_{t\in T} X_t$ satisfies $Var(Z) \leq \sigma^2$, and for all $u > 0$, we have*

$$\mathbb{P}\Big(\big|Z - \mathbb{E}[Z]\big| \geq u\Big) \leq e^{-u^2/(2\sigma^2)}.$$

6.6 Covering and Packing Numbers

"Mit den Grenzen der Wissenschaften gegeneinander verhält es sich ähnlich, wie mit denen der Meere. Sie sind künstlich und aus praktischen Gründen angenommen. Alles Wissen steht miteinander im Zusammenhange, und eine Untersuchung abzubrechen, um nicht eine solche Grenze zu überschreiten, wäre nicht zu rechtfertigen.[5]"
Postcard from Gottlob Frege to Heinrich Rickert in 1 July 1911.

This section has, in a first moment, no clear relation with probability. Here we will define two important notions for metric spaces that will be used in subsequent chapters: covering numbers and packing numbers. However, in a second moment, these notions - together with probabilistic techniques - proved to be fundamental to demonstrate important theorems [Pisier '89]. Also, they play a fundamental role in understanding the behavior of stochastic processes. Usually one wants to infer some properties of a stochastic process X_t, with $t \in \Lambda$. Knowing the structure of the set Λ and one might ask the dependence of X_t on Λ. An example of such phenomenon is Dudley's Theorem in Stochastic Process.

Besides, in many situations, we want to manipulate and quickly compute some properties of a data set. This set is sometimes modeled as a (compact) metric space and this manipulation could be translated as the use of "sparse", typically discrete, object. The role of this object is to approximately capture the geometry or complexity of a (subset of) a metric space. It is in this context that the concepts of covering number and packing number arise. They are highly used, for example, in Error-Corrector Coding [Candès & Randall '08], Quantization [Boufonos '12] and Geometric Functional Analysis [Pisier '89].

Definition 6.48. Let X be a subset of a metric space (M, d). For $t > 0$, the *covering number* $N_\varepsilon = N(X, d, \varepsilon)$ is the smallest N such that X can be covered with balls of radius ε, that is, we cover X with balls $B_\varepsilon(x) = \{x \in M, d(x, x_i) \leq \varepsilon\}$. The logarithm of the covering number is called the *metric entropy* of the metric space X.

Definition 6.49. The *packing number* $P(X, d, \varepsilon)$ is defined, for $\varepsilon > 0$, as the maximal integer P such that there are points $x_i \in X$, $i = \{1, \ldots, P\}$ which are ε-separated, i.e., $d(x_i, x_k) > \varepsilon$ for all $i, k \in \{1, \ldots, P\}$, $i \neq k$.

Proposition 6.50. *We have some basic properties for the covering number. The packing number will also satisfy them.*

a.) For arbitrary sets $X, Y \subset M$, we have $N(X \cup Y, d, \varepsilon) \leq N(X, d, \varepsilon) + N(Y, d, \varepsilon)$.

b.) For any $\alpha > 0$, we have $N(X, \alpha d, \varepsilon) = N(X, d, \varepsilon/\alpha)$.

c.) If $M = \mathbb{R}^n$ and d is a metric induced by a norm $||.||$, then $N(\alpha X, \alpha d, \varepsilon) = N(X, d, \varepsilon/\alpha)$.

d.) If $\tilde{d}$ is another metric on M that satisfies $\tilde{d}(x, y) \leq d(x, y)$ for all $x, y \in X$, then

$$N(X, \tilde{d}, \varepsilon) \leq N(X, d, \varepsilon).$$

Proof. All of them follow from distance properties such as triangular inequality as well as the definition of the covering/packing number. □

The two notions are related, as the next theorem shows.

Theorem 6.51. *Let X be a subset of a metric space (M, d) and let $t > 0$. Then*

$$P(X, d, 2\varepsilon) \leq N(X, d, \varepsilon) \leq P(X, d, \varepsilon)$$

[5] In a free translation: *The boundaries between the disciplines of science are somewhat like those between the seas. They are artificial and adopted for practical purposes. All knowledge is interrelated, and it would be unjustifiable to abort an investigation in order not to cross such a boundary.* The fac simile version of the postcard and more details about it can be found at [Schlotter & Wehmeier '13].

Proof. In order to prove the right-hand side, let $P = P(X, d, \varepsilon)$. Then there exists P points $\{x_1, \dots, x_P\}$ that are ε-separated in X. By the maximality of P, it follows that any other $x \in X$ satisfies $d(x, x_i) \leq \varepsilon$ for some index i. Thus $X \subseteq \cup_{i=1}^{m} B_\varepsilon(x_i)$ and so $N(X, d, \varepsilon) \leq P(X, d, \varepsilon)$.

For the left-hand side, we just need to use the pigeonhole principle. Let $\overline{P} = P(X, d, 2\varepsilon)$. Then there exists $\overline{P}$ points $\{x_1, \dots, x_{\overline{P}}\} \subset X$ that are 2ε-separated, i.e., $d(x_i, x_j) > 2\varepsilon$ if, $i \neq j$. Suppose, by contradiction, that $\overline{P} > N(X, d, \varepsilon)$, then X is covered by fewer than $\overline{P}$ balls of radius ε. So, at least two of the $\overline{P}$ distinct points x_i and x_j must lie in the same ball $B_\varepsilon(\tilde{x})$ with a certain center $\tilde{x}$. By the triangle inequality, we have $d(x_i, x_j) \leq d(x_i, \tilde{x}) + d(x_j, \tilde{x})$ But this is a contradiction since x_i and x_j are 2ε-separated. Therefore $\overline{P} \leq N(X, d, \varepsilon)$.

□

The next question one might ask is how to calculate these quantities for certain sets on metric spaces. Usually this is a difficult problem and one can just provide some estimates instead of a closed formula. The set of notes [Wainwright '15] provides many examples and a nice explanation about these quantities and how to use them. Here, we state just one example that will be useful in the following chapters.

Proposition 6.52. *Let $||.||$ be some norm on $\mathbb{R}^n$ and let U be a subset of the unit ball $B = \{x \in \mathbb{R}^n, ||x|| \leq 1\}$. Then the packing and the covering numbers satisfy, for $\varepsilon > 0$,*

$$N(U, ||.||, \varepsilon) \leq P(U, ||.||, \varepsilon) \leq \left(1 + \frac{2}{\varepsilon}\right)^n.$$

In the case that $U = B$, we have the lower estimate $\left(\frac{1}{2\varepsilon}\right)^n \leq N(U, ||.||, \varepsilon)$.

Proof. The inequality $N(U, ||.||, \varepsilon) \leq P(U, ||.||, \varepsilon)$ is just Theorem 6.51. For the other inequality, let $\{x_1, \dots, x_P\} \subset U$ be a maximal ε-packing. This tells us that the balls $B(x_i, \varepsilon/2)$ have empty intersection and, even more, they are contained in the ball $(1 + \varepsilon/2)B$. What remains is to compare the Lebesgue measure (aka volume) of these balls.

$$\mathrm{vol}\left(\bigcup_{i=1}^{P} B(x_i, \varepsilon/2)\right) = \sum_{i=1}^{P} \mathrm{vol}(B(x_i, \varepsilon/2)) = P\mathrm{vol}((\varepsilon/2)B) \leq \mathrm{vol}\big((1 + \varepsilon/2)B\big).$$

Also, we have the homogeneity of volumes in $\mathbb{R}^n$ which tells us that $\mathrm{vol}(\varepsilon B) = \varepsilon^n \mathrm{vol}(B)$. Therefore we conclude that $P(\varepsilon/2)^n \mathrm{vol}(B) \leq (1 + \varepsilon/2)^n \mathrm{vol}(B)$. Dividing both sides by $(\varepsilon/2)^n \mathrm{vol}(B)$ leads to $P \leq (1 + 2/\varepsilon)^n$.
Now, for the case $U = B$, in view of Proposition 6.52, we just need to prove that if $\overline{P} = P(B, ||.||, \varepsilon)$, then $\overline{P} \geq (1/\varepsilon)^n$ since this implies in $N(B, ||.||, \varepsilon) \geq P(B, ||.||, 2\varepsilon) \geq (1/2\varepsilon)^n$. If $\overline{P} = P(B, ||.||, \varepsilon)$, there must exists $\overline{P}$ points $\{x_1, \dots, x_{\overline{P}}\}$ that are ε-separated in B. By the maximality of $\overline{P}$, it follows that any point $x \in B$ satisfies $d(x, x_i) \leq \varepsilon$ foor some i. Hence, $B \subset \cup_{j=1}^{\overline{P}} \mathrm{vol}(B_\varepsilon(x_j))$ and $\sum_{j=1}^{\overline{P}} \mathrm{vol}(B_\varepsilon(x_j)) \geq \mathrm{vol}(B)$. By the homogeneity of the volume, we have $\overline{P}\varepsilon^n \geq 1$. This leads to $\overline{P} \geq (1/\varepsilon)^n$ as we wanted to show. □

These definitions allow us to reduce the complexity of the calculations of norm operators, for example. The next theorem illustrates this fact. In order to calculate it, it is necessary to maximize over the whole sphere S^{n-1}. Using coverings, we replace by the maximum over a finite set.

Theorem 6.53. *Let A be an $N \times m$ matrix, and let N_ε be an ε-covering of S^{n-1} for some $\varepsilon \in [0, 1)$. Then*

$$\sup_{x \in N_\varepsilon} ||Ax||_2 \leq ||A||_{2\to 2} \leq (1 - \varepsilon)^{-1} \sup_{x \in N_\varepsilon} ||Ax||_2$$

In the case of a symmetric $n \times n$ matrix, we have

$$||A||_{2\to 2} = \sup_{x \in S^{n-1}} |\langle Ax, x\rangle| \leq (1 - 2\varepsilon)^{-1} \sup_{x \in N_\varepsilon} |\langle Ax, x\rangle|$$

Proof. In the first case, the lower bound follows from the definition of an operator's norm. For the upper bound, let us fix $x \in S^{n-1}$ which attains the norm, i.e.,$||A||_{2\to 2} = ||Ax||_2$. Now, choose y in the ε-covering which approximates x as $||x-y||_2 \leq \varepsilon$. Using the triangle inequality, we have $||Ax - Ay||_2 \leq ||A||_{2\to 2}||x-y||_2 \leq \varepsilon||A||_{2\to 2}$. It follows that

$$||Ay||_2 \geq ||Ax||_2 - ||Ax - Ay||_2 \geq ||Ax||_2 - ||A||_{2\to 2}||x-y||_2 \geq ||A||_{2\to 2} - \varepsilon||A||_{2\to 2} = (1-\varepsilon)||A||_{2\to 2}.$$

Taking the maximum over all y in the ε-covering in this inequality completes the proof. For the second statement, the proof is quite similar. Take $x \in S^{n-1}$ for which $||A||_{2\to 2} = |\langle Ax, x\rangle|$ and choose y in the ε-covering which approximates x as $||x-y||_2 \leq \varepsilon$. Again, by the triangle inequality we have

$$\begin{aligned} \big|\langle Ax, x\rangle - \langle Ay, y\rangle\big| &= \big|\langle Ax, x-y\rangle + \langle A(x-y), y\rangle\big| \\ \leq ||A|||_{2\to 2}||x||_2||x-y||_2 + ||A|||_{2\to 2}||y||_2||x-y||_2 &\leq 2\varepsilon||A|||_{2\to 2}. \end{aligned}$$

It follows that $|\langle Ay, y\rangle| \geq |\langle Ax, x\rangle| - 2\varepsilon||A|||_{2\to 2} = (1-2\varepsilon)||A|||_{2\to 2}$. Taking the maximum over all y in the ε-covering in this inequality yields the desired result. □

For generalizations related to the results of this section, one can consult Chapter 4 of [Pisier '89].

Chapter 7

Matrices Which Satisfy The RIP

The 8P Rule: Proper Prior Planning and Preparation Prevents Piss Poor Performance.
One of the many variations of a US Marine Corps adage

7.1 Introduction

In Chapter 5 we discussed the importance of having matrices that satisfy the restricted isometry property. For linear systems associated to these matrices, the search for sparse vectors is successful through various algorithms, such as BP, IHT, OMP, COSaMP, etc. As a consequence of Lemma 5.21 from [Cai, Wang & Xu I '10], we deduced, in the case of ℓ_1 minimization, that $\delta_{2s} < 0.6246$ suffices for a robust and stable reconstruction. However, there is little hope in improving this bound and make $\delta \to 1$, due to Theorem 5.27, proved by [Davies & Gribonval '09]. Nevertheless this quantity is very important since it can ensure the existence of matrices with the small (and optimal) number of measurements. This optimality will be explored through this chapter with the aid of probabilistic arguments.

However, as we saw in Section 4.10, so far we can only conclude the existence of matrices satisfying RIP with a lower bound in the number of rows given by $m \geq Cs^2$. Considering that the vectors we seek when solving the linear system have information only in s directions, it was expected that the number of measurements should vary linearly, and not quadratically, in s, the sparsity.

This is an indication that there could be another way of constructing matrices with δ_s small enough but, at the same time, with m significantly smaller than Cs^2. In this chapter will use the power of the probability theorems developed in Chapter 6, such as *concentration of measure* and *Gordon's Lemma*, to prove the existence of such matrices when $m \geq C_{\delta_s} s \ln(N/s)$. This impressive result will be proven to be *sharp* in Chapter 8. In the literature it is common to say thatt "m varying linearly with s up to a logarithmic factor."

Such kind of probabilistic argument has also played a central role in many fields. Random matrices and projections have been important tools in some fields like asymptotic theory of finite-dimensional normed spaces and convex geometry since the seventies. Milman, Schechtman, Szarek, Gluskin, Garnaev, Kashin, Rudelson and Vershynin can be mentioned in the development of these ideas. It is important to note that Rudelson and Vershynin introduced a lot of probabilistic techniques into the field of Compressive Sensing. For more on these, see [Schechtman '03], [Artstein-Avidan, Giannopoulos & Milman '15] and references therein.

Some ideas from these fields appear in modern Data Science techniques. We can cite the Johnson-Lindenstrauss Lemma, which essentially states that with high probability, Lipschitz projections do not disturb the geometry of a point cloud when they are projected onto a space of dimension logarithmic in the number of points. We will see that ideas from compressive sensing appear, in some hidden form, in the development of this profound theorem. It will be shown that there exists a deep connection between the restricted isometry property and dimensionality reduction through the JL Lemma.

Also, we can understand the idea behind the motto *random projections act as almost norm preserving*

for some subsets of the sphere, used nowadays in the Machine Learning community.

In this chapter, we introduce an ensemble of matrices for which the restricted isometry property holds with high probability and we show the importance of Johnson-Lindenstrauss' Lemma and its connections with the Restricted Isometry Property.

7.2 RIP for Subgaussian Ensembles

While developing the probabilistic tools on the last chapter, subgaussian random variables emerged as a generalization of Gaussian variables, preserving many good properties. We will now consider matrices $A \in \mathbb{R}^{m\times N}$ having random variables in each of its entries. These are called *random matrices* or a *random matrix ensemble*.

Definition 7.1. Let $A \in \mathbb{R}^{m\times N}$ be a random matrix

[i.] If the entries of A are independent Rademacher variables, A is called a *Bernoulli random matrix*

[ii.] If the entries of A are independent standard gaussian random variables, A is a *Gaussian random matrix*.

[iii.] If the entries of A are independent zero mean subgaussian random variables with variance 1 and same subgaussian parameters β and κ, *i.e.*,

$$\mathbb{P}(|A_{i,j}| \geq t) \leq \beta e^{-\kappa t^2} \qquad \forall t > 0, \quad i \in [m],\ j \in [N],$$

A is called a *subgaussian random matrix*.

We showed that Gaussian and Rademacher random variables are particular cases of subgaussian variables. The same holds for matrices: Gaussian and Bernoulli matrices are subgassian matrices. Besides, note that entries of a subgassian matrix do not have to be identically distributed.

We are interested in the RIP constant of A. Henceforth we will work with $\frac{1}{\sqrt{m}}A$ instead of A because

$$\mathbb{E}\left[\left|\left|\frac{1}{\sqrt{m}}Ax\right|\right|_2^2\right] = ||x||_2^2$$

for any vector x and a subgaussian matrix A (since all entries have variance 1). This is the same as saying that the expectation of the squared ℓ_2-norm of the columns of A is 1. In this case, the RIP constant δ_s is a measure of the deviation of $||\frac{1}{\sqrt{m}}Ax||_2^2$ from its mean, uniformly over all s-sparse vectors x.

Our main result in the first part of this chapter, due to [Baraniuk et al. '08] and the seminal work of [Mendelson, Pajor & Tomczak-Jaegermann '08], is the following.

Theorem 7.2. *Let $A \in \mathbb{R}^{m\times N}$ be a subgaussian random matrix and $\varepsilon > 0$. Suppose*

$$m \geq C\delta^{-2}(s\ln(eN/s) + \ln(2\varepsilon^{-1}))$$

for a constant $C > 0$ (depending only on the subgaussian parameters β, κ). Then, with probability at least $1-\varepsilon$, δ_s, the restricted isometry constant of $\frac{1}{\sqrt{m}}A$ satisfies $\delta_s \leq \delta$.

In fact we will prove a more general result. For this, we need some definitions.

Definition 7.3. Let Y be a random vector in $\mathbb{R}^N$.

[i.] If $\mathbb{E}\big[|\langle Y, x\rangle|^2\big] = ||x||^2 \quad \forall x \in \mathbb{R}^N$, then Y is called *isotropic*.

[ii.] If, $\forall x \in \mathbb{R}^N$ with $||x||_2 = 1$, the random variable $\langle Y, x\rangle$ is subgaussian with a subgaussian parameter c independent of x, that is,

$$\mathbb{E}\big[\exp(\theta\langle Y, x\rangle)\big] \leq \exp(c\theta^2), \quad \forall \theta \in \mathbb{R} \quad \text{and } \forall x \text{ with } ||x||_2 = 1,$$

then Y is called a *subgaussian random vector*.

We quickly realize that the constant c should ideally be independent of N. But note that if Y is a randomly selected row or column of an orthogonal matrix, then Y is isotropic and subgaussian but c has a dependence in N. We now state the more general result we will prove.

Theorem 7.4. *Let $A \in \mathbb{R}^{m\times N}$ be a random matrix with independent, isotropic and subgaussian rows with the same subgaussian parameter c in the definition above. If*

$$m \geq C\delta^{-2}(s\ln(eN/s) + \ln(2\varepsilon^{-1})),$$

where $C = 2(4\beta + 2\kappa)/3\kappa^2$ and β, κ are some parameters depending only on c, then with probability at least $1-\varepsilon$, the restricted isometry constant of $\frac{1}{\sqrt{m}}A$ satisfies $\delta_s \leq \delta$.

The proof of this theorem relies on a concentration inequality which, in turn, is a consequence of Bernstein's inequality for subexponential random variables, proved in the last chapter.

Lemma 7.5. *Let $A \in \mathbb{R}^{m\times N}$ be a random matrix with independent, isotropic, and subgaussian rows with the same subgaussian parameter c as in Definition 7.3. Then, $\forall x \in \mathbb{R}^N$ and $\forall t \in (0,1)$,*

$$\mathbb{P}\left[\left|m^{-1}||Ax||_2^2 - ||x||_2^2\right| \geq t||x||_2^2\right] \leq 2\exp(-Kt^2m)$$

where K depends only on c.

Proof. Let $x \in \mathbb{R}^N$. We may assume that $||x||_2 = 1$. Denoting the rows of A by $Y_1, \ldots, Y_m \in \mathbb{R}^N$ we can define the following random variables

$$Z_i = |\langle Y_i, x\rangle|^2 - ||x||_2^2, \quad i \in [m].$$

By hypothesis, Y_i is isotropic, so $\mathbb{E}[Z_i] = 0$. Moreover, as $\langle Y_i, x\rangle$ is subgaussian, Z_i is subexponential, that is, $\mathbb{P}(|Z_i| \geq r) \leq \beta\exp(-\kappa r)$ for all $r > 0$ and some parameters β, κ depending only on c. Now, after a simple calculation we obtain

$$\frac{1}{m}||Ax||_2^2 - ||x||_2^2 = \frac{1}{m}\sum_{i=1}^m (|\langle Y_i, x\rangle|^2 - ||x||_2^2) = \frac{1}{m}\sum_{i=1}^m Z_i.$$

Since the Y_i are independent, Z_i are also independent, it follows from Bernstein's inequality for subexponential random variables, Theorem 6.23, that

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^m Z_i\right| \geq t\right) = \mathbb{P}\left(\left|\sum_{i=1}^m Z_i\right| \geq tm\right) \leq 2\exp\left(-\frac{\kappa^2 m^2 t^2/2}{2\beta m + \kappa m t}\right) \leq 2\exp\left(-\frac{\kappa^2}{4\beta + 2\kappa}t^2 m\right),$$

where we used that $t \in (0,1)$ in the last step. Defining $K = \frac{\kappa^2}{4\beta+2\kappa}$, the concentration inequality follows. □

Now, in order to prove Theorem 7.4, we start by showing that a submatrix of a random matrix is well conditioned, with high probability, if we make some restriction on its size.

Theorem 7.6. *Suppose that a random matrix $A \in \mathbb{R}^{m\times N}$ is drawn according to a probability distribution for which the concentration inequality of the lemma above holds, that is, for $t \in (0,1)$,*

$$\mathbb{P}\left[\left|||Ax||_2^2 - ||x||_2^2\right| \geq t||x||_2^2\right] \leq 2\exp(-Kt^2m) \qquad \forall x \in \mathbb{R}^N.$$

Suppose also that for $\delta, \varepsilon \in (0,1)$ we have

$$m \geq C\delta^{-2}(7s + 2\ln(2\varepsilon^{-1}))$$

where $C = 2/(3K)$. Then, for $S \subset [N]$ with $\#S = s$, we have

$$\mathbb{P}(||A_S^* A_S - Id||_{2\to 2} < \delta) \geq 1 - \varepsilon$$

Proof. From the combinatorial arguments developed in Theorem 6.51 we know that that for $\rho \in (0, 1/2)$, there exists a finite subset U from the unit ball $B_S = \{x \in \mathbb{R}^N, \text{supp } x \subset S, ||x||_2 \leq 1\}$ which satisfies

$$\#U \leq \left(1 + \frac{2}{\rho}\right)^s \qquad \text{and} \qquad \min_{u \in U} ||z - u||_2 \leq \rho \quad \forall z \in B_S.$$

Now, using the concentration inequality of the hypothesis, we have, for $t \in (0, 1)$ and δ and ρ (they will be determined later)

$$\mathbb{P}\left[\left|||Au||_2^2 - ||u||_2^2\right| \geq t||u||_2^2 \quad \text{for some } u \in U\right] \leq \sum_{u \in U} \mathbb{P}\left[\left||Au||_2^2 - ||u||_2^2\right| \geq t||u||_2^2\right]$$

$$\leq 2\#U \exp(-Kt^2m) \leq 2\left(1 + \frac{2}{\rho}\right)^s \exp(-Kt^2m).$$

Suppose, indeed, that the realization of the random matrix A yields the opposite inequality

$$\left|||Au||_2^2 - ||u||_2^2\right| < t||u||_2^2 \qquad \forall u \in U. \tag{7.1}$$

The calculations above showed that

$$\mathbb{P}\left(\left|||Au||_2^2 - ||u||_2^2\right| < t||u||_2^2 \quad \forall\, u \in U\right) \geq 1 - 2\left(1 + \frac{2}{\rho}\right)^s \exp(-Kt^2m). \tag{7.2}$$

We are going to show that, for the right choice of ρ and t, (7.1) implies $\left|||Ax||_2^2 - ||x||_2^2\right| \leq \delta$ for all $x \in B_S$, i.e., the well conditioning of the submatrices, $||A_S^*A_S - Id||_{2\to 2} \leq \delta$. Defining $B = A_S^*A_S - \text{Id}$, the equation (7.1) is the same as $|\langle Bu, u\rangle| < t \quad \forall u \in U$. This happens because

$$\left|||Au||_2^2 - ||u||_2^2\right| < t||u||_2^2 \implies |\langle Au, Au\rangle - \langle u, u\rangle| < t||u||_2^2$$

$$\implies |\langle A^*Au, u\rangle - \langle u, u\rangle| < t||u||_2^2 \implies |\langle (A^*A - Id)u, u\rangle| < t||u||_2^2 \implies |\langle Bu, u\rangle| < t||u||_2^2.$$

Consider a vector $x \in B_S$ and choose $u \in U$ such that $||x - u||_2 \leq \rho < 1/2$. We obtain

$$|\langle Bx, x\rangle| = |\langle Bu, u\rangle + \langle B(x + u), x - u\rangle| \leq |\langle Bu, u\rangle| + |\langle B(x + u), x - u\rangle|$$

$$< t + ||B||_{2\to 2}||x + u||_2||x - u||_2 \leq t + 2\rho||B||_{2\to 2}.$$

Taking the maximum over all $x \in B_S$, we see that

$$||B||_{2\to 2} < t + 2\rho||B||_{2\to 2}, \qquad \text{i.e.} \qquad ||B||_{2\to 2} \leq \frac{t}{1 - 2\rho}.$$

We would like that $||B||_{2\to 2} < \delta$, so with the choice $t = (1 - 2\rho)\delta$ the aid of (7.2), we conclude

$$\mathbb{P}\left(||A_S^*A_S - Id||_{2\to 2} \leq \delta\right) \geq 1 - 2\left(1 + \frac{2}{\rho}\right)^s \exp(-K(1 - 2\rho)^2\delta^2m). \tag{7.3}$$

Now, this probability is at least $1 - \varepsilon$ if

$$\varepsilon \geq 2\left(1 + \frac{2}{\rho}\right)^s \exp(-K(1 - 2\rho)^2\delta^2m).$$

And this is the same as

$$m \geq \frac{1}{K(1 - 2\rho)^2}\delta^{-2}(\ln(1 + 2/\rho)s + \ln(2\varepsilon^{-1})).$$

To finish, we need to choose the value of ρ. Taking $\rho = 2/(e^{7/2} - 1) \approx 0.0623$ so that $1/(1-2\rho)^2 \leq 4/3$ and $\ln(1 + 2/\rho)/(1 - 2\rho)^2 \leq 14/3$, the inequality above will be satisfied if

$$m \geq \frac{2}{3K}\delta^{-2}(7s + 2\ln(2\varepsilon^{-1})).$$

This estimative concludes the proof.

□

What remains to be done is to complete the gaps. We first need to show that subgaussian matrices have isotropic and subgaussian rows. After this, we have to prove that matrices drawn from a distribution that satisfies the concentration inequality have small RIP constant of optimal order. For the first part the following lemma suffices.

Lemma 7.7. *Let $Y \in \mathbb{R}^N$ be a random vector with independent, zero mean and subgaussian entries with variance 1 and same subgaussian parameter c. Then Y is an isotropic and subgaussian random vector with the same subgaussian parameter c.*

Proof. Let $x \in \mathbb{R}^N$ with $||x||_2 = 1$. Since Y_i are independent, zero mean and have unit variance, we have

$$\mathbb{E}\big[|\langle Y, x\rangle|^2\big] = \sum_{i=1}^{N}\sum_{i'=1}^{N} x_i x_{i'}\mathbb{E}\big[Y_i Y_{i'}\big] = \sum_{i=1}^{N} x_i^2 = ||x||_2^2.$$

This allows us to conclude that Y is isotropic. We need to prove that $Z = \langle Y, x\rangle = \sum_{i=1}^{N} x_i Y_i$ is subgaussian. By the hypothesis of independence, we have

$$\begin{aligned}\mathbb{E}\big[\exp(Z\theta)\big] &= \mathbb{E}\left[\exp\left(\theta\sum_{i=1}^{N} x_i Y_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{N}\exp(\theta x_i Y_i)\right] = \prod_{i=1}^{N}\mathbb{E}\big[\exp(\theta x_i Y_i)\big] \\ &\leq \prod_{i=1}^{N}\exp(c\theta^2 x_i^2) = \exp(c||x||_2^2\theta^2).\end{aligned}$$

Hence Y is a subgaussian random vector with parameter independent of N.

□

Finally, we prove that matrices which satisfy the concentration inequality have small RIP of optimal order in the number of parameters.

Theorem 7.8. *Suppose that a random matrix $A \in \mathbb{R}^{m\times N}$ is drawn according to a probability distribution for which the concentration inequality holds, that is, for $t \in (0,1)$,*

$$\mathbb{P}\left[\Big|||Ax||_2^2 - ||x||_2^2\Big| \geq t||x||_2^2\right] \leq 2\exp(-Kt^2 m), \qquad \forall x \in \mathbb{R}^N.$$

Suppose further that, for $\delta, \varepsilon \in (0,1)$, the number of rows (measurements) satisfies

$$m \geq C\delta^{-2}\big[s\big(9 + 2\ln(N/s)\big) + 2\ln(2\varepsilon^{-1})\big],$$

where $C = 2/(3K)$. Then with probability $1-\varepsilon$, the restricted isometry constant δ_s of A satisfies $\delta_s < \delta$.

Proof. We start by recalling the equivalent definition of the RIC, i.e. $\delta_s = \sup_{S\subset[N],\ \#S=s} ||A_S^* A_S - Id||_{2\to 2}$. So we must analyze all subsets of $[N]$ with fixed cardinality s. By taking the union bound in all $\binom{N}{s}$ subsets $S \subset [N]$ of cardinality s we obtain

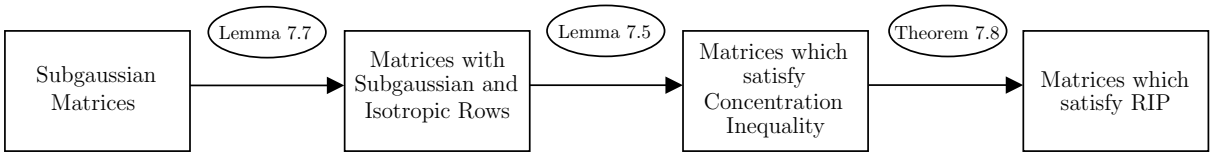
$$\begin{aligned}\mathbb{P}(\delta_s \geq \delta) &\leq \sum_{S\subset[N],\ \#S=s}\mathbb{P}\Big[||A_S^* A_S - Id||_{2\to 2} \geq \delta\Big] \leq 2\binom{N}{s}\left(1 + \frac{2}{\rho}\right)^s \exp(-K\delta^2(1-2\rho)^2 m) \\ &\leq 2\left(\frac{eN}{s}\right)^s\left(1 + \frac{2}{\rho}\right)^s \exp(-K\delta^2(1-2\rho)^2 m).\end{aligned}$$

Choosing $\rho = 2/(e^{7/2} - 1) \approx 0.0623$ as in Theorem 7.6 yields $\mathbb{P}(\delta_s < \delta) \geq 1 - \varepsilon$ if

$$m \geq \frac{1}{C\delta^2}\left(\frac{4}{3}s\ln(eN/s) + \frac{14}{3}s + \frac{4}{3}\ln(2\varepsilon^{-1})\right) = \frac{2}{3K}\delta^{-2}[s(9 + 2\ln(N/s)) + 2\ln(2\varepsilon^{-1})].$$

□

The diagram below shows the chain of ideas developed in this section:



In all of the previous Theorems above, we can choose the more convenient value of ε. In particular, if we set $\varepsilon = 2\exp(-m/2C_2)$ in Theorem 7.8, with probability $1 - 2\exp(-m/2C_2)$ *all s-sparse vectors* will be recovered via ℓ_1-minimization using an $m \times N$ subgaussian random matrix for which

$$m \geq 2C_1 s\ln(eN/s)$$

So, at least theoretically, instead of using matrices built with high precision to recover sparse information, some coins with a certain subgaussian distribution can be drawn to be used as the entries of the matrix. This will recover sparse vectors with high probability. Albeit this approach with subgaussian variables seems very promising and many papers written by mathematicians fully establish this as the *status quo* in sparse recovery, in practical terms the implementation is more subtle. This is currently a very active research area, see for example [Haboba et al. '12], where the relative strengths and weaknesses of hardware implementation of Gaussian and Bernoulli circuits is discussed.

7.3 Universal Recovery

Usually we need to find the right basis to represent signals in a sparse way. But not all sparsity phenomena occur with respect to the canonical basis. In many situations we need some other (orthonormal) basis such as discrete Fourier, wavelets, shearlets, curvelets, etc to achieve a compact representation of the signals. In mathematical terms, this means that the phenomenon we are studying is written as a vector $z = Ux$ with a orthogonal $N \times N$ matrix and a s-sparse vector $x \in \mathbb{C}^N$. So the act of taking measurements is represented by

$$y = Az = AUx.$$

In order to recover z, we need to recover first x and then make $z = Ux$. In practical terms, this means that the compressive sensing problem has a measurement matrix of the form $\tilde{A} = AU$. The matrix A, in this context, is a random $m \times N$ matrix while the one changing basis is a fixed and deterministic matrix $U \in \mathbb{R}^{N\times N}$. Then we need to adapt the results concerning small RIP constant to this more realistic situation. This is easily made in the corollary below.

Corollary 7.9. *Let $U \in \mathbb{R}^{N\times N}$ be a fixed orthogonal matrix. Suppose that an $m \times N$ random matrix A is drawn according to a probabilistic distribution for which the concentration inequality*

$$\mathbb{P}\left[\left|||Ax||_2^2 - ||x||_2^2\right| \geq t||x||_2^2\right] \leq 2\exp(-Kt^2m) \qquad \forall x \in \mathbb{R}^N$$

holds for all $t \in (0,1)$ and $x \in \mathbb{R}^N$. Suppose also that for a given $\delta, \varepsilon \in (0,1)$ we have

$$m \geq C\delta^{-2}[s(9 + 2\ln(N/s)) + 2\ln(2\varepsilon^{-1})],$$

for $C = 2/(3K)$. *Then, with probability* $1 - \varepsilon$, *the restricted isometry constant* δ_s *of* UA *satisfies* $\delta_s < \delta$.

Proof. As U is an orthogonal matrix, the concentration inequality holds with A replaced by UA. Let $x \in \mathbb{R}^N$ and $\tilde{x} = Ux$. We have

$$\mathbb{P}\Big(\big|||AUx||_2^2 - ||x||_2^2\big| \geq t||x||_2^2\Big) = \mathbb{P}\Big(\big|||A\tilde{x}||_2^2 - ||\tilde{x}||_2^2\big| \geq t||\tilde{x}||_2^2\Big) \leq 2\exp(-Kt^2m).$$

□

One can see that the matrix U in this theorem is arbitrary. So we say that sparse recovery with subgaussian matrices is a *universal* phenomenon regarding the orthogonal basis where signals are represented in a sparse way. The implication is that since the measurements are taken in the form $y = AUx$, one does not need to know the matrix U in advance to perform the sensing (enconding) stage. The knowledge of U is necessary only when applying the recovery techniques at the decoding stage.

The theorem proved in this section only states that for a fixed matrix U, a random choice of A will work with high probability. Universality does not mean that one single matrix measurement matrix A can recover all vectors represented in a sparse way on any basis since all vectors are 1-sparse in some basis. Of course that for every measurement scheme, we can construct a matrix of basis transform U such that the recover will fail.

Universality is a very interesting feature which occurs not only here but permeates all of the probability theory. The excellent post [Tao's Blog - 09/14/2010] gives a glimpse of this search for universality in a very didactic way.

7.4 The Curious Case of Gaussian Matrices

"Our lives are defined by opportunities, even the ones we miss."
Eric Roth in The Curious Case of Benjamin Button by F. Scott Fitzgerald

From a mathematical point of view, the original approach of Candes and Tao in [Candès & Tao I '06] was very different from the one shown, relying on the condition number estimate for Gaussian matrices. Since it uses tools as the *Gaussian concentration of measure* and *Slepian-Gordon lemma*, Theorem 6.37 and Theorem 6.30 respectively, it has the disadvantage of working *only* for this particular set of matrices. On the other hand, it has the advantage of providing better estimates due to a harder analysis.

It is of major importance to find better constants and thereby find the suitable balance between the number of measurements and dimension of the signal. The knowledge of these scales is necessary to apply the techniques of compressive sensing.

To prove the theorem which asserts that Gaussian matrices have small RIP constant with high probability we need some tools from the area of *non-asymptotic theory of random matrices*, especially the estimates for the extremal singular values of Gaussian random matrices. For a great overview of this area, see [Rudelson & Vershynin '10] and [Rudelson '14]. In order to prove these estimates, we need the following two lemmas.

Lemma 7.10. *For all matrices* A *and* B *in* $\mathbb{R}^{n\times n}$, *their smallest and largest singular values* σ_{min} *and* σ_{max} *satisfy*

$$|\sigma_{max}(A)-\sigma_{max}(B)| \leq ||A-B||_{2\to 2} \leq ||A-B||_F \quad \textit{and} \quad |\sigma_{min}(A)-\sigma_{min}(B)| \leq ||A-B||_{2\to 2} \leq ||A-B||_F.$$

Proof. For the largest singular value, we just need to identify it with the operator norm:

$$|\sigma_{\max}(A) - \sigma_{\min}(B)| = \big|||A||_{2\to 2} - ||B||_{2\to 2}\big| \leq ||A - B||_{2\to 2}.$$

For the smallest singular vales, we have:

$$\sigma_{\min}(A) = \inf_{||x||_2=1} ||Ax||_2 \leq \inf_{||x||_2=1} \big(||Bx||_2 + ||(A-B)x||_2\big)$$
$$\leq \inf_{||x||_2=1} \big(||Bx||_2 + ||(A-B)||_{2\to 2}\big) = \sigma_{\min}(B) + ||A-B||_{2\to 2}.$$

Therefore, we have $\sigma_{\min}(A) - \sigma_{\max}(B) \leq ||A-B||_{2\to 2}$ and the lemma follows from the symmetry between A and B. The second inequality is just the domination of the operator norm by the Frobenius norm. □

Lemma 7.11. *For integers $n \geq s \geq 1$,*

$$\sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)} - \sqrt{2}\frac{\Gamma((s+1)/2)}{\Gamma(s/2)} \geq \sqrt{n} - \sqrt{s}.$$

Proof. It suffices to show that, for any $n \geq 1$,

$$\alpha_{n+1} - \alpha_n \geq \sqrt{n+1} - \sqrt{n} \qquad \text{where} \quad \alpha_n = \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}.$$

If follows from $\Gamma((n+2)/2) = (n/2)\Gamma(n/2)$ that $\alpha_{n+1}\alpha_n = n$. Then, multiplying the inequality we want to prove by α_n and rearranging, we see that we need to prove that

$$\alpha_m^2 + (\sqrt{m+1} - \sqrt{m})\alpha_m - m \leq 0.$$

So we need to prove that α_m does not exceed the positive root of the polynomial $z^2 + (\sqrt{m+1} - \sqrt{m})z - m$, *i.e.* that

$$\alpha_m \leq \beta_m = \frac{-(\sqrt{m+1} - \sqrt{m}) + \sqrt{(\sqrt{m+1} - \sqrt{m})^2 + 4m}}{2}.$$

First, we will bound α_m from above and then bound β_m from below.. For the first part, we will use the Gauss hypergeometric theorem (see Theorem 2.2.2 of [Andrews, Askey & Roy '99]), from which we obtain that if $\mathrm{Re}(c-a-b) > 0$, then

$$\frac{\Gamma(c)\Gamma(c-a-b)}{\Gamma(c-a)\Gamma(c-b)} = {}_2F_1(a,b;c;1),$$

where ${}_2F_1(a,b;c;z)$ denotes the Gauss hypergeometric function

$${}_2F_1(a,b;c;z) = \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}.$$

and $(x)_n$ is the *Pochhammer symbol* defined by $(x)_0 = 1$ and $(x)_n = x(x+1)\ldots(x+n-1)$ for $n \geq 1$. Hence, choosing $a = -1/2$, $b = -1/2$ and $c = (m-1)/2$, we derive

$$\alpha_m^2 = 2\frac{\Gamma((m+1)/2)^2}{\Gamma(m/2)^2} = (m-1)\frac{\Gamma((m+1)/2)\Gamma((m-1)/2)}{\Gamma(m/2)^2}$$
$$= (m-1)\sum_{n=0}^{\infty} \frac{\big((-1/2)(1/2)\ldots(-1/2+n-1)\big)^2}{((m-1)/2)((m+1)/2)\ldots((m+2n-3)/2)} \frac{1}{n!}$$
$$= m - \frac{1}{2} + \frac{1}{8}\frac{1}{m+1} + \sum_{n=3}^{\infty} \frac{2^{n-2}\big((1/2)\ldots(n-3/2)\big)^2}{n!(m+1)(m+3)\ldots(m+2n-3)}.$$

Defining γ_m as

$$\gamma_m := (m+1)\left(\alpha_m^2 - m + \frac{1}{2} - \frac{1}{8}\frac{1}{m+1}\right) = \sum_{n=3}^{\infty} \frac{2^{n-2}\big((1/2)\ldots(n-3/2)\big)^2}{n!(m+3)\ldots(m+2n-3)},$$

we see that γ_m decreases with m. Thus, there is an integer m_0 (that will be chosen later) such that

$$\alpha_m^2 \leq m - \frac{1}{2} + \frac{1}{8}\frac{1}{m+1} + \frac{\gamma_{m_0}}{m+1} \qquad m \geq m_0. \tag{7.4}$$

Now, let us bound β_m from below. Let

$$\beta_m \geq \frac{-(\sqrt{m+1}-\sqrt{m})+\sqrt{4m}}{2} = \frac{3\sqrt{m}-\sqrt{m-1}}{2} = \delta_m.$$

Expanding $\sqrt{m(m+1)}$ and using that $\delta_m^2 = (10m+1-6\sqrt{m(m+1)})/4$, we have

$$\sqrt{m(m+1)} = (m+1)\sqrt{1-\frac{1}{m+1}} = (m+1)\left(1+\sum_{n=1}^{\infty}\frac{(1/2)(-1/2)\dots(1/2-n+1)}{n!}\left(\frac{-1}{m+1}\right)^n\right)$$

$$= m + \frac{1}{2} - \frac{1}{8}\frac{1}{m+1} - \frac{1}{2}\sum_{n=3}^{\infty}\frac{(1/2)\dots(1/2-n+1)}{n!}\left(\frac{-1}{m+1}\right)^{n-1}.$$

Therefore

$$\delta_m^2 = \frac{10m+1}{4} - \frac{3}{2}\sqrt{m(m+1)} = \frac{10m+1}{4} - \frac{3}{2}\left(m+\frac{1}{2}-\frac{1}{8}\frac{1}{m+1}-\frac{1}{2}\sum_{n=3}^{\infty}\frac{(1/2)\dots(1/2-n+1)}{n!}\left(\frac{-1}{m+1}\right)^{n-1}\right).$$

Thus

$$\beta_m^2 \geq \delta_m^2 \geq m - \frac{1}{2} + \frac{3}{16}\frac{1}{m+1}. \tag{7.5}$$

Subtracting (7.4) from

$$\beta_m^2 - \alpha_m^2 \geq \frac{1}{16}\frac{1}{m+1} - \frac{\gamma_{m_0}}{m+1} \qquad m \geq m_0.$$

To finish, we choose m_0 to be the smallest integer such that $\gamma_{m_0} \leq 1/16$, that is, $m_0 = 3$, so that $\beta_m^2 \geq \alpha_m^2$ for all $m \geq 3$. The numerical cases $m = 1$ and $m = 2$ are straightforward to verify.

□

We are now able to prove the *non-asymptotic* estimates of the singular values of a Gaussian matrix. This result is a consequence of estimates due to Gordon and Slepian[1] and was obtained by [Ledoux '01].

Theorem 7.12. *Let $A \in \mathbb{R}^{m\times s}$ be a Gaussian matrix with $m > s$ and let σ_{min} and σ_{max} be the smallest and largest singular values of the renormalized matrix $\frac{1}{\sqrt{m}}A$. Then, for $t > 0$,*

$$\mathbb{P}(\sigma_{max} \geq 1 + \sqrt{s/m} + t) \leq e^{-mt^2/2} \quad \textit{and} \quad \mathbb{P}(\sigma_{min} \leq 1 - \sqrt{s/m} - t) \leq e^{-mt^2/2}.$$

Proof. Let us start by using Lemma 7.10 and by noticing that the extremal singular values are 1-Lipschitz functions with respect to the Frobenius norm. So, by the concentration of measure inequality, Theorem 6.37, we have the following relations between $\sigma_{\max}$ and its expected value,

$$\mathbb{P}\big(\sigma_{\max} \geq \mathbb{E}[\sigma_{\max}] + r\big) \leq e^{-r^2/2}. \tag{7.6}$$

It remains to estimate the expected value above. For this, we use Slepian-Gordon's Lemma, Theorem 6.30. As

[1] *Asymptotic* estimates are know since the eighties, see [Geman '80] and [Yin, Bai & Krishnaiah '88] for $\sigma_{\max}$ and [Silverstein '85] for $\sigma_{\min}$. Also, see [Marcenko & Pastur '67].

$$\sigma_{\max} = \sup_{x\in S^{s-1}} \sup_{y\in S^{m-1}} \langle Ax, y\rangle,$$

we define the following two special Gaussian processes

$$X_{x,y} = \langle Ax, y\rangle \quad \text{and} \quad Y_{x,y} = \langle g, x\rangle + \langle h, y\rangle,$$

where $g \in \mathbb{R}^s$ and $h \in \mathbb{R}^m$ are two independent standard Gaussian vectors. Now we will compare these two processes. Taking $x, \tilde{x} \in S^{s-1}$ and $y, \tilde{y} \in S^{m-1}$ and using that A_{ij} are independent and of variance 1, we obtain

$$\mathbb{E}\big|X_{x,y} - X_{\tilde{x},\tilde{y}}\big|^2 = \mathbb{E}\Big|\sum_{i=1}^m\sum_{j=1}^s A_{ij}(x_jy_i - \tilde{x}_j\tilde{y}_i)\Big|^2 = \sum_{i=1}^m\sum_{j=1}^s (x_jy_i - \tilde{x}_j\tilde{y}_i)^2 = \sum_{i=1}^m\sum_{j=1}^s (x_j^2y_i^2 - 2x_j\tilde{x}_jy_i\tilde{y}_i + \tilde{x}_j^2\tilde{y}_i^2)$$

$$= ||x||_2^2||y||_2^2 + ||\tilde{x}||_2^2||\tilde{y}||_2^2 - 2\langle x, \tilde{x}\rangle\langle y, \tilde{y}\rangle = 2 - 2\langle x, \tilde{x}\rangle\langle y, \tilde{y}\rangle.$$

Using independence and isotropy of the standard multivariate Gaussian, we have

$$\mathbb{E}\big|Y_{x,y} - Y_{\tilde{x},\tilde{y}}\big|^2 = \mathbb{E}|\langle g, x - \tilde{x}\rangle + \langle h, y - \tilde{y}\rangle|^2 = \mathbb{E}|\langle g, x - \tilde{x}\rangle|^2 + \mathbb{E}|\langle h, y - \tilde{y}\rangle|^2 = ||x - \tilde{x}||_2^2 + ||y - \tilde{y}||_2^2$$

$$= ||x||_2^2 + ||\tilde{x}||_2^2 - 2\langle x, \tilde{x}\rangle + ||y||_2^2 + ||\tilde{y}||_2^2 - 2\langle y, \tilde{y}\rangle = 4 - 2\langle x, \tilde{x}\rangle - 2\langle y, \tilde{y}\rangle.$$

Therefore we conclude that

$$\mathbb{E}\big|Y_{x,y} - Y_{\tilde{x},\tilde{y}}\big|^2 - \mathbb{E}\big|X_{x,y} - X_{\tilde{x},\tilde{y}}\big|^2 = 2(1 - \langle x, \tilde{x}\rangle - \langle y, \tilde{y}\rangle + \langle x, \tilde{x}\rangle\langle y, \tilde{y}\rangle) = 2(1 - \langle x, \tilde{x}\rangle)(1 - \langle y, \tilde{y}\rangle) \geq 0.$$

In this last inequality we used Cauchy-Schwarz inequality and the fact that the vectors belong to the unit sphere (so equality holds in and only if $\langle y, \tilde{y}\rangle = 1$ or $\langle x, \tilde{x}\rangle = 1$). Thus, we showed that

$$\mathbb{E}\big|X_{x,y} - X_{\tilde{x},\tilde{y}}\big|^2 \leq \mathbb{E}\big|Y_{x,y} - Y_{\tilde{x},\tilde{y}}\big|^2.$$

Then, *Slepian-Gordon's Lemma* implies that

$$\mathbb{E}\sigma_{\max} = \mathbb{E}\sup_{x\in S^{s-1}}\sup_{y\in S^{m-1}} X_{x,y} \leq \mathbb{E}\sup_{x\in S^{s-1}}\sup_{y\in S^{m-1}} Y_{x,y} = \mathbb{E}\sup_{x\in S^{s-1}}\langle g, x\rangle + \mathbb{E}\sup_{y\in S^{m-1}}\langle h, y\rangle$$

$$= \mathbb{E}||g||_2 + \mathbb{E}||h||_2 \leq \sqrt{\mathbb{E}||g||_2^2} + \sqrt{\mathbb{E}||h||_2^2} \leq \sqrt{s} + \sqrt{m},$$

where we used Jensen's inequality. Using this in (7.6) we obtain that

$$\mathbb{P}\big(\sigma_{\max}(A) \geq \sqrt{s} + \sqrt{m} + r\big) \leq e^{-r^2/2}.$$

To finish, we just need to rescale A by $\frac{1}{\sqrt{m}}$. This gives us a estimate for the largest singular value of $\frac{A}{\sqrt{m}}$. The estimate for $\sigma_{\min}(A) = \inf_{x\in S^{s-1}} ||Ax||_2$ is more intricate. We need to measure the Gaussian width of S^{s-1} and then use *Gordon's escape through the mesh*, Theorem 6.32. Recall that for a given standard Gaussian vector $g \in \mathbb{R}^s$ we have

$$\ell(S^{s-1}) = \mathbb{E}\Big[\sup_{x\in S^{s-1}}\langle g, x\rangle\Big] = \mathbb{E}||g||_2.$$

If the entries of g are independent, then $||g||_2^2$ has χ^2 distribution with n degrees of freedom. Then

$$\mathbb{E}||g||_2 = \mathbb{E}\Big(\sum_{i=1}^n g_i^2\Big)^{1/2} = \frac{1}{2^{n/2}\Gamma(n/2)}\int_0^\infty u^{1/2}u^{(n/2)-1}e^{-u/2}du = \frac{2^{(n+1)/2}}{2^{n/2}\Gamma(n/2)}\int_0^\infty t^{(n/2)-1/2}e^{-t}dt$$

$$= \sqrt{2}\frac{\Gamma((n+1)/2)}{\Gamma(n/2)}.$$

Let us recall that for $g \in \mathbb{R}^s$ we introduced the notation $\mathbb{E}||g||_2 = E_s$. Then, with the aid of Lemma 7.11, we obtain

$$E_m - \ell(S^{s-1}) = E_m - E_s = \sqrt{2}\frac{\Gamma((m+1)/2)}{\Gamma(m/2)} - \sqrt{2}\frac{\Gamma((s+1)/2)}{\Gamma(s/2)} \geq \sqrt{m} - \sqrt{s}.$$

To conclude, using Theorem 6.32 we obtain

$$\mathbb{P}\Big(\sigma_{\min}(A) \leq \sqrt{m} - \sqrt{s} - t\Big) \leq \mathbb{P}\Big(\inf_{x \in S^{s-1}} ||Ax||_2 \leq E_m - \ell(S^{s-1}) - t\Big) \leq e^{-t^2/2}.$$

Finally, one just needs to rescale A by $\frac{1}{\sqrt{m}}$ to finish the proof.

□

We can state the main Theorem of this section.

Theorem 7.13. *Let $A \in \mathbb{R}^{m \times N}$ be a Gaussian matrix with $m < N$. Let $\eta, \varepsilon \in (0,1)$, and assume that*

$$m \geq 2\eta^{-2}\big(s\ln(eN/s) + \ln(2\varepsilon^{-1})\big).$$

Then we have

$$\mathbb{P}\left[\delta_s\left(\frac{1}{\sqrt{m}}A\right) \leq 2\left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)\eta + \left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)^2\eta^2\right] \geq 1 - \varepsilon.$$

Proof. The proof is very similar to the one for Theorem 7.8. Let $S \subset [N]$ such that $\#S = s$. The submatrix A_S is an $m \times s$ Gaussian matrix and the eigenvalues of $\frac{1}{m}A_S^*A_S - \mathrm{Id}$ are inside the interval $[\sigma_{\min}^2 - 1, \sigma_{\max}^2 - 1]$ where $\sigma_{\min}$ and $\sigma_{\max}$ are the extremal singular values of $\frac{1}{\sqrt{m}}A_S$. From Theorem 7.12, we obtain

$$\mathbb{P}\left(\left|\left|\frac{1}{m}A_S^*A_S - \mathrm{Id}\right|\right|_{2\to 2} \leq \max\left\{\left(1 + \sqrt{s/m} + \eta\right)^2 - 1, \left(1 - (\sqrt{s/m} + \eta)\right)^2\right\}\right)$$

$$= \mathbb{P}\left(\left|\left|\frac{1}{m}A_S^*A_S - \mathrm{Id}\right|\right|_{2\to 2} \leq 2(\sqrt{s/m} + \eta) + (\sqrt{s/m} + \eta)^2\right) \geq 1 - 2\exp(-m\eta^2/2).$$

Again, in view of the definition $\delta_s = \sup_{S\subset[N],\ \#S=s} ||A_S^*A_S - Id||_{2\to 2}$, we take the union bound over all $\binom{N}{s}$ subsets of $[N]$ with cardinality s:

$$\mathbb{P}\Big[\delta_2 > 2(\sqrt{s/m} + \eta) + (\sqrt{s/m} + \eta)^2\Big] \leq 2\binom{N}{s}e^{-m\eta^2/2} \leq 2\left(\frac{eN}{s}\right)^s e^{-m\eta^2/2} \leq \varepsilon,$$

by the hypothesis on the number of rows and the Stirling's formula. This same hypothesis also tells us that $\sqrt{s/m} \leq \frac{\eta}{\sqrt{2\ln(eN/s)}}$ and so

$$\mathbb{P}\left[\delta_s\left(\frac{1}{\sqrt{m}}A\right) > 2\left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)\eta + \left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)^2\eta^2\right]$$

$$\leq \mathbb{P}\Big[\delta_2 > 2(\sqrt{s/m} + \eta) + (\sqrt{s/m} + \eta)^2\Big] < \varepsilon.$$

This concludes the proof.

□

The reader should ask why we expend so much effort in the Gaussian case, proving even technical lemma such as Lemma 7.11. The reason is related to the constant C in $m \geq Cs\ln(N/s)$ and here a difference between Theorem 7.2 and Theorem 7.13 appears. In the Gaussian case, the number of measurements are much less than in the general subgaussian case and in this case we have the best known constant so far. The reason behind it is related to Theorem 6.32. The reader can consult [Vershynin '15] for more details on this comparison. Also, the techniques developed for the Gaussian case are so powerful that we can not only prove that they satisfy the RIP with small constant but we also can directly establish that they satisfy NSP, rather than relying on RIP and use it as a sufficient condition to NSP. The proof is even more laborious and appeared in the literature for the first time in the book [Rauhut & Foucart '13].

Theorem 7.14. (Theorem 9.29 of [Rauhut & Foucart '13]: *Let $A \in \mathbb{R}^{m\times N}$ be a random drawn of a Gaussian matrix. Assume that*

$$\frac{m^2}{m+1} \geq 2s\ln(eN/s)\left(1+\rho^{-1}+D(s/N)+\sqrt{\frac{\ln(\varepsilon^{-1})}{s\ln(eN/s)}}\right)^2,$$

where D is a function that satisfies $D(\alpha) \leq 0.92$ for all $\alpha \in (0,1]$ and $\lim_{\alpha\to 0} D(\alpha) = 0$. The, with probability at least $1-\varepsilon$ the matrix A satisfies the stable null space property of order s with constant ρ.

Proof. See Theorem 9.29 and Corollary 9.34 from [Rauhut & Foucart '13]. □

7.5 Johnson-Lindenstrauss Embeddings and the RIP

Assume we have a cloud of data in a high dimensional space, *i.e.*, M points $\{x_1,\ldots,x_M\} \in \mathbb{R}^N$. If $N > M$, these points lie in a subspace of dimension M. Then we can consider the projection of these points into this subspace without distorting their mutual distance, that is, $\forall x_i, x_j$ we have a projection f satisfying $||f(x_i)-f(x_j)||_2 = ||x_i - x_j||_2$. This is the same as saying that we have a natural isometry (only for the data points) into this subspace.

Usually it is computationally expensive to process data when it lies in a high dimensional space. Moreover the existing algorithms scale very poorly when increasing the dimension of the data. So this projection must be done to avoid the *curse of dimensionality*, extremely important in these "Big Data" days. At the same time, we need to project them while preserving some kind of geometric structure.

Now, consider the situation where we allow a bit of distortion when projecting this cloud of data, that is, instead of looking for isometric projections, we look for ε-isometries

$$(1-\varepsilon)||x_i-x_j||_2^2 \leq ||f(x_i)-f(x_j)||_2^2 \leq (1+\varepsilon)||x_i-x_j||_2^2. \tag{7.7}$$

When we have measurements with errors, it is reasonable to allow this kind to projection. The question is then the following: provided we have distortion, can we do beter than $N = M$?, that is, can we project on a lower dimensional space? [Johnson & Lindenstrauss '84] proved that this is possible.

Theorem 7.15. (Lemma 1 of [Johnson & Lindenstrauss '84]): *Let $x_1,\ldots,x_M \in \mathbb{R}^N$ be an arbitrary set of points and $\varepsilon > 0$. If*

$$m > K\varepsilon^{-2}\ln(M),$$

then there exists a matrix $A \in \mathbb{R}^{m\times N}$ such that

$$(1-\varepsilon)||x_i-x_j||_2^2 \leq ||A(x_i-x_j)||_2^2 \leq (1+\varepsilon)||x_i-x_j||_2^2,$$

for all $i,j \in [M]$. The constant $K > 0$ is universal.

Proof. If we look at the set of mutual distances $E = \{x_j - x_i \ : \ 1 \leq j \leq i \leq M\}$ with cardinality $\#E \leq M(M-1/2)$, we just need to show the existence of a matrix A such that

$$(1-\varepsilon)||x||_2^2 \leq ||Ax||_2^2 \leq (1+\varepsilon)||x||_2^2, \qquad \forall x \in E. \tag{7.8}$$

Take $\tilde{A} = \frac{1}{\sqrt{m}}A \in \mathbb{R}^{m\times N}$, where A is a random drawn of a subgaussian matrix. Then Lemma 7.5 implies the existence of a constant C such, for any fixed $x \in E$, we have

$$\mathbb{P}\left(\left|||\tilde{A}x||_2^2 - ||x||_2^2\right| \geq \varepsilon||x||_2^2\right) \leq 2\exp(-Cm\varepsilon^2).$$

Now, taking the union bound in such a way that Equation (7.8) holds for all $x \in E$

$$\mathbb{P}\left(\bigcup_{x\in E}\left|||\tilde{A}X||_2^2 - ||x||_2^2\right| \geq \varepsilon||x||_2^2\right) \leq \sum_{x\in E}\mathbb{P}\left(\left|||\tilde{A}X||_2^2 - ||x||_2^2\right| \geq \varepsilon||x||_2^2\right) \leq 2\frac{M(M-1)}{2}\exp(-Cm\varepsilon^2).$$

And so

$$\mathbb{P}\left((1-\varepsilon)||x||_2^2 \leq ||Ax||_2^2 \leq (1+\varepsilon)||x||_2^2 \qquad \forall x \in E\right) \geq 1 - M^2e^{-Cm\varepsilon^2}$$

As we want $M^2e^{-Cm\varepsilon^2}$ smaller than a certain η, we just need to take $m \geq C^{-1}\varepsilon^{-2}\ln(M^2/\eta)$ and then the inequality above holds with probability $1-\eta$. We conclude the existence of the Johnson-Lindenstrauss map when $\eta < 1$. Considering the limit $\eta \to 1$, this gives the claim with $K = 2C^{-1}$. □

This deep theorem tells us that if we relax the condition on the isometry and ask for a quasi-isometry, we can project our cloud of data in a much lower dimensional space, whose dimension is logarithmic in the number of vectors and can be controlled by ε, our error relative to geometric structure preservation. And it is remarkable because the target dimension is independent of the ambient dimension N. Moreover, the proof presented here is that subgaussian matrices work as this kind of projection.

The Johnson-Lindenstrauss Lemma has a lot of applications and is widely used nowadays to transform a high-dimensional problem into a low-dimensional one in such a way that the optimal solution to the problem in low dimension can be lifted to a nearly optimal solution to the high dimensional one. Besides its connection with compressive sensing, we could cite aplications in combinatorial optimization, data streams, fast low-rank approximations, nearest neighbor search, etc. For variations on the theme and applications, see [Matousek '08] and [Vempala '04].

It would be interesting, for computational purposes, to find other ensemble of matrices which can be used as projections, other than the subgaussians ones. The next Theorem, proved in [Krahmer & Ward '12] shows that given a matrix A satisfying the RIP, a randomization of the column signs of A provides a Johnson-Lindenstrauss embedding and this will allow fast computational embeddings.

Theorem 7.16. ([Krahmer & Ward '12]): *Let $E \subset \mathbb{R}^N$ be a finite point set such that $\#E = M$. For $\varepsilon, \eta \in (0,1)$, let $A \in \mathbb{R}^{m\times N}$ with RIC satisfying $\delta_{2s} \leq \varepsilon/4$ for some $s \geq 16\ln(4M/\eta)$ and let $\epsilon = (\epsilon_1, \ldots, \epsilon_N)$ be a Rademacher sequence. Also define D_ϵ as the diagonal matrix with ϵ on the diagonal. Then with probability at least $1-\eta$*

$$(1-\varepsilon)||x||_2^2 \leq ||AD_\epsilon x||_2^2 \leq (1+\varepsilon)||x||_2^2, \qquad \forall x \in E. \tag{7.9}$$

Proof. We may assume, without loss of generality, that for all $x \in E$, we have $||x||_2 = 1$. For a fixed $x \in E$, we will partition $[N]$ into blocks of size s as a nonincreasing rearrangement of x. More precisely, let us take $S_1 \subset [N]$ as the index set of the s largest absolute entries of x. Now, take $S_2 \subset [N]\backslash S_1$ as the index set of the s largest absolute entries of x in $[N]\backslash S_1$ and so on. With this construction, we can write

$$\left|\left|AD_\epsilon x\right|\right|_2^2 = \left|\left|AD_\epsilon \sum_{j\geq 1} x_{S_j}\right|\right|_2^2 = \sum_{j\geq 1}\left|\left|AD_\epsilon x_{S_j}\right|\right|_2^2 + 2\left\langle AD_\epsilon x_{S_1}, AD_\epsilon x_{\overline{S_1}}\right\rangle + \sum_{\substack{j,l\geq 2\\ j\neq l}}\left\langle AD_\epsilon x_{S_j}, AD_\epsilon x_{S_l}\right\rangle.$$

By hypothesis, A satisfies RIP with $\delta_s \leq \delta_{2s} \leq \varepsilon/4$ and since $||D_\epsilon x_{S_j}|| = ||x_{S_j}||_2$ we can estimate the first term by

$$(1-\varepsilon/4)||x||_2^2 = (1-\varepsilon/4)\sum_{j\geq 1}||x_{S_j}||_2^2 \leq \sum_{j\geq 1}\big|\big|AD_\epsilon x_{S_j}\big|\big|_2^2 \leq (1+\varepsilon/4)||x||_2^2.$$

To estimate the second term, we use the fact that $D_\epsilon x = D_x \varepsilon$. Considering the random variable

$$X = \Big\langle AD_\epsilon x_{S_1}, AD_\epsilon x_{\overline{S_1}}\Big\rangle = \Big\langle AD_{x_{S_1}}\epsilon, AD_{x_{\overline{S_1}}}\epsilon\Big\rangle = \Big\langle D_{x_{S_1}}A^*_{x_{S_1}}A_{S_1}D_{x_{S_1}}\epsilon_{S_1}, \epsilon_{\overline{S_1}}\Big\rangle = \Big\langle v, \epsilon_{\overline{S_1}}\Big\rangle = \sum_{i\neq S_1}\epsilon_i v_i,$$

where $v \in \mathbb{R}^{\overline{S_1}}$ is given by

$$v = D_{x_{S_1}}A^*_{x_{S_1}}A_{S_1}D_{x_{S_1}}\epsilon_{S_1}.$$

Now, note that v and $\epsilon_{\overline{S_1}}$ are independent because one of them just depends on S_1 and the other just on $\overline{S_1}$. Our goal is to estimate this second term by the Hoeffding's inequality for Rademacher variables, Equation (6.8), that says

$$\mathbb{P}\Big(\sum_{i=1}^M \epsilon_i v_i \geq ||v||_2 u\Big) \leq 2\exp(-u^2/2).$$

Therefore we need to estimate the 2-norm of the vector v, so

$$||v||_2 = \sup_{||z||_2\leq 1}\langle v, z\rangle = \sup_{||z||_2\leq 1}\sum_{i\geq 2}\Big\langle z_{S_i}, D_{x_{S_i}}A^*_{S_i}A_{S_1}D_{x_{S_1}}\epsilon_{S_1}\Big\rangle$$

$$= \sup_{||z||_2\leq 1}\sum_{i\geq 2}\Big\langle z_{S_i}, D_{x_{S_i}}A^*_{S_i}A_{S_1}D_{\epsilon_{S_1}}x_{S_1}\Big\rangle \leq \sup_{||z||_2\leq 1}\sum_{i\geq 2}||z_{S_i}||_2||D_{x_{S_i}}A^*_{S_i}A_{S_1}D_{\epsilon_{S_1}}||_{2\to 2}||x_{S_1}||_2$$

$$\leq \sup_{||z||_2\leq 1}\sum_{i\geq 2}||z_{S_i}||_2||D_{x_{S_i}}||_{2\to 2}||A^*_{S_i}A_{S_1}||_{2\to 2}||D_{\epsilon_{S_1}}||_{2\to 2}||x_{S_1}||_2$$

$$\leq \sup_{||z||_2\leq 1}\sum_{i\geq 2}||z_{S_i}||_2||x_{S_i}||_\infty||A^*_{S_i}A_{S_1}||_{2\to 2}||x_{S_1}||_2.$$

Here we have used that $||D_{x_{S_i}}||_{2\to 2} = ||x_{S_i}||_\infty$ and that $||D_{\epsilon_{S_1}}||_{2\to 2} = ||\epsilon_{S_1}||_\infty = 1$. Moreover, from the way we construct $S_1, S_2, \ldots$, etc, it follows from Lemma 5.18 that $||x_{S_i}||_\infty \leq s^{-1/2}||x_{S_{i-1}}||_2$. Besides that, $||x_{S_1}||_2 \leq ||x||_2 = 1$ and from Proposition 5.5 we have, for $i \geq 2$, that $||A^*_{S_i}A_{S_1}||_{2\to 2} \leq \delta_{2s}$. It then follows that

$$||v||_2 \leq \frac{\delta_{2s}}{\sqrt{s}}\sup_{||z||_2\leq 1}\sum_{i\geq 2}||z_{S_i}||_2||x_{S_{i-1}}||_2 \leq \frac{\delta_{2s}}{\sqrt{s}}\sup_{||z||_2\leq 1}\sum_{i\geq 2}\frac{1}{2}\big(||z_{S_i}||_2^2 + ||x_{S_{i-1}}||_2^2\big) \leq \frac{\delta_{2s}}{\sqrt{s}}, \tag{7.10}$$

where we have used the geometric-arithmetic mean inequality and that $\sum_{j\geq 2}||x_{S_j}||_2^2 \leq ||x||_2^2 = 1$. Now, Hoeffding's inequality, Equation (6.8), conditionally on ϵ_{S_1} yields, for $t > 0$

$$\mathbb{P}\Big(|X| \geq ||v||_2 u\Big) = \mathbb{P}\Big(\sum_{i=1}^M \epsilon_i v_i \geq ||v||_2 u\Big) \leq 2\exp(-u^2/2),$$

which, using Equation (7.10), $\delta_{2s} \leq \varepsilon/4$ and $||v||_2 u = t$, is the same as

$$\mathbb{P}\Big(|X| \geq t\Big) \leq \exp(-t^2/2||v||_2^2) \leq \exp(-t^2 s/2\delta_{2s}^2) \leq \exp(-8st^2/\varepsilon). \tag{7.11}$$

To finish the proof, we need to estimate the third term. In order to do it, let us consider the random variable Y define by

$$Y = \sum_{\substack{j,l\geq 2\\ j\neq l}} \Big\langle AD_\epsilon x_{S_j}, AD_\epsilon x_{S_l}\Big\rangle = \sum_{j,l=s+1}^{N} \epsilon_j\epsilon_l B_{j,l} = \boldsymbol{\epsilon}^*\boldsymbol{B}\boldsymbol{\epsilon},$$

where $\boldsymbol{B}\in\mathbb{R}^{N\times N}$ is a symmetric matrix with zero diagonal given by

$$B_{j,l} = \begin{cases} x_j\boldsymbol{a}_j^*\boldsymbol{a}_l x_l & \text{if } j,l\in\overline{S_1} \text{ and } j,l \text{ belong to different blocks } S_k\\ 0 & \text{otherwise}\end{cases}$$

We will bound the tail of this Rademacher chaos using Theorem 6.25. Therefore, we need now to estimate the Frobenius and the spectral norms of $\boldsymbol{B}$. As it is symmetric, its spectral norm can be estimated by

$$||\boldsymbol{B}||_{2\to 2} = \sup_{||z||_2\leq 1}\langle \boldsymbol{B}z, z\rangle = \sup_{||z||_2\leq 1}\sum_{\substack{j,l\geq 2\\ j\neq l}}\Big\langle z_{S_j}, D_{x_{S_j}}A_{S_j}^*A_{S_l}D_{x_{S_1}}z_{S_l}\Big\rangle$$

$$\leq \sup_{||z||_2\leq 1}\sum_{\substack{j,l\geq 2\\ j\neq l}} ||z_{S_j}||_2||z_{S_l}||_2||x_{S_j}||_\infty||x_{S_l}||_\infty||A_{S_j}^*A_{S_l}||_{2\to 2}$$

$$\leq \delta_{2s}\sup_{||z||_2\leq 1}\sum_{\substack{j,l\geq 2\\ j\neq l}} ||z_{S_j}||_2||z_{S_l}||_2 s^{-1/2}||x_{S_{j-1}}||_2 s^{-1/2}||x_{S_{l-1}}||_2$$

$$\leq \frac{\delta_{2s}}{4s}\sup_{||z||_2\leq 1}\sum_{\substack{j,l\geq 2\\ j\neq l}} \big(||x_{S_{j-1}}||_2^2 + ||z_{S_j}||_2^2\big)\big(||x_{S_{l-1}}||_2^2 + ||z_{S_l}||_2^2\big) \leq \frac{\delta_{2s}}{s}.$$

The Frobenius norm obeys the same bound,

$$||\boldsymbol{B}||_F = \sum_{\substack{j,l\geq 2\\ j\neq l}}\sum_{i\in S_j}\sum_{l\in S_k}(x_j\boldsymbol{a}_j^*\boldsymbol{a}_l x_l)^2 = \sum_{\substack{j,l\geq 2\\ j\neq l}}\sum_{i\in S_j} x_i^2\big|\big|D_{x_{S_k}}A_{S_k}^*\boldsymbol{a_i}\big|\big|_2^2$$

$$\leq \sum_{\substack{j,l\geq 2\\ j\neq l}}\sum_{i\in S_j} x_i^2||x_{S_k}||_\infty^2||A_{S_k}^*\boldsymbol{a_i}||_2^2 \leq \delta_{2s}^2\sum_{\substack{j,l\geq 2\\ j\neq l}}||x_{S_j}||_2^2 s^{-1}||x_{S_k}||_2^2 \leq \frac{\delta_{2s}^2}{s},$$

where we have used the fact that $||A_{S_k}^*\boldsymbol{a_i}||_2^2 = ||A_{S_k}^*||_{2\to 2}\leq \delta_{s+1}\leq\delta_{2s}$. So, by Inequality (6.14), the tail of the third term can be estimate, for any $t>0$, by

$$\mathbb{P}\Big(|Y|\geq t\Big) \leq 2\exp\left(-\min\left\{\frac{3t^2}{128||\boldsymbol{B}||_F^2}, \frac{t}{32||\boldsymbol{B}||_{2\to 2}}\right\}\right)$$

$$\leq 2\exp\left(-\min\left\{\frac{3st^2}{128\delta_{2s}}, \frac{st}{32\delta_{2s}}\right\}\right) \leq 2\exp\left(-\min\left\{\frac{3t^2}{8\varepsilon^2}, \frac{t}{8\varepsilon}\right\}\right). \tag{7.12}$$

Choosing $t=\varepsilon/6$ in (7.11) and $t=\varepsilon/2$ in (7.12) and taking into account the bounds for the three terms simultaneously, we conclude that for a fixed $x\in E$, with probability at least

$$1-2\exp(-s/8)-2\exp(-s\min\{3/32, 1/16\}) \geq 1-\exp(-s/16),$$

we have

$$(1-\varepsilon)||x||_2^2 \leq ||AD_\epsilon x||_2^2 \leq (1+\varepsilon)||x||_2^2.$$

Taking the union bound over all $x\in E$ and using the hypothesis $s\geq 16\ln(2M/\eta)$, we conclude that this holds for all $x\in E$ with probability at least

$$1 - 4M\exp(-s/16) \geq 1 - \eta.$$

This concludes the proof. □

We note that this theorem allows us to conclude the existence of Johnson-Lindenstrauss embeddings for other types of matrices. In particular, matrices that satisfy RIP provide such embeddingsm as is the case of random partial Fourier matrices. It is an open problem *to show directly that such type os matrices are JL embeddings* because all the proofs involve, in some way, subgaussian matrices. Without the randomization of columns, this Theorem is false. To see this, just take the points of E belonging to the kernel of A then the lower bound cannot hold. Thus the randomization of columns signs ensures that the probability of the intersection of E and the kernel of AD_ϵ not being empty is very small.

Additionally, the question about sharpness of the results emerges, that is, if we could improve the fact that $m = O(\varepsilon^{-2}\ln(M))$ or, in other words, if there exist some set E such that for any such map f, as defined in (7.7), we must have $m = \Omega(\varepsilon^{-2}\ln(M))$.

In their original paper, [Johnson & Lindenstrauss '84] proved the first lower bound of $m = \Omega(\ln N)$ when ε is smaller than some constant. This was improved by [Alon '03], who showed that any JL map must embed into dimension $m = \Omega(\min\{N, \varepsilon^{-2}\ln N/\ln(1/\varepsilon)\})$, where the first term in the min is achieved by the identity map. This was later improved to $m = \Omega(\varepsilon^{-2}\min\{\ln N, (\ln N/\ln(1/\varepsilon))^2\})$. Finally, [Larsen & Nelson '14] closed the gap by proving the following result.

Theorem 7.17. *There exists $C > 0$ such that for any $N > 1$ and $0 < \varepsilon < 1/2$, there is a subset $E \subset \mathbb{R}^N$ with $\#E = N + N^3$ such that any embedding linear $f : \mathbb{R}^N \to \mathbb{R}^m$ satisfying*

$$(1-\varepsilon)||x||_2^2 \leq ||f(x)||_2^2 \leq (1+\varepsilon)||x||_2^2, \qquad \forall x \in E,$$

must have $m \geq C\min\{N, \varepsilon^{-2}\ln(N)\}$. In other words, the JL lemma is optimal in the case where the projection f is linear.

Proof. Theorem 3 from [Larsen & Nelson '14]. □

It is very curious that all methods for dimensionality reduction through JL lemma are via linear maps. Moreover, the sharpness result is also for linear maps. Thus, circumventing the lower bound would require a fundamentally new approach with nonlinear projections.

To conclude this chapter, we address the subject of the time needed to generate suitable JL projections. This is known as the search for *Fast Johnson-Lindenstrauss Transform(FJLT)*, a term coined in [Ailon & Chazelle '06]. They define a probability distribution over a product of matrices through the following product

$$A_{FJLT} = PH_N D_\epsilon$$

where (remembering that we are dealing with vectors $E = \{x_1, \ldots, x_M\} \in \mathbb{R}^N$, that is, we have M vectors in a space whose dimension is N)

- D_ϵ is a $N \times N$ matrix where each D_{ii} is drawn independently from $\{-1, 1\}$ with probability $1/2$, as in Theorem 7.16. Note that D_ϵ is an isometry for the 2-norm.
- H_N is the $N \times N$ Hadamard matrix (we are assuming that N is a power of 2) with entries given by $(H_d)_{ij} = (-1)^{\langle i-1, j-1\rangle}/\sqrt{N}$ where for i, j each in $\{0, ..., d-1\}$ we write them in binary and treat them as $\log_2 d$-dimensional vectors. Note that H_N is also an isometry for the 2-norm.
- P is a $m \times t$ matrix whose elements are independently distributed as follows. With probability $1-\eta$ set P_{ij} as 0, and otherwise (with the remaining probability η) draw P_{ij} from a normal distribution of expectation 0 and variance η^{-1}. The sparsity constant q is given as

$$q = \min\left\{\Theta\left(\frac{\varepsilon^{-2}\log M}{t}\right), 1\right\}.$$

We need to choose t in order to complete the construction. It will be $t = \varepsilon^{-2}\ln(1/\eta)\ln(N/\eta)$, with η as defined above being the probability of failure in the JL lemma. The running time to apply A to some vector $x \in \mathbb{R}^N$ is

- D_ϵ takes $O(N)$ time.
- Due to the recurrence relation $H_1 = (1)$ and $H_N = \frac{1}{\sqrt{2}}\begin{pmatrix} H_{N/2} & H_{N/2} \\ H_{N/2} & -H_{N/2} \end{pmatrix}$, it takes $O(N\ln N)$.
- Applying P takes $O(mt)$ time.

So the total time to apply A_{FJLT} is $O\Big(N\ln N + \varepsilon^{-4}\ln^2(1/\eta)\ln(N/\eta)\Big)$. To complete the argument, an analysis must be done showing that this is an JL isometry. This is given by the following lemmas:

Lemma 7.18. *For any $x \in \mathbb{R}^N$ with $||x||_2 = 1$, and for any $0 < \eta < 1/2$,*

$$\mathbb{P}\left[\Big|\Big|H_N D_\epsilon x\Big|\Big|_\infty > \sqrt{\frac{\ln(N/\eta)}{N}}\right] < \eta.$$

Proof. Lemma 1 from [Ailon & Chazelle '09]. □

Lemma 7.19. *If $||x||_2 = 1$ has $||x||_\infty \leq \sqrt{\ln(N/\eta)/N}$, then*

$$\mathbb{P}\Big[1 - \varepsilon \leq ||Px||_2 \leq 1 + \varepsilon\Big] > \eta.$$

Proof. Lemma 7 from [Nelson '??][2]. □

This approach is known as *FJLT via Fast Hadamard*. For more details and a complete analysis, see [Ailon & Chazelle '06]. Other methods to obtain similar results include the use of FFT and also Toeplitz and circulant matrices. Some of them do not need to assume that matrices have dimension $N = 2^k$ for some k, see [Vybiral '11], [Hinrichs & Vybiral '11], [Ailon & Liberty '11], [Dasgupta, Kumar & Sarlós '10], and references therein. Besides, Jelani Nelson points out the following open question, the most important in the computational tractability of using JL as a tool for dimension reduction:

Open Problem: Obtain a probability distribution for the JL lemma with embedding time $O(N\ln m)$ (or even $O(N\ln N)$) and the dimension of the target space being $m = O(\varepsilon^{-2}\ln(N))$

In this chapter we proved that for matrices for which $m \geq C_\delta s\ln(N/s)$, RIP holds. It turns out that *subgaussian matrices* provide a good ensemble for the algorithms of sparse recovery to work. We also saw that, roughly speaking, this scale is optimal in order to do dimension reduction via Johnson-Lindenstrauss lemma.

Now we would like to know whether this scale is optimal and if there are some special situations where it could be improved. For example, can we design, for some specific sparse signals, smart measurement schemes such that there will be improvement in the art of measurement? In the next Chapter we provide an answer through ideas based on the work of Kolmogorov and Gelfand in approximation theory.

[2] It is not indicated in the website of Jelani Nelson, `http://people.seas.harvard.edu/~minilek/`, when these notes were written, so we have the interrogation on the reference.

Chapter 8

Optimality in the Number of Measurements

> É estranho que tu, sendo homem do mar, me digas isso, que ja não há ilhas desconhecidas. Homem da terra sou eu, e não ignoro que todas as ilhas, mesmo as conhecidas, são desconhecidas enquanto não embarcamos nelas.[1]
>
> *José Saramago in O Conto da Ilha Desconhecida*

8.1 Introduction

Along this dissertation we developed the theory of Compressive Sensing and showed that it is possible to sense signals using much fewer measurements than stated by the usual sampling paradigm. In the introduction of [Donoho '06], he asks *"why go to so much effort to acquire* ***all*** *the data when* ***most*** *of what we get will be thrown away? Can't we just* ***directly*** *measure the part that won't end up being thrown away?"*

We saw that this new kind of sensing can be done in a nonadaptive way, that is, we do not require the knowledge of the signal in advance. We just need to know that data is sparse or compressible when expressed in some basis. Several techniques were required in order to show when and how this is possible. We passed through Mathematical Optimization, Nonasymptotic Probability and sophisticated ideas from Linear Algebra and Harmonic Analysis. This shows us that often circumventing dominant ideas or breaking paradigms can be a difficult task.

After many steps, we arrived at a point where it was possible to show how to design (random) sensing matrices and perform optimal sampling with them. This was a breakthrough introduced by Donoho, Tao, Candès and Romberg. They discovered that $m \approx s \ln(N/s)$ is a rule of thumb regarding the number of measurements. In other words, the information acquired is linear in the information content (sparsity S) and logarithmic in the ambient dimension.

In Section 1.4 of his seminal work[2], Donoho says that the estimation of error measurement of compressible objects *"concerns the geometry of high-dimensional convex and nonconvex balls...to develop this geometric viewpoint further, we consider two notions of n-width"* and argues that there is an equivalence between optimal recovery of nonadaptive information and objects in approximation theory.

The purpose of this chapter is to explore this equivalence. Despite the fact that this type of observation was well known in the Approximation Theory community (see [DeVore '06] and [Micchelli & Rivlin '76]), it was a new information for people working on Signal Processing. Another fundamental paper of Compressive Sensing concerning this error reconstruction questions is [Cohen, Dahmen & DeVore '09].

In this final chapter, we will define the two notions of n-width that Donoho was referring to. In order to do this, we first need to discuss some ideas of Approximation Theory. After, we prove the Theorem

[1] A free translation of Saramago's quote is: *It is strange that you, being a man from the sea, tell me this, that there are no unknown islands. I am a man from the land, and I am not unaware that all the islands, even the known ones, are unknown until we land on them.*

[2] According to Google Scholar, it has more than 16500 citations.

of Kashin-Garnaev-Gluskin in Geometry of Banach Spaces and explain why this deep theorem is fundamental for Compressive Sensing. We derive, as its consequence, the optimal number of measurements for sparse recovery, regardless of the reconstruction scheme.

The first part of this chapter was highly inspired by [Gamkrelidze '90] and [Pinkus '85]. The technical results where extracted from [Rauhut & Foucart '13] and original papers. Also, the historical part was taken from many papers written by Vladimir Tikhomirov[3].

8.2 n-Widths in Approximation Theory

In any computational study where the available memory is finite, it is necessary to represent functions using a finite number of parameters, for example, coefficients of some truncated series expansion. Moreover, these coefficients, real numbers, must be transformed into a finite alphabet and many times this alphabet will be represented by binary digits. This is the problem of quantization. With the bits obtained we could use coding techniques to achieve a final compression.

Many ideas for this kind of representation (or approximation) came from the Soviet school [Steffens '06]. In particular, the abstract notions of metric entropy and widths are linked to the name of Kolmogorov, who introduced them, and Tikhomirov, one of his students. The latter developed many techniques related to widths and solved many problems left open by Kolmogorov and collaborators. In this context, the names of Mityagin, Vitushkin, Kashin, Gluskin, Maiorov, Solomyak should also be mentioned [Pietsch '07].

One of the first problems studied in approximation theory is the following: given a point and a subset in a metric space, find the point (or points) in the subset which best approximates the given point. As well as in the quantization problem, in any approximation problem it is necessary to characterize what *best approximation* (or quantization) means. This will measure the accuracy with which an element from the given set can be recovered. For example, given $x \in X$, where X is a Banach space, and given X_n, an n-dimensional subspace of X, we wish to find the best approximation of x in X_n. Also we would like to estimate the value of the error, i.e., the measure of the distance between x and its best approximant.

If we denote the distance by

$$d(x, X_n) = \inf\{||x - y||_X : y \in X_n\},$$

we ask if there exists a $y^* \in X_n$ for which $d(x, X_n) = ||x - y^*||$ and also ask for its characterization and uniqueness, when possible. Note that as X_n is a finite dimensional subspace of X, a best approximation always exist.

Now, instead of a point, we could ask the same for a given subset $K \subset X$, that is, how well one can approximate K by X_n, an n-dimensional subspace of X. Sometimes this is called the *deviation* of A from X_n. We also allow the possibility of varying the n-dimensional subspaces X_n. This was first done by [Kolmogorov '36], despite Uryson in 1922 and Aleksandrov in 1933 having created similar definitions for the study of geometric problems. Geometry was also the motivation behind the theory of widths developed in the Soviet Union, see Chapter 8 of [Charpentier, Lesne & Nikolski '07].

Definition 8.1. The Kolmogorov m-width of a subset K of a normed space X is defined as

$$d_m(K, X) = \inf_{X_m} \left\{ \sup_{x \in K} \inf_{y \in X_m} ||x - y||, \ X_m \text{ subspace of } X \text{ with } \dim(X_m) \leq m \right\}$$

As Tikhomirov points out in [Gamkrelidze '90], *"undoubtedly, of most importance in approximation theory is the Kolmogorov width which is the most closely connected with the basic direction of classical approximation theory".*

He also wrote, *"the determination of the entropy[4] and widths of function classes can have several goals, and these were actually all noted by Kolmogorov himself in 1956. Firstly, it can lead to invariants*

[3] The interested reader can consult the page of Tikhomirov at Math-Net: `http://www.mathnet.ru/eng/person8555`.

[4] This is a reference to entropy numbers, which are related to the concept of covering numbers defined in Section 6.6.

enabling us to distinguish function sets of different massiveness. The meaning of the most fundamental concept of "number of variables" often manifests itself in this way. Secondly, computations of widths and entropy make it possible to find new untraditional methods of approximation and to argue in favour of their expediency. Thirdly, the exact solution of the corresponding extremal problems makes it possible to enrich means of analysis by starting from the conviction that what is good does not merely turn out to be good for a single question, but usually turns out to be useful for other questions. Fourthly, and finally, this can be of interest for computational mathematics by giving guidelines for the creation of the most expedient algorithms for solving practical problems." [Tikhomirov '83].

Tikhomirov points out that in a discussion with Gel'fand, the former expressed the idea that there must be a dual to the Kolmogorov width. So this was introduced in 1965 by Tikhomirov through the following definition (it was introduced independently by S. Smolyak and I. Sharygin in the same year).

Definition 8.2. The Gel'fand m-widht of a subset K of a normed space X is defined as

$$d^m(K,X) = \inf_{X_m} \left\{ \sup_{x \in K \cap X_m} ||x||,\ X_m \text{ subspace of } X \text{ with } \operatorname{codim}(X_m) \leq m \right\}.$$

A subspace X_m of X is of codimension at most m in and only if there exists linear functionals $\lambda_1, \ldots, \lambda_m : X \to \mathbb{R}$ in the dual space X^* such that

$$X_m = \{x \in X : \lambda_i(x) = 0 \quad \text{for all } i \in [m]\} = \ker A,$$

where $A : X \to \mathbb{R}^m$, $x \mapsto [\lambda_1(x), \ldots, \lambda_m(x)]$. Using A, we have the following alternative definition

$$d^m(K,X) = \inf_{X_m} \left\{ \sup_{x \in K \cap \ker A} ||x||,\ A : X \to \mathbb{R}^m \text{ linear} \right\}.$$

This means that the width measures the extent to which one can determine $x \in X$, given the value of m functionals applied on x.

Remark 38. *In Compressive Sensing, the Gel'fand widths will be more important than the Kolmogorov ones, as we will see through this chapter.*

The following Theorem establishes the duality of these widths, as suggested by Gel'fand.

Theorem 8.3. *For $1 \leq p, q \leq \infty$, let p^*, q^* be such that $1/p^* + 1/p^* = 1$ and $1/q + 1/q^* = 1$. Then*

$$d_m(B_p^N, \ell_q^N) = d^m(B_{q^*}^N, \ell_{p^*}^N)$$

In order to prove this result, we need the definition of a dual norm and a Lemma.

Definition 8.4. Let $||.||$ be a norm on a Banach space. Its dual norm is defined by $||x||_* = \sup_{||y|| \leq 1} |\langle x, y \rangle|$.

Lemma 8.5. *Let Y be a finite-dimensional subspace of a normed space X. Given $x \in X \backslash Y$ and $y^* \in Y$, the following statements are equivalent:*

1. *y^* is a best approximation to x from Y, that is, $||y^* - x|| \leq ||y - x||,\ \forall y \in Y$.*
2. *$||x - y^*|| = \lambda(x)$ for some linear functional λ vanishing on Y and satisfying $||\lambda|| \leq 1$.*

Proof. Assume that 1) holds. Let us define the linear functional $\tilde{\lambda}$ on the space $Y \oplus \operatorname{span}(x)$ by

$$\tilde{\lambda}(y + tx) = t||x - y^*||, \qquad \forall\ y \in Y \text{ and } t \in \mathbb{R},$$

When setting $t = 0$, we see that $\tilde{\lambda}$ vanishes on Y. Furthermore, for $y \in Y$ and $t \neq 0$, we have

$$|\tilde{\lambda}(y + tx)| = |t|\ ||x - y^*|| \leq |t|\ ||x - (-y/t)|| = ||y + tx||,$$

since y^* is the best approximation to x. Dividing both sides of this inequality (which is obviously valid for $t = 0$) by $||y + tx||$, we have $||\tilde{\lambda}|| \leq 1$. Using the Hahn-Banach Extension Theorem, the functional $\lambda : X \to \mathbb{R}$ we are looking for is just the extension to X provided by this theorem.

Now assume that 2) holds. First observe that $\lambda(y) = 0$, $\forall y \in Y$ and so

$$||x - y^*|| = \lambda(x) = \lambda(x - y) \leq ||\lambda||\ ||x - y|| \leq ||x - y||, \qquad \forall\ y \in Y.$$

□

Now we are able to prove the duality relation between Gel'fand and Kolmogorov widths.

Proof. (Theorem 8.3) Recall that ℓ_q^N denotes $\mathbb{R}^N$ with the $||.||_q$ norm. Given a subspace X_m of ℓ_q^N with $\dim(X_m) \leq m$ and a vector $x \in B_p^N$, Lemma 8.5 shows that there exists a functional $\lambda : \ell_q^N \to \mathbb{R}$ with $||\lambda||_{q^*} \leq 1$ and $\lambda(X_m) = 0$ such that

$$\inf_{z \in X_m} ||x - z||_q = ||x - z^*||_q = \lambda(x) \leq ||\lambda||_{q^*} ||x||_q \leq ||x||_q = \sup_{||u||_{q^*}=1} \langle u, x \rangle = \sup_{u \in B_{q^*}^N} \langle u, x \rangle = \sup_{u \in B_{q^*}^N \cap X_m^\perp} \langle u, x \rangle,$$

where in the last equality we used the fact that the functional satisfies $\lambda(X_m) = 0$ and so $\langle u, z \rangle = 0$ $\forall\ z \in X_m$ implies $u \in X_m^\perp$. For $z \in X_m$, we have

$$\sup_{u \in B_{q^*}^N} \langle u, x \rangle = \sup_{u \in B_{q^*}^N \cap X_m^\perp} \langle u, x - z \rangle \leq \sup_{u \in B_{q^*}^N \cap X_m^\perp} ||u||_{q^*} ||x - z||_q = ||x - z||_q.$$

Therefore we deduce the equality

$$\inf_{z \in X_m} ||x - z||_q = \sup_{u \in B_{q^*}^N \cap X_m^\perp} \langle u, x \rangle.$$

It follows that

$$\sup_{x \in B_p^N} \inf_{z \in X_m} ||x - z||_q = \sup_{x \in B_p^N} \sup_{u \in B_{q^*}^N \cap X_m^\perp} \langle u, x \rangle = \sup_{u \in B_{q^*}^N \cap X_m^\perp} \sup_{x \in B_p^N} \langle u, x \rangle = \sup_{u \in B_{q^*}^N \cap X_m^\perp} ||u||_{p^*}$$

Now, note that there is a correspondence between the subspaces $X_m^\perp$ and the subspaces L_m with $\mathrm{codim}(L_m) \leq m$. Taking the infimum over all X_m with $\dim(X_m) \leq m$ yields

$$d_m(B_p^N, \ell_q^N) = d^m(B_{q^*}^N, \ell_{p^*}^N).$$

□

It is difficult to prove general results about widths. There is essentially only one general result available. It was proved in [Tikhomirov '66]. All other results are proved for very particular cases.

Theorem 8.6. (Tikhomirov '66): *For every sequence $\{\alpha_i\}_{i \in \mathbb{N}}$ satisfying $\alpha_1 > \cdots > \alpha_n$ and $\alpha_n \to 0$, there exists a Banach space X and a compact set $K \subset X$ such that*

$$d_n(K, X) = \alpha_n \qquad\qquad n = 0, 1, \ldots$$

The same holds for the Kolmogorov width $d^m(K, X)$.

It is interesting to note that the reference which transformed the theory of widths, a theory previously studied only in Soviet Union, into a well known field in western mathematics was the book [Lorentz '66]. The reason, as described in [Pietsch '07], is that *"G.G. Lorentz (1910-2006) emigrated from Leningrad to the USA. After, Lorentz wrote the book which opened up the concepts of widths and entropy to the "Free World" and has become a standard reference."*

8.3 Compressive Widths

The power of the theory of widths enters compressive sensing through the following definition, in which a way of "measuring" the worst-case reconstruction errors is given.

Definition 8.7. The (non-adaptive) compressive m-width of a subset K of a normed space X is defined as

$$E^m(K,X) = \inf\left\{ \sup_{x\in K} ||x - \Delta(Ax)||,\ A : X \to \mathbb{R}^m \text{ linear} \ \text{ and } \ \Delta : \mathbb{R}^m \to X \right\}$$

In this definition, no assumptions are made on the reconstruction map $\Delta : \mathbb{R}^m \to X$: it could be ℓ_0-minimization, a greedy algorithm or even a thresholding algorithm. It also can vary with the measurement matrix A or be fixed. Moreover, the measurement map A is nonadaptive. It is given by m fixed linear functionals $\lambda_1, \ldots, \lambda_m$, that is, the action of A is $Ax = [\lambda_1(x), \ldots, \lambda_m(x)]$. We could also consider an adaptive map where a choice of a measurement depends on the previous one through a specific rule. This adaptive map $F : X \to \mathbb{R}^m$, is represented by

$$F(x) = \left[\lambda_1(x), \lambda_{2;\lambda_1(x)}, \ldots, \lambda_{m;\lambda_1(x),\ldots,\lambda_{m-1}(x)}(x)\right], \tag{8.1}$$

that is, the linear functional $\lambda_{j;\lambda_1(x),\ldots,\lambda_{j-1}(x)}(x)$ is allowed to depend on previous $\lambda_{j-1;\lambda_1(x),\ldots,\lambda_{j-2}(x)}(x)$.

With this definition in mind, we introduce a notion of width for this kind of measurements.

Definition 8.8. The adaptive compressive m-widht of a subset K of a normed space X is defined as

$$E^m_{\text{ada}}(K,X) = \inf\left\{ \sup_{x\in K} ||x - \Delta(Ax)||,\ A : X \to \mathbb{R}^m \text{ adaptive} \ \text{ and } \ \Delta : \mathbb{R}^m \to X \right\}.$$

Remark 39. *As [Rauhut & Foucart '13] points out, intuitively, we expect that the notion of adaptivity should improve the measurement/reconstruction scheme. However this is false and is a particular case of the general philosophy from information-based complexity that "adaptivity does not help". For more information, see [Novak & Wozniakowski '08].*

In their seminal paper [Cohen, Dahmen & DeVore '09], they introduced the notion of *null space property* described in Chapter 4 and proved a result relating adaptive and non-adaptive reconstruction (see also [Donoho '06], where many similar results are proved). When considering the worst case reconstruction over K, this theorem shows that adaptive and non-adaptive compressive sensing widths are equivalent. Even more, under reasonable conditions on K, they are equivalent to the Gel'fand widths.

Theorem 8.9. *Let X be a normed space. If $K \subset X$, then*

$$E^m_{ada}(K,X) \le E^m(K,X).$$

If $-K = K$, then

$$d^m(K,X) \le E^m_{ada}(K,X).$$

Furthermore, if $K + K \subset aK$ for some positive constant a, then

$$E^m(K,X) \le a\ d^m(K,X).$$

Proof. The first inequality is obvious, as any linear map $A : X \to \mathbb{R}^m$ can be considered adaptive. For the second inequality, assume that we have an adaptive map $A_{\text{ada}} : X \to \mathbb{R}^m$ as described by (8.1) and a reconstruction map $\Delta : \mathbb{R}^m \to X$. We will consider subspaces of X generated by the kernel of a map $A : X \to \mathbb{R}^m$ defined by $A(x) = [\lambda_1(x), \lambda_{2;0}(x), \ldots, \lambda_{m;0,\ldots,0}(x)]$, so we set $X_m = \ker A$. As

$$\dim(\text{Im}(A)) + \dim(\ker(A)) = \dim X \qquad \text{and} \qquad \dim(\ker(A)) + \text{codim}(\ker(A)) = \dim X,$$

then $\operatorname{codim}(\ker(A)) = \dim(\operatorname{Im}(A)) \leq m$. So, by the definition of the Gel'fand widths, we have

$$d^m(K,X) = \inf_{X_m} \left\{ \sup_{x \in K \cap X_m} ||x||,\ X_m \text{ subspace of } X \text{ with } \operatorname{codim}(X_m) \leq m \right\} \leq \sup_{x \in K \cap \ker A} ||x||. \quad (8.2)$$

For the image of $v \in \ker A$, the image of the adaptative map A_{ada} will be zero, because

$$\lambda_1(v) = 0 \implies \lambda_{2;\lambda_1(v)}(v) = \lambda_{2;0}(v) = 0 \implies \cdots \implies \lambda_{m;\lambda_1(v),\ldots,\lambda_{m-1}(v)}(v) = \lambda_{m;0,\ldots,0}(v) = 0$$
$$\implies A_{\text{ada}}(v) = 0.$$

Thus, taking $v \in K \cap \ker A$, we have

$$||v - \Delta(0)|| = ||v - \Delta(A_{\text{ada}}(v))|| \leq \sup_{x \in K} ||x - \Delta(A_{\text{ada}}(x))||.$$

By the hypothesis, $K = -K$, which implies $-v \in K \cap \ker A$ and then

$$||-v - \Delta(0)|| = ||-v - \Delta(A_{\text{ada}}(v))|| \leq \sup_{x \in K} ||x - \Delta(A_{\text{ada}}(x))||.$$

Then, for any $x \in K \cap \ker A$,

$$||x|| = \left|\left| \frac{1}{2}(x - \Delta(0)) - \frac{1}{2}(-x - \Delta(0)) \right|\right| \leq \frac{1}{2}||x - \Delta(0)|| + \frac{1}{2}||-x - \Delta(0)|| \leq \sup_{x \in K} ||x - \Delta(A_{\text{ada}}(x))||.$$

Incorporating this in (8.2), we obtain

$$d^m(K,X) \leq \sup_{x \in K} ||x - \Delta(A_{\text{ada}}(x))||.$$

Taking the infimum over all possible A_{ada} and Δ, we conclude that $d^m(K,X) \leq E^m_{\text{ada}}(K,X)$.

Now let us use the same construction as before, namely, consider a map $A : X \to \mathbb{R}^m$ and a subspace of X given by $X_m = \ker A$. We need an appropiate choice of a reconstruction map. Define $\Delta : \mathbb{R}^m \to X$ such that

$$\Delta(y) \in K \cap A^{-1}(y) \qquad \forall\ y \in A(K).$$

This choice implies that

$$E^m(K,X) \leq \sup_{x \in K} ||x - \Delta(A(x))|| \leq \sup_{x \in K} \left[\sup_{z \in K \cap A^{-1}(Ax)} ||x - z|| \right]$$

Analyzing $x - z$, for $x \in K$ and $z \in K \cap A^{-1}(Ax)$, we see that it belongs to $K + (-K)$ and also to $\ker A = X_m$. Our hypothesis says that $K + K \subset aK$ for some positive constant a, hence we have

$$E^m(K,X) \leq \sup_{u \in aK \cap X_m} ||u|| \leq \sup_{v \in K \cap X_m} ||v||.$$

Taking the infimum, this time over X_m, we obtain $E^m(K,X) \leq a d^m(K,X)$. □

Now that we know that the comparison between widths makes sense, the question of which are the "good" sets K to consider in the context of Compressive Sensing arises. We saw in Proposition 1.5 that the unit balls B_p^N in ℓ_p^N with small p are good models for compressible vectors. Then maybe we should consider these sets and try to estimate their widths. If it is possible to estimate $d_m(B_p^N, \ell_p^N)$ for small p then, by the Theorem 8.9, we will be able to estimate $E^m(B_p^N, \ell_p^N)$. This will provide constrains to the

quantities involved in the coding/decoding process. Therefore the sharp relation between the number of measurements and the ambient space of the signals will be established.

In the quest to estimate the Gel'fand widths $d_m(B_p^N, \ell_p^N)$ for small p, some difficulties arise when $p < 1$, since in this case we do not have a norm, but an ℓ_p-quasinorm. Here we will only treat the case $p = 1$ and refer to the corresponding literature for the other cases. Nevertheless, the case $p = 1$ allows us to establish a sharp relation between the number of measurements and the dimension of the signal.

The estimates of $d_m(B_p^N, \ell_1^N)$ are the content of an important theorem, which will be proved in Section 8.5. It was proved by [Gluskin '84] and [Garnaev & Gluskin '84] following work by [Kashin '77]. In the latter, the author explored the duality relation and developed some results about Gel'fand widths in the course of determining the Kolmogorov widths of some Sobolev spaces. Before proving this very important result, let us analyze what happens if we try to use basis pursuit as a method of sparse recovery.

8.4 Optimal Number of Measurements

In this chapter we are interested in comparing m and N, this is, the relation between the amount of measured information and the dimension of the signals under the process of measurement and recover. We start this section by showing a necessary condition of combinatorial flavor for Basis Pursuit.

Theorem 8.10. *Given a matrix $A \in \mathbb{R}^{m\times N}$, if every $2s$-sparse vector $x \in \mathbb{R}^N$ is a minimizer of $||z||_1$ subject to $Az = Ax$, then*

$$m \geq C_1 s \ln\left(\frac{N}{C_2 s}\right),$$

where $C_1 = 1/\ln 9$ and $C_2 = 4$.

In order to prove this result, we need to estimate the amount of sets with a given intersection between them. This was done in the context of Compressive Sensing by [Foucart, Pajor, Rauhut & Ullrich '10] and their proof follows [Mendelson, Pajor & Rudelson '05]. However, it seems that this combinatorial lemma was already known by the community of researchers in combinatorics and complexity theory. See for example [Graham & Sloane '80] and [Noam & Avi '94].

Lemma 8.11. *Given integers $s < N$, there exist*

$$n \geq \left(\frac{N}{4s}\right)^{s/2}$$

subsets $S_1, \ldots, S_n$ of $[N]$ such that each S_j has cardinality s and

$$\#(S_i \cap S_j) < \frac{s}{2} \qquad \text{for } i \neq j.$$

Proof. Let us assume that $s \leq N/4$, because otherwise we have $1 > \frac{N}{4s}$ and it suffices to take $n = 1$ subset of $[N]$. Denote by $\mathcal{A}_s$ the family of subsets of $[N]$ having cardinality s. Let us draw a fixed element $S_1 \in \mathcal{A}_s$ and cluster all the sets $S \in \mathcal{A}_s$ such that $\#(S \cap S_1) \geq \frac{s}{2}$ into a family $\mathcal{A}_{1s}$. The cardinality of this family is estimated by

$$\#(\mathcal{A}_{1s}) = \sum_{k=\lceil s/2\rceil}^{s} \binom{s}{k}\binom{N-s}{s-k} \leq \max_{\lceil s/2\rceil \leq k \leq s} \binom{N-s}{s-k} \sum_{k=\lceil s/2\rceil}^{s} \binom{s}{k} \leq \max_{\lceil s/2\rceil \leq k \leq s} \binom{N-s}{s-k} \sum_{k=0}^{s} \binom{s}{k}$$

$$\leq 2^s \max_{\lceil s/2\rceil \leq k \leq s} \binom{N-s}{s-k} = 2^s \binom{N-s}{\lfloor s/2\rfloor},$$

where the last equality holds because it is known that a general binomial number attains its maximum when $k = N/2$ so, in our case, where $s \leq N/4$ the maximum of $s - k$ is $s/2$, which is still smaller than

$(N-s)/2$. Looking to the complement of $\mathcal{A}_{1s}$, it is clear, by definition, that any $S \in \mathcal{A}_s \backslash \mathcal{A}_{1s}$ satisfies $\#(S \cap S_1) < s/2$, provided the latter is nonempty. Then we take a new element $S_2 \in \mathcal{A}_s \backslash \mathcal{A}_{1s}$ and make a new cluster by putting together all the sets $S \in \mathcal{A}_s \backslash \mathcal{A}_{1s}$ such that $\#(S \cap S_2) \geq \frac{s}{2}$. This new family will be named $\mathcal{A}_{2s}$. By the same argument, we have

$$\#(\mathcal{A}_{2s}) \leq 2^s \binom{N-s}{\lfloor s/2 \rfloor}.$$

Next, observe that any set $S \in \mathcal{A}_s \backslash (\mathcal{A}_{1s} \cup \mathcal{A}_{2s})$ satisfies $\#(S \cap S_1) < s/2$ and $\#(S \cap S_2) < s/2$ simultaneously. Inductively we repeat this construction and selection of sets $S_1, S_2, \ldots, S_n$ until $\mathcal{A}_s \backslash (\mathcal{A}_{1s} \cup \cdots \cup \mathcal{A}_{ns})$ is empty. Hence all the constructed sets must satisfy $\#(S_i \cap S_j) < s/2$ for $i \neq j$. Moreover, this construction was done in such a way that $\cup_i \mathcal{A}_{is} = \mathcal{A}_s$ and so $\sum_{i=1}^n \#\mathcal{A}_{is} \geq \#\cup(\mathcal{A}_{is}) = \#(A_s)$. With this observation in mind, we can estimate the cardinality of $\mathcal{A}_s$ through

$$\#(\mathcal{A}_s) \leq \sum_{i=1}^n \#(\mathcal{A}_{is}) \leq n \left(\max_{1 \leq i \leq n} \#(\mathcal{A}_{is}) \right).$$

Then

$$\begin{aligned} n &\geq \frac{\#(\mathcal{A}_s)}{\max_{1 \leq i \leq n} \#(\mathcal{A}_{is})} \geq \frac{\binom{N}{s}}{2^s \binom{N-s}{\lfloor s/2 \rfloor}} \\ &= \frac{1}{2^s} \frac{N(N-1)\ldots(N-s+1)}{(N-s)(N-s-1)\ldots(N-s-\lfloor s/2 \rfloor+1)} \frac{1}{s(s-1)\ldots(\lfloor s/2 \rfloor+1)} \\ &\geq \frac{1}{2^s} \frac{N(N-1)\ldots(N-\lfloor s/2 \rfloor+1)}{(s)(s-1)\ldots(s-\lfloor s/2 \rfloor+1)} \geq \frac{1}{2^s} \left(\frac{N}{s} \right)^{\lfloor s/2 \rfloor} \geq \left(\frac{N}{4s} \right)^{s/2}. \end{aligned}$$

□

Now we are able the prove Theorem 8.10, which says that: *"if Basis Pursuit works, then the minimal number of measurements is given by $m \geq C_1 s \ln\big(N/(C_2 s)\big)$ for all s-sparse vectors"* or, in other words, what is the maximum discrepancy between the number of rows and the number of columns when we have a unique sparse solution of a linear system $Ax = b$.

Proof. (Theorem 8.10) Consider the quotient space

$$X = \ell_1^N / \ker A = \{[x] = x + \ker A, x \in \mathbb{R}^N\},$$

endowed with the norm $||[x]|| = \inf_{v \in \ker A} ||x - v||_1$. If we have a $2s$-sparse vector $x \in \mathbb{R}^N$, then for $v \in \ker A$ and $z = x - v$, we have $Az = A(x - v) = Ax$ and thus $||[x]|| = ||x||_1$. Let $S_1, \ldots S_n$ be sets as the ones introduced in the Lemma 8.11 and let us define s-sparse vectors $x^1, \ldots, x^n \in \mathbb{R}^N$ with unit ℓ_1-norms by

$$x_k^i = \begin{cases} 1/s & \text{if } k \in S_i, \\ 0 & \text{if } k \notin S_i. \end{cases}$$

As the vector $x^i - x^j$ is $2s$-sparse, we have, for $1 \leq i \neq j \leq n$, $||[x^i] - [x^j]|| = ||[x^i - x^j]|| = ||x^i - x^j||_1$ and since $|x_k^i - x_k^j| = 1/s$ if $k \in S_i \Delta S_j = (S_i \cup S_j) \backslash (S_i \cap S_j)$, vanishes otherwise and $\#(S_i \Delta S_j) > s$, we have $||x^i - x^j||_1 > 1$. This implies

$$\big|\big|[x^i] - [x^j]\big|\big| > 1 \qquad \forall\ 1 \leq i \neq j \leq n,$$

and shows that $\{[x^1], \ldots, [x^n]\}$ is a 1-separated subset of the unit sphere of X, which has dimension $r = \operatorname{rank}(A) \leq m$, since it is isomorphic to the image of A. Theorem 6.51 implies that $n \leq 3^r \leq 3^m$. By Lemma 8.11, the number of sets in this construction is at least

$$\left(\frac{N}{4s}\right)^{s/2} \leq n$$

and this implies

$$\left(\frac{N}{4s}\right)^{s/2} \leq 3^m.$$

The conclusion follows by taking the logarithm on both sides.

□

8.5 The Theorem of Kashin, Garnaev and Gluskin

An important problem posed by Kolomogorov in 1950s was to compute widths in Sobolev spaces. In the sixties, Tikhomirov and Babenko solved some of these problems using ideas of duality and Gel'fand widths. Kashin made important progress when he reduced the width calculation for function spaces to the finite-dimensional case of B_p^N inside ℓ_q^N [Tikhomirov '03]. It is important to note that [Kashin '77] was the one of the first papers in which random matrices are explicitly used to study the geometry of Banach spaces. See [Davidson & Szarek '01]. This motivated many soviet mathematicians like Garnaev, Gluskin, Temlyakov to work on this type of estimates.

For the width of the ℓ_1-ball, Kashin provided only the upper bound in Theorem 8.12 and did not achieve the best exponent in the logarithm, whereas Garnaev and Gluskin proved the lower bounds and provided the correct exponent. Kashin's construction of subspaces X_m involves $m \times N$ Bernoulli random matrices with independent entries. On the other hand, the construction of Garnaev and Gluskin for the lower bound used Gaussian matrices. More recently, [Milman & Pajor '03] realized that Bernoulli matrices could be used to yield the upper bound and they also generalized the result to an arbitrary compact and convex body $K \subset \mathbb{R}^N$.

Theorem 8.12. *For $1 < p \leq 2$ and $m < N$, there exists constants $c_1, c_2 > 0$ depending only on p such that*

$$c_1 \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p} \leq d^m(B_1^N, \ell_p^N) \leq c_2 \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p}.$$

We split the proof of this result into two parts: the upper and the lower bound, and present a modern version based on Compressive Sensing techniques. In many results in sparse recovery, we find estimates of the form $m \geq cs\ln(eN/m)$. We first need a lemma about changing the m inside the logarithm by s.

Lemma 8.13. *Let $N \geq m \geq s$ be positive integers, and c, d two positive real numbers. If $m \geq cs\ln(dN/m)$ then*

1. *$m \geq c's\ln(dN/s)$ with $c' = ec/(e+c)$,*
2. *$m \geq \tilde{c}s\ln(dN/s)$ with $\tilde{c} = c/(1+\ln(c))$ provided $c, d \geq e$.*

Proof. Rewriting the hypothesis $m \geq cs\ln(dN/m)$ we obtain:

$$m \geq cs\ln\left(\frac{dN}{s}\right) + cs\ln\left(\frac{s}{m}\right) = cs\ln\left(\frac{dN}{m}\right) + cm\frac{s}{m}\ln\left(\frac{s}{m}\right).$$

Taking $x = \frac{s}{m}$, consider $f(x) = x\ln(x)$, which is decreasing on $(0, 1/e)$ and increasing on $(1/e, +\infty)$ and has a minimum value of $-1/e$. We conclude that

$$m \geq cs\ln\left(\frac{dN}{m}\right) + cm\frac{s}{m}\ln\left(\frac{s}{m}\right) \geq cs\ln\left(\frac{dN}{m}\right) - \frac{cm}{e},$$

and this implies the first inequality

$$m \geq \left(1 + \frac{c}{e}\right)^{-1} cs \ln\left(\frac{dN}{s}\right).$$

From $m \geq cs\ln(dN/s)$, we have

$$\frac{s}{m} \leq \frac{1}{c\ln(dN/m)} \leq \frac{1}{c}.$$

As $f(x) = x\ln x$ is a decreasing function on $(0, 1/e)$, this leads to $f\left(\frac{s}{m}\right) \geq f\left(\frac{1}{c}\right) = -\frac{\ln c}{c}$ and then

$$m \geq cs\ln\left(\frac{dN}{m}\right) + cm\frac{s}{m}\ln\left(\frac{s}{m}\right) = cs\ln\left(\frac{dN}{m}\right) + cmf\left(\frac{s}{m}\right) \geq cs\ln\left(\frac{dN}{m}\right) - cm\frac{\ln c}{c}.$$

The result of the second part follows. □

Proof. (Upper Bound of Theorem 8.12): For $x \in \mathbb{R}^N$, when using the inequality $||x||_p \leq ||x||_1$ in the definition of Gel'fand width, it yields

$$d^m(B_1^N, \ell_p^N) = \inf_{X_m}\left\{\sup_{x\in B_1^N \cap X_m} ||x||_p,\ X_m \text{ subspace of } \ell_p^N \text{ with } \operatorname{codim}(X_m) \leq m\right\} \leq d^0(B_1, \ell_1^N) = 1.$$

So if $m \leq c\ln(eN/m)$ then

$$d^m(B_1^N, \ell_p^N) \leq \min\left\{1, \frac{c\ln(eN/m)}{m}\right\}^{1-1/p}. \tag{8.3}$$

On the other hand, if $m > c\ln(eN/m)$, let us define $s \geq 1$ to be the largest integer smaller than $m/(c\ln(eN/m))$, so that

$$\frac{1}{2}\frac{m}{c\ln(eN/m)} \leq s < \frac{m}{c\ln(eN/m)}.$$

Recall Lemma 8.13 we have that for $c > e$, $m > cs\ln(eN/m)$ implies $m > \tilde{c}s\ln(eN/s)$, with $\tilde{c} = c/(1+\ln(c))$. Choosing $c = 1160$,

$$m > \frac{1160}{1+\ln(1160)} s\ln(eN/s) > 144s\ln(eN/s).$$

Using Theorem 7.13, which says that if $m \geq 2\eta^{-2}\big(s\ln(eN/s) + \ln(2\varepsilon^{-1})\big)$, then we have

$$\mathbb{P}\left[\delta_s\left(\frac{1}{\sqrt{m}}A\right) \leq 2\left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)\eta + \left(1 + \frac{1}{\sqrt{2\ln(eN/s)}}\right)^2\eta^2\right] \geq 1 - \varepsilon,$$

for Gaussian random matrices. So we choose $\eta = 1/6$ and $\varepsilon = 2\exp(-m/144)$ and our hypothesis ("If $m \geq cs\ln(dN/m)$ then") turns into

$$m \geq 2\cdot 36\big(s\ln(eN/s) + \ln(\exp(-m/144)^{-1})\big) = 72\left(s\ln(eN/s) + \frac{m}{144}\right).$$

This implies $m \geq 144s\ln(eN/s)$, by Lemma 8.13. Using the estimate $1 + 1/\sqrt{2\ln(eN/s)} \leq 2$ in Theorem 7.13, we can guarantee the existence of a measurement matrix[5] $B \in \mathbb{R}^{m\times N}$ with restricted isometry constant given by

[5] In fact, what we guarantee is the existence of a ℓ_2-normalized matrix with $\delta_s(A/\sqrt{m}) \leq \delta$. As only the existence is important, we just take $B = A/\sqrt{m}$.

$$\delta_s(B) \le \delta = 4\eta + 4\eta^2 = \frac{4}{6} + \frac{4}{36} = \frac{7}{9}.$$

Now, we split the index set $[N]$ as a disjoint union $S_0 \cup S_1 \cup S_2 \cup \dots$ of index sets of size s in such a way that $|x_i| \ge |x_j|$ whenever $i \in S_{k-1}, j \in S_k$ and $k \ge 1$. By Lemma 5.18 we have $||x_{S_k}||_2 \le ||x_{S_{k-1}}||_1/\sqrt{s}$ for all $k \ge 1$. Hence, for $x \in X_m = \ker A$, we have

$$\begin{aligned}
||x||_p &\le \sum_{k\ge 0} ||x_{S_k}||_p \ \le \sum_{k\ge 0} s^{1/p-1/2}||x_{S_k}||_2 \le \sum_{k\ge 0} \frac{s^{1/p-1/2}}{\sqrt{1-\delta}}||A(x_{S_k})||_2 \\
&= \frac{s^{1/p-1/2}}{\sqrt{1-\delta}}\left[\left|\left|A\Big(-\sum_{k\ge 1} x_{S_k}\Big)\right|\right|_2 + \sum_{k\ge 1}||A(x_{S_k})||_2\right] \le \frac{s^{1/p-1/2}}{\sqrt{1-\delta}}\left[2\sum_{k\ge 1}||A(x_{S_k})||_2\right] \\
&\le 2\sqrt{\frac{1+\delta}{1-\delta}}s^{1/p-1/2}\sum_{k\ge 1}||x_{S_k}||_2 \le 2\sqrt{\frac{1+\delta}{1-\delta}}s^{1/p-1/2}\sum_{k\ge 1}||x_{S_{k-1}}||_1/\sqrt{s} \\
&= 2\sqrt{\frac{1+\delta}{1-\delta}}\frac{1}{s^{1-1/p}}\sum_{k\ge 1}||x_{S_{k-1}}||_1 \le 2\sqrt{\frac{1+\delta}{1-\delta}}\left(\frac{2c\ln(eN/m)}{m}\right)^{1-1/p}||x||_1,
\end{aligned}$$

where in the second inequality we used that $1 < p \le 2$. Choosing $\delta = 7/9$ and using $2^{1-1/p} \le 2$, it follows that, for all $x \in B_1^N \cap X^m$,

$$||x||_p \le 8\sqrt{2}\left(\frac{c\ln(eN/m)}{m}\right)^{1-1/p}.$$

This shows that, if $m > c\ln(eN/m)$, then

$$d^m(B_1^N, \ell_p^N) \le 8\sqrt{2}\min\left\{1, \frac{c\ln(eN/m)}{m}\right\}^{1-1/p}. \tag{8.4}$$

From 8.3 and 8.4, we conclude that

$$d^m(B_1^N, \ell_p^N) \le C\min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p},$$

with $C = 1160 \cdot 8\sqrt{2} = 9280\sqrt{2}$. □

We now establish the lower bound for the Gel'fand widths of ℓ_1-balls in ℓ_p^N for $1 < p \le \infty$.

Proof. (Lower Bound of Theorem 8.12): Let $c' = 2/(1 + 4\ln 9)$. First, we need to understand why it is sufficient to show that

$$d^m(B_1^N, \ell_p^N) \ge \frac{1}{2^{2-1/p}}\min\left\{1, \frac{c'\ln(eN/m)}{m}\right\}^{1-1/p}.$$

Letting $K = 1/2^{2-1/p}$ and $\varphi := \frac{\ln(eN/m)}{m}$, we need to see why

$$d^m(B_1^N, \ell_p^N) \ge K\min\{1, c'\varphi\}^{1-1/p},$$

implies , that there is a positive x such that

$$d^m(B_1^N, \ell_p^N) \ge c\min\{1, \varphi\}^{1-1/p}.$$

This is just the following sequence of logical implications:

$$d^m(B_1^N, \ell_p^N) \geq K \min\{1, c'\varphi\}^{1-1/p}$$

implies

$$d^m(B_1^N, \ell_p^N) \geq K \quad \text{or} \quad d^m(B_1^N, \ell_p^N) \geq K(c'\varphi)^{1-1/p}$$

that implies

$$d^m(B_1^N, \ell_p^N) \geq \min\{K, K(c')^{1-1/p}\} \quad \text{or} \quad d^m(B_1^N, \ell_p^N) \geq \min\{K, K(c')^{1-1/p}\}\varphi^{1-1/p}$$

which, in turn, implies

$$d^m(B_1^N, \ell_p^N) \geq c \min\{1, \varphi\}^{1-1/p} \quad \text{with} \quad c = \min\{K, K(c')^{1-1/p}\}.$$

Define $\mu = \min\left\{1, \frac{c' \ln(eN/m)}{m}\right\}$. Now, assume by contradiction that $d^m(B_1^N, \ell_p^N) \leq \mu^{1-1/p}/2^{2-1/p}$. Then there exists a subspace X_m of $\mathbb{R}^N$ with $\operatorname{codim}(X_m) \leq m$ such that, for all $v \in X_m \backslash \{0\}$,

$$||v||_p < \frac{\mu^{1-1/p}}{2^{2-1/p}} ||v||_1.$$

Again, as in the proof of Theorem 8.9, let us consider a matrix $A \in \mathbb{R}^{m \times N}$ such that $\ker A = X_m$. Let $s = \lfloor 1/\mu \rfloor \geq 1$, so that $1/(2\mu) < s \leq 1/\mu$. So, for all $v \in \ker A \backslash \{0\}$,

$$||v||_p < \frac{1}{2} \left(\frac{1}{2s}\right)^{1-1/p} ||v||_1.$$

By the Hölder inequality, we have $||v||_1 \leq N^{1-1/p} ||v||_p$ and then $1 < (N/2s)^{1-1/p}$. Hence, we can deduce that $2s < N$. Then, again by the Hölder inequality, for $S \subset [N]$ with $\#S \leq 2s$ and for $v \in \ker A \backslash \{0\}$, we have

$$||v_S||_1 \leq (2s)^{1-1/p} ||v_S||_p \leq (2s)^{1-1/p} ||v||_p < \frac{1}{2} ||v||_1.$$

The conclusion is that A has the *null space property* of order $2s$. Due to Theorem 3.3, every $2s$-sparse vector $x \in \mathbb{R}^N$ is uniquely recovered from $y = Ax$ via ℓ_1-minimization. By Theorem 8.10, the following inequality holds:

$$m \geq c_1 s \ln\left(\frac{N}{c_2 s}\right), \qquad \text{with } c_1 = \frac{1}{\ln 9} \text{ and } c_2 = 4.$$

Also we have that $m \geq 2(2s) = 4s = c_2 s$ as a consequence of Theorem 1.11. Thus

$$m \geq c_1 s \ln\left(\frac{N}{c_2 s}\right) \geq c_1 s \ln\left(\frac{N}{m}\right) = c_1 s \ln\left(\frac{eN}{m}\right) - c_1 s > \frac{c_1}{2\mu} \ln\left(\frac{eN}{m}\right) - \frac{c_1}{4} m.$$

Rearranging leads to

$$m > \frac{2c_1}{4 + c_1} \frac{\ln(eN/m)}{\min\{1, c' \ln(eN/m)/m\}} \geq \frac{2c_1}{4 + c_1} \frac{\ln(eN/m)}{c' \ln(eN/m)/m} = \frac{2(\ln 9)^{-1}}{4 + (\ln 9)^{-1}} \frac{\ln(eN/m)}{c' \ln(eN/m)/m} = m.$$

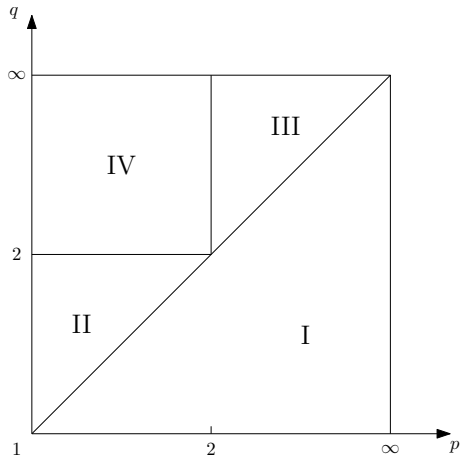
This is the contradiction we are looking for. So we have the result about the lower estimate of Gel'fand widths. □

This theorem was extended in [Donoho '06] to Gel'fand widths of ℓ_p-balls with $p < 1$. Unfortunately, his proof of the lower bound contains a gap. In the same paper, he proved the upper bound with $\log(N)$ instead of $\log(N/m)$. Then, [Vybiral '08] provided the correct upper bound when $p \leq 1$. After this, [Foucart, Pajor, Rauhut & Ullrich '10] used methods based on compressive sensing techniques (RIP, for example), to establish the lower bound for the case $p \leq 1$. This was shown in Theorem 8.12, where we exhibited their proof for the case $p = 1$. The final result is the following.

Theorem 8.14. ([Foucart, Pajor, Rauhut & Ullrich '10]): *For $0 < p \leq 1$ and $p < q \leq 2$ and $m < N$, there exists constants $c_{p,q}, C_{p,q} > 0$ depending only p and q such that*

$$c_{p,q} \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1/p-1/q} \leq d^m(B_p^N, \ell_q^N) \leq C_{p,q} \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1/p-1/q}.$$

Remark 40. *For $1 < q < p \leq \infty$, the study of widths is typically divided into four regions. We summarize here the results known for Gel'fand widths. See Chapter 14 in [Lorentz, von Golitschek & Makovoz '96] for the proof of these results and more results concerning widths in general.*



I: $1 \leq q \leq p \leq \infty$
II: $1 \leq p \leq q \leq 2$
III: $2 \leq p \leq q \leq \infty$
IV: $1 \leq p \leq 2 \leq q \leq \infty$

Region II: If $1 < p < q \leq 2$

$$d^m(B_p^N, \ell_q^N) \asymp \min\left\{1, \frac{N^{1-1/p}}{m^{1/2}}\right\}.$$

Region III: If $2 \leq p < q \leq \infty$

$$d^m(B_p^N, \ell_q^N) \asymp \max\left\{\frac{1}{N^{1/p-1/q}}, \left(1 - \frac{m}{N}\right)^{\frac{1/p-1/q}{1-2/q}}\right\}.$$

Region IV: If $1 < p \leq 2 < q \leq \infty$

$$d^m(B_p^N, \ell_q^N) \asymp \min\left\{\frac{1}{N^{1/p-1/q}}, \left(1 - \frac{m}{N}\right)^{1/2} \min\left(1, \frac{N^{1-1/p}}{m^{1/2}}\right)\right\}.$$

8.6 Connections Between Widths

With these estimates in hands, we can relate Gel'fand widths with compressive widths.

Corollary 8.15. *For $1 < p \leq 2$ and $m < N$, the adaptive and nonadaptive compressive widths satisfy*

$$E_{ada}^m(B_1^N, \ell_p^N) \asymp E^m(B_1^N, \ell_p^N) \asymp \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p}.$$

Proof. Since $-B_1^N = B_1^N$ and $B_1^N + B_1^N \subset 2B_1^N$, Theorem 8.9 implies

$$d^m(B_1^N, \ell_p^N) \leq E_{\text{ada}}^m(B_1^N, \ell_p^N) \leq E^m(B_1^N, \ell_p^N) \leq 2d^m(B_1^N, \ell_p^N).$$

By the theorem of Kashin, Gluskin and Garnaev, the results follows.

□

Proposition 8.16. *Let $1 < p \leq 2$. Suppose that there exists a matrix $A \in \mathbb{R}^{m\times N}$ and a map $\Delta : \mathbb{R}^m \to \mathbb{R}^N$ such that, $\forall x \in \mathbb{R}^N$,*

$$||x - \Delta(Ax)||_p \leq \frac{C}{s^{1-1/p}}\sigma_s(x)_1. \tag{8.5}$$

Then, for some constants $c_1, c_2 > 0$ depending only on C, we have

$$m \geq c_1 s \ln\left(\frac{eN}{s}\right),$$

provided that $s > c_2$. The same statement holds for an adaptive map $F : \mathbb{R}^N \to \mathbb{R}^m$ in place of a linear map A.

Proof. Again, it suffices to prove the proposition for an adaptive map $F : \mathbb{R}^N \to \mathbb{R}^m$, since every linear map could be considered adaptive. Hypothesis (8.5) implies

$$E^m_{\text{ada}}(B_1^N, \ell_p^N) \leq \frac{C}{s^{1-1/p}} \sup_{x\in B_1^N} \sigma_s(x)_1 \leq \frac{C}{s^{1-1/p}}.$$

By the Corollary 8.15, there exists a constant $c > 0$ such that

$$c \min\left\{1, \frac{\ln(eN/m)}{m}\right\}^{1-1/p} \leq E^m_{\text{ada}}(B_1^N, \ell_p^N) \leq \frac{C}{s^{1-1/p}}$$

Hence, there is some constant $\tilde{c} > 0$ such that

$$\tilde{c} \min\left\{1, \frac{\ln(eN/m)}{m}\right\} \leq \frac{1}{s}.$$

We conclude that either $s \leq 1/\tilde{c}$ or $m \geq \tilde{c}s\ln(eN/m)$. Setting $c_2 = 1/\tilde{c} < s$, we must have the second case(we will deal with the case $s < c_2$ in Section 8.7 because it is more delicate). To finish the proof, we need to prove that $m \geq \tilde{c}s\ln(eN/m)$ implies $m \geq c_1 s\ln(eN/s)$ with $c_1 = \tilde{c}e/(\tilde{c}+e)$. But this is the content of Lemma 8.13.

□

We will remove the restrictions $s > c_2$ and $p > 1$ in Section 8.7. Now, accepting that this theorem is true for all values of s and p, we can state the result on the minimal number of measurements for the best known values which ensures that RIP is satisfied.

Corollary 8.17. *If the $2s$-th restricted isometry constant of $A \in \mathbb{R}^{m\times N}$ satisfies $\delta_{2s} < 4/\sqrt{41} \approx 0.6246$, then necessarily*

$$m \geq cs\ln\left(\frac{eN}{s}\right)$$

for some constant $c > 0$ depending only on δ_{2s}.

Proof. If $\delta_{2s} < 0.6246$ and if Δ is *basis pursuit*, i.e., the ℓ_1-minimization reconstruction map, we proved in Theorem 5.19 that

$$||x - \Delta(Ax)||_2 \leq \frac{C}{s^{1/2}}\sigma_s(x)_1$$

holds for some constant C depending only on δ_{2s}. By Theorem 8.16, the conclusion follows. □

8.7 Instance Optimality

Suppose that we have a measurement-reconstruction scheme chosen for s-sparse recovery. We will compare the reconstruction error in ℓ_p with the best s-term approximation error in ℓ_p. In order to do this, we define the concept of ℓ_p-instance optimality for a pair of measurement matrix and reconstruction map.

Definition 8.18. Given $p \geq 1$, a pair of a measurement matrix $A \in \mathbb{C}^{m \times N}$ and a reconstruction map $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ is *ℓ_p-instance optimal* of order s with constant $C > 0$ if

$$||x - \Delta(Ax)||_p \leq C\sigma_s(x)_p \qquad \text{for all } x \in \mathbb{C}^N.$$

Some examples of ℓ_1-instance optimal pairs were shown in Chapter 6, i.e., a matrix A with restricted isometry constant δ_{2s}, δ_{6s} or δ_{26s} less than a certain value and a reconstruction map Δ corresponding to *Basis Pursuit*, *IHT* or *OMP* respectively.

We can generalize this notion of instance optimality for two different norms, one for the reconstruction error and another one for the best s-term approximation. This has already appeared in the analysis of the algorithms in Chapter 6.

Definition 8.19. Given $q \geq p \geq 1$, a pair of a measurement matrix $A \in \mathbb{C}^{m \times N}$ and a reconstruction map $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ is *mixed (ℓ_q, ℓ_p)-instance optimal of order s* with constant $C > 0$ if

$$||x - \Delta(Ax)||_q \leq \frac{C}{s^{1/p-1/q}}\sigma_s(x)_p \qquad \text{for all } x \in \mathbb{C}^N.$$

Remark 41. *Why does the term $s^{1/p-1/q}$ appears in the definition of the mixed instance optimality for two different norms? Clearly the error of reconstruction $||x - \Delta(Ax)||_q$ should be comparable to the error of best approximation $\sigma_s(x)_q$ for vectors $x \in B_r^N$ with $r < 1$. Regarding Proposition 1.5 as a "change of variables", we have*

$$\sup_{x \in B_r^N} \sigma_s(x)_q \asymp \frac{1}{s^{1/r-1/q}} \asymp \frac{1}{s^{1/p-1/q}} \sup_{x \in B_r^N} \sigma_s(x)_p.$$

This justifies the comparison between $||x - \Delta(Ax)||_q$ and $\sigma_s(x)_p / s^{1/p-1/q}$.

The next theorem shows that vectors belonging to the kernel of a measurement matrix are controlled by the best approximation error if and only if the measurement matrix is part of a mixed instance optimal scheme.

Theorem 8.20. ([Cohen, Dahmen & DeVore '09]): *Let $q \geq p \geq 1$ and a measurement matrix $A \in \mathbb{C}^{m \times N}$. If there exists a reconstruction map Δ making the pair (A, Δ) a mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant $C > 0$, then*

$$||v||_q \leq \frac{C}{s^{1/p-1/q}}\sigma_{2s}(v)_p \qquad \text{for all } v \in \ker A \tag{8.6}$$

Conversely, if Equation (8.6) holds, then there exists a reconstruction map Δ which makes the pair (A, Δ) a mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant $2C$.

Proof. Let us start by assuming that (A, Δ) is a mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant C. For $v \in \ker A$, let S be an index set of s largest absolute entries of v. From the instance optimality, we have that $||v_S - \Delta(Av_S)||_p \leq C\sigma_s(v_S)_1 = 0$ implies $v_S - \Delta(A(v_S)) = 0$, which is the same as $-v_S = \Delta(A(-v_S))$. Besides, we have that $Av = 0$ leads to $A(v_S + v_{\overline{S}}) = 0$, hence $A(-v_S) = A(v_{\overline{S}})$. We then deduce that $-v_S = \Delta(A(v_{\overline{S}}))$. With this, Equation (8.6) follows from

$$||v||_q = ||v_{\overline{S}} + v_S||_q = ||v_{\overline{S}} - \Delta(A(v_{\overline{S}}))||_q \leq \frac{C}{s^{1/p-1/q}}\sigma_s(v_{\overline{S}})_p = \frac{C}{s^{1/p-1/q}}\sigma_{2s}(v_S)_p.$$

Now, suppose that (8.6) holds for some measurement matrix A. The reconstruction map we are looking for is given by

$$\Delta(y) = \operatorname{argmin}\{\sigma_s(z)_p \text{ subject to } Az = y\}.$$

For any $x \in \mathbb{C}^N$, applying (8.6) to $v = x - \Delta(Ax) \in \ker A$ and using the triangular inequality $\sigma_{2s}(u+v)_p \leq \sigma_s(u)_p + \sigma_s(v)_p$, yields

$$||x - \Delta(Ax)||_q \leq \frac{C}{s^{1/p-1/q}} \sigma_{2s}(x - \Delta(Ax))_p \leq \frac{C}{s^{1/p-1/q}} \big(\sigma_s(x)_p + \sigma_{2s}(\Delta(Ax))_p\big) \leq \frac{2C}{s^{1/p-1/q}} \sigma_s(x)_p.$$

This proves that (A, Δ) is a mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant $2C$. □

The next theorem asserts that ℓ_2-instance optimality is not a good concept to work with. More specifically, if we expect this property to hold, then the number m of measurements is comparable to the dimension N of the signal. This holds even if we ask for instance optimality of order $s = 1$.

Theorem 8.21. ([Cohen, Dahmen & DeVore '09]): *If a pair of measurement matrix $A \in \mathbb{C}^{m \times N}$ and reconstruction map $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ is ℓ_2-instance optimal of order $s \geq 1$ with constant C, then*

$$m \geq cN,$$

for some constante c depending only on C.

Proof. From Theorem 8.20 we have that the measurement matrix A in the instance optimal pair satisfies

$$||v||_2 \leq C\sigma_s(v)_2 \qquad \forall\ v \in \ker A.$$

Taking $s = 1$ and using the triangular inequality we obtain

$$||v||_2^2 \leq C^2(||v||_2^2 - |v_j|^2), \qquad \forall\ v \in \ker A \text{ and } j \in [N].$$

If $\{e_1, \ldots, e_N\}$ represents the canonical basis of $\mathbb{C}^N$, the last inequality can be rewritten as

$$|\langle v, e_j \rangle| \leq \sqrt{(C^2 - 1)/C^2}||v||_2 \qquad \forall\ v \in \ker A \text{ and } j \in [N].$$

So, denoting the orthogonal projection onto $\ker A$ by P, we have

$$N - m \leq \dim(\ker A) = \operatorname{tr}(P) = \sum_{i=1}^{N} \langle Pe_i, e_i \rangle \leq \sum_{i=1}^{N} \sqrt{(C^2 - 1)/C^2}||Pe_i||_2 \leq \Big(\sqrt{(C^2 - 1)/C^2}\Big)N.$$

The conclusion follows by taking $c = 1 - \sqrt{(C^2 - 1)/C^2}$. □

Theorem 8.21 shows that it is better to abandon ideas related to ordinary least squares and ℓ_2-minimization in some situations involving compression, and instead look for some convex relaxation like Basis Pursuit. The next theorem confirms this expectation when looking at ℓ_1-instance optimality. Moreover, we will remove the restrictions imposed in Proposition 8.16, i.e. $q > 1$ and s larger than some constant.

Theorem 8.22. ([Foucart, Pajor, Rauhut & Ullrich '10]): *If a pair of measurement matrix $A \in \mathbb{C}^{m \times N}$ and reconstruction map $\Delta : \mathbb{C}^m \to \mathbb{C}^N$ is ℓ_1-instance optilmal of order $s \geq 1$ with constant C, then*

$$m \geq cs \ln(eN/s),$$

for some constant c depending only on C.

Proof. We know from Theorem 8.11 that there exists $n \geq (N/4s)^{s/2}$ index sets $S_1, \dots, S_n$ os size s satisfying $\#(S_i \cap S_j) < s/2$ for all $1 \leq i \neq j \leq n$. Let us consider the same construction of vectors $x^1, \dots, x^n$ from Theorem 8.10, i.e.

$$x_k^i = \begin{cases} 1/s & \text{if } k \in S_i, \\ 0 & \text{if } k \notin S_i. \end{cases}$$

First, notice that $||x^i||_1 = 1$ and $||x^i - x^j||_1 > 1$ for all $1 \leq i \neq j \leq n$. Next, we will prove that if $\rho = 1/(2(C+1))$, then $\{A(x^i + \rho B_1^N)\}_{i=1}^N$ is a disjoint collection of subsets of $A(\mathbb{C}^N)$ which has dimension $d \leq m$. Suppose, by contradiction, that we have two elements from this collection which are equal, that is, there exist indices $i \neq j$ and vectors $z, \tilde{z} \in \rho B_1^N$ such that $A(x^i + z) = A(x^j + \tilde{z})$. Then

$$\begin{aligned} ||x^i - x^j||_1 &= \Big|\Big|\big(x^i + z - \Delta(A(x^i + z))\big) - \big(x^j + \tilde{z} - \Delta(A(x^j + \tilde{z}))\big) - z + \tilde{z}\Big|\Big|_1 \\ &\leq \Big|\Big|x^i + z - \Delta(A(x^i + z))\Big|\Big|_1 + \Big|\Big|x^j + \tilde{z} - \Delta(A(x^j + \tilde{z}))\Big|\Big|_1 + ||z||_1 + ||\tilde{z}||_1 \\ &\leq C\sigma_s(x^i + z)_1 + C\sigma_s(x^j + \tilde{z})_1 + ||z||_1 + ||\tilde{z}||_1 \leq C||z||_1 + C||\tilde{z}||_1 + ||z||_1 + ||\tilde{z}||_1 \leq 2(C+1)\rho = 1. \end{aligned}$$

Now, using the fact that the collection $\{A(x^i + \rho B_1^N)\}_{i=1}^N$ is contained in $(1+\rho)A(B_1^N)$, we deduce

$$\sum_{i \in [N]} \operatorname{vol}(A(x^i + \rho B_1^N)) \leq \operatorname{vol}\Big((1+\rho)A(B_1^N)\Big).$$

We are working on the d-dimensional complex case, which is equivalent to working in the $2d$-dimensional real case. Thus, using the homogeneity and the translation invariance of the volume, we have

$$n\rho^{2d} \operatorname{vol}(A(B_1^N)) \leq (1+\rho)^{2d} \operatorname{vol}(A(B_1^N)).$$

This leads to

$$\left(\frac{N}{4s}\right)^{s/2} \leq n \leq \left(1 + \frac{1}{\rho}\right)^{2s} = (2C+3)^{2d} \leq (2C+3)^{2m}.$$

Taking logarithms, we obtain

$$\frac{m}{s} \geq \frac{\ln(N/4s)}{2\ln(2C+3)}.$$

Besides, we know by hypothesis that the pair (A, Δ) is ℓ_1-instance optimal of order s and so it ensures exact recovery of s-sparse vectors. Based on it, we have $m \geq 2s$. Combining both inequalities leads to

$$\left(4\ln(2C+3) + 2\right)\frac{m}{s} \geq \ln(N/4s) + 4 = \ln(N/4s) + \ln(e^4) = \ln(e^4 N/4s) \geq \ln(eN/s).$$

Defining $c = 1/(4(\ln(2C+3)+2))$, we have the result. □

Now we can prove that mixed (ℓ_q, ℓ_p)-instance optimality is preserved when we decrease q. We will use this to prove that the same number of measurements is imposed when we have (ℓ_q, ℓ_1)-instance optimality for $q > 1$. Equivalently, no matter in which norm we bound the error of our reconstruction, we can expect the same number of measurements in all error norms.

Lemma 8.23. *Given $q \geq q' \geq p \geq 1$, if a pair (A, Δ) is mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant C, then there is a reconstruction map Δ' making the pair (A, Δ') mixed $(\ell_{q'}, \ell_p)$-instance optimal of order s with constant C' depending only on C.*

Proof. By hypothesis, the pair (A, Δ) is a mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant C. Then, for a vector $v \in \ker A$, Theorem 8.20 yields

$$||v||_q \leq \frac{C}{s^{1/p-1/q}} \sigma_{2s}(v)_p.$$

If S denotes an index set of the $3s$ largest entries of v in modulus, we have

$$\begin{aligned} ||v_s||_{q'} \leq (3s)^{1/q'-1/q} ||v_S||_q \leq (3s)^{1/q'-1/q} ||v||_q \leq (3s)^{1/q'-1/q} \frac{C}{s^{1/p-1/q}} \sigma_{2s}(v)_p \\ = \frac{3^{1/q'-1/q} C}{s^{1/p-1/q'}} \sigma_{2s}(v)_p \leq \frac{3C}{s^{1/p-1/q'}} \sigma_{2s}(v)_p. \end{aligned}$$

Also, from Proposition 1.5, we have

$$||v_{\overline{S}}||_{q'} \leq \frac{1}{s^{1/p-1/q'}} \sigma_{2s}(v)_p.$$

Thus,

$$||v||_{q'} \leq ||v_S||_{q'} + ||v_{\overline{S}}||_{q'} \leq \frac{3C+1}{s^{1/p-1/q'}} \sigma_{2s}(v)_p = \frac{\tilde{C}}{s^{1/p-1/q'}} \sigma_{2s}(v)_p.$$

Therefore, by the converse part of Theorem 8.20, the result holds with $\tilde{C} = 2(3C+1)$. □

With the same techniques, we can prove an analogous result for p instead of q, that is, mixed (ℓ_q, ℓ_p)-instance optimality is also preserved when we decrease p instead of q. This is the content of the following Lemma.

Lemma 8.24. *For $q \geq p \geq p' \geq 1$, if a pair (A, Δ) is mixed (ℓ_q, ℓ_p)-instance optimal of order s with constant C, then it is also mixed $(\ell_q, \ell_{p'})$-instance optimal of order $\lceil s/2 \rceil$ with constant C' depending only on C.*

Corollary 8.25. *Given $q > 1$, if a pair of measurement matrix and reconstruction map is mixed (ℓ_q, ℓ_1)-instance optimal of order s with constant C, then*

$$m \geq cs \ln(eN/s)$$

for some constant c depending only on C.

Proof. This is a simple consequence of Theorem 8.22 and Lemma 8.23. □

Remark 42. *What happens in the case of a mixed (ℓ_1, ℓ_p)-instance optimal of order s with constant C for $p > 1$? Without loss of generality, suppose that we are on the region II described at the end of Section 8.5. So we have*

$$\min\left\{1, \frac{N^{1-1/p}}{m^{1/2}}\right\}^{\frac{1/p-1/q}{1/p-1/2}} \leq d^m(B_p^N, \ell_q^N) \leq E_{ada}^m(B_p^N, \ell_q^N) \leq \frac{C}{s^{1/p-1/q}}.$$

Considering only reasonable values of parameters (i.e. excluding the case where the minimum is 1), we have

$$\frac{C}{s^{1/p-1/q}} \geq \left\{\frac{N^{1-1/p}}{m^{1/2}}\right\}^{\frac{1/p-1/q}{1/p-1/2}}.$$

Now, raising both side to $2(1/p-1/2)/(1/p-1/q)$, we obtain

$$m \geq C^{\frac{-2(1/p-1/2)}{1/p-1/q}} s^{2(1/p-1/2)} N^{2-2/p} = \tilde{C} s^{2(1/p-1/2)} N^{2-2/p}$$

So, we can conclude that, for $p > 1$, a measurement/decoder scheme which is (ℓ_q, ℓ_p)-instance optimal is not achievable in the regime $m \asymp s \ln(eN/m)$. The same analysis could be done to regions III and IV. Therefore, we need to look for $p \leq 1$ in order to use reasonable techniques for compressive sensing.

Bibliography

[] Books:

[Andrews, Askey & Roy '99] George Andrews, Richard Askey & Ranjay Roy, *Special Functions*, Encyclopedia of Mathematics and Its Applications, Volume 71, Cambridge University Press, Cambridge, 1999.

[Arora & Barak '07] Sanjeev Arora & Boaz Barak, *Computational Complexity: A Modern Approach*, Cambridge University Press, Cambridge, 2007.

[Artstein-Avidan, Giannopoulos & Milman '15] Shiri Artstein-Avidan, Apostolos Giannopoulos & Vitali D. Milman, *Asymptotic Geometric Analysis*, Mathematical Surveys and Monographs, American Mathematical Society, 2015.

[Bertsekas '16] Dimitri Bertsekas, *Nonlinear Programming*, Athena Scientific; 3rd edition, 2016.

[Björck '96] Ake Björck, *Numerical Methods for Least Squares Problems*, SIAM, Philadelphia, 1996.

[Boche et at. '15] Holger Boche (Ed), Robert Calderbank (Ed), Gitta Kutyniok (Ed), Jan Vybíral (Ed), *Compressed Sensing and its Applications: MATHEON Workshop 2013*, Birkhäuser, 2015.

[Boucheron, Lugosi & Massart '13] Stephane Boucheron, Gábor Lugosi & Pascal Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence*, Oxfod University Press, Oxford, 2013.

[Boyd & Vanderberghe '04] Stephen Boyd & Lieven Vanderberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[Buldygin & Kozachenko '98] Valerii V. Buldygin & Yuri V. Kozachenko, *Metric Characterization of Random Variables and Random Process*, Translations of Mathematical Monographs, Volume 188, American Mathematical Society, Providence, Rhode Island, 1998.

[Burrus, Gopinath & Guo '97] C. Sidney Burrus, Ramesh A. Gopinath & Haitao Guo, *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice-Hall, New Jersey, 1997.

[Casazza & Kutyniok '13] Peter G. Casazza(Ed) & Gitta Kutyniok(Ed) *Finite Frames: Theory and Applications*, Birkhäuser; 2013.

[Charpentier, Lesne & Nikolski '07] Eric Charpentier (Ed), Annick Lesne (Ed), Nikolaï K. Nikolski(Ed), *Kolmogorov's Heritage in Mathematics*, Springer, 2007.

[Christensen '08] Ole Christensen, *Frames and Bases: An Introductory Course*, Birkhäuser, 2008.

[Cohen '94] Leon Cohen, *Time Frequency Analysis: Theory and Applications*, Prentice Hall, 1994.

[Damelin & Miller '12] Steven B. Damelin & Willard Miller Jr, *The Mathematics of Signal Processing*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, 2012.

[Daubechies '92] Ingrid Daubechies, *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.

[DeGroot & Schervish '11] Morris H. DeGroot & Mark J. Schervish, *Probability and Statistics*. Pearson, Upper Saddle River, USA, 2011.

[Dubhashi & Panconesi '12] Devdatt P. Dubhashi & Alessandro Panconesi, *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge University Press, Cambridge, 2012.

[Durrett '10] Rick Durrett, *Probability: Theory and Examples*, Cambridge University Press, Cambridge, 2010.

[Elad '10] Michael Elad, *Sparse and Reduntant Representations: From Theory to Applications in Signal and Imaging Processing*, Springer, 2010.

[Eldar '15] Yonina Eldar, *Sampling Theory: Beyond Bandlimited Systems*, Cambridge University Press, Cambridge, 2015.

[Eldar & Kutyniok '12] Yonina C. Eldar(Ed) & Gitta Kutyniok(Ed), *Compressed Sensing: Theory and Applications*, Cambridge University Press, 2012.

[Feller '68] William Feller, *An Introduction to Probability Theory and Its Applications, Volume 1*, 3rd Edition, John Wiley & Sons, New York, 1968.

[Feller '72] William Feller, *An Introduction to Probability Theory and Its Applications, Volume 2*, 3rd Edition, John Wiley & Sons, New York, 1972.

[Fischer '11] Hans Fischer, *A History of the Central Limit Theorem: From Classical to Modern Probability Theory*, Sources and Studies in the History of Mathematics and Physical Sciences, Springer, New York, 2011.

[Gamkrelidze '90] Revaz Gamkrelidze, *Analysis II: Convex Analysis and Approximation Theory*, Encyclopaedia of Mathematical Sciences (Book 14). Springer, 1990.

[Garey & Johnson '79] M. R. Garey & D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-completeness*. Freeman and Company, 1979.

[Gass & Assad '04] Saul I. Gass & Arjang A. Assad, *An Annotated Timeline of Operations Research: An Informal History*. Springer Verlag, 2004.

[Goldreich '08] Oded Goldreich, *Computational Complexity: A conceptual Perspective*. Cambridge University Press, Cambridge, 2008.

[Graham, Knuth & Patashnik '94] Ronald L. Graham, Donald E. Knuth & Oren Patashnik, *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Professional, 1994.

[Gröchenig '01] Karlheinz Gröchenig, *Foundations of Time-Frequency Analysis*, Birkhäuser, 2001.

[Guillemin & Pollack] Victor Guillemin & Allan Pollack, *Differential Topology*. Prentice Hall, 1974.

[Hacking '06] Ian Hacking, *The Emergence of Probability: A Philosophical Study of Early Ideas About Probability Induction and Statistical Inference*. Cambridge University Press, Cambridge, 2006.

[Han et al. '07] Deguang Han, Keri Kornelson, David Larson & Eric Weber *Frames for Undergraduates*, American Mathematical Society, 2007.

[Hardy, Littlewood & Polya] Godfrey H. Hardy, John E. Littlewood & George Polya, *Inequalities"*, Cambridge University Press, Second Edition, 1952.

[Hastie, Tibshirani & Wainwright '15] Trevor Hastie, Robert Tibshirani & Martin Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.

[Havin & Jöricke '94] Victor Havin & Burglind Jöricke, *The Uncertainty Principle in Harmonic Analysis*, Springer, 1994.

[Jammer '96] Max Jammer, *The Conceptual Development of Quantum Mechanics*, McGraw-Hill, 1966.

[Jammer '74] Max Jammer, *The Philosophy of Quantum Mechanics: The Interpretations of Quantum Mechanics in Historical Perspective*, Wiley, 1974.

[Kirsch '11] Andreas Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Applied Mathematical Sciences, Vol. 120, Springer, 2011.

[Ledoux '01] Michel Ledoux, *The Concentration of Measure Phenomenon*, Mathematical Surveys and Monographs 89, American Mathematical Society, Providence, RI, 2001.

[Lorentz '66] G. G. Lorentz, *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[Lorentz, von Golitschek & Makovoz '96] G. G. Lorentz, M von Golitschek & Y. Makovoz, *Constructiive Approximation: Advanced Problems*, Springer, 1996.

[Marvasti '01] Farokh Marvasti, *Nonuniform Sampling: Theory and Practice*. Information Technology: Transmission, Processing and Storage Series, Springer, New York, 2001.

[Mehra & Rechenberg '01] Jagdish Mehra & Helmut Rechenberg, *The Historical Development of Quantum Mechanics*, Springer, 2001.

[Micchelli & Rivlin '76] Charles A. Micchelli & Theodore J. Rivlin, *Optimal Estimation in Approximation Theory*, Springer ,1976.

[Nesterov & Nemirovskii '94] Yurii E. Nesterov & Arkadii S. Nemirovskii, *Interior Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, 1994.

[Novak & Wozniakowski '08] Erich Novak & Henryk Wozniakowski, '*Tractability of Multivariate Problems: Linear Information*, European Mathematical Society, 2008.

[Novak & Wozniakowski '10] Erich Novak & Henryk Wozniakowski, *Tractability of Multivariate Problems: Standard Information for Functionals*, European Mathematical Society, 2010.

[Novak & Wozniakowski '12] Erich Novak & Henryk Wozniakowski, *Tractability of Multivariate Problems: Standard Information for Operators*, European Mathematical Society, 2012.

[Pinkus '85] Allan Pinkus, *N-Width in Approximation Theory*, Springer, Berlin, 1985.

[Pisier '89] Gillies Pisier, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge Tracts in Mathematics 94, Cambridge University Press, Cambridge, 1989.

[Pietsch '07] Albrecht Pietsch, *History of Banach Spaces and Linear Operators*, Birkhäuser, 2007.

[Rasmussen & Williams '05] Carl Edward Rasmussen & Christopher K. I. Williams, *Gaussian Processes for Machine Learning*, The MIT Press, 2005.

[Rauhut & Foucart '13] Holger Rauhut & Simon Foucart, *A Mathematical Introduction to Compressive Sensing*, Birkhäuser, 2013.

[Resnick '13] Sidney Resnick, *A Probability Path*, Birkhäuser, 2013.

[Rish & Grabarnik '14] Irina Rish & Genady Grabarnik, *Sparse Modeling: Theory, Algorithms and Applications*, CRC Press, 2014.

[Ruszczynski] Andrzej Ruszczynski, *Nonlinear Optimization*, Princeton University Press, 2006.

[Schulz, da Silva & Velho '09] Adriana Schulz, Eduardo A. B. da Silva & Luiz Velho, *Compressive Sensing*, 27° Colóquio Brasileiro de Matemática, IMPA, Rio de Janeiro, 2009.

[Sra, Nowozin & Wright '11] Suvrit Sra, Sebastian Nowozin & Stephen J. Wright, *Optimization for Machine Learning*, The MIT Press, 2011.

[Steffens '06] Karl-Georg Steffens, *The History of Approximation Theory: From Euler to Bernstein*, Birkhäuser, 2006.

[Stein & Shakarchi '05] Elias Stein & Rami Shakarchi, *Real Analysis: Measure Theory, Integration, and Hilbert Spaces*. Princeton Lectures in Analysis, Volume III, Princeton University Press, Nova Jersey, 2005.

[Stein & Shakarchi '11] Elias Stein & Rami Shakarchi, *Functional Analysis: Introduction to Further Topics in Analysis*. Princeton Lectures in Analysis, Volume IV, Princeton University Press, Nova Jersey, 2005.

[Stigler '90] S. M. Stigler, *The History of Statistics: The Measurement of Uncertainty Before 1900*. Belknap Press, 1990.

[Temlyakov '11] *Greedy Approximation*, Cambridge University Press, 2011.

[Tong '80] Yung Liang Tong, *Probability Inequalities in Multivariate Distributions*, Academic Press, New York, 1980.

[Tropp '15] Joel Tropp, *An Introduction to Matrix Concentration Inequalities*, Foundations and Trends in Machine Learning Volume 8, Now Publishers Inc, 2015.

[von Plato '98] Jan von Plato, *Creating Modern Probability: Its Mathematics, Physics and Philosophy in Historical Perspective*, Cambridge University Press , Cambridge, 1998.

[Young '80] Robert Young, *An Introduction to Non-Harmonic Fourier Series*. Academic Press, New York, 1980.

[Varga '04] Richard S. Varga, *Gershgorin and His Circles*. Springer-Verlag, Berlin, 2004.

[Vempala '04] Santosh Vempala, *The Random Projection Method*. AMS, 2004.

[Vershynin '16] Roman Vershynin, *High Dimensional Probability: An Introduction with Applications in Data Science*, to be published by Cambridge University Press, 2016. Available at `http://www-personal.umich.edu/~romanv/papers/HDP-book/HDP-book.pdf`.

Historical:

[Borges '60] Jorge Luis Borges, *El Hacedor*, Biblioteca Borges, Alianza Editorial, 1960.

[Carathéodory 1911] Constantin Carathéodory, *Über den Variabilitätsbereich der Fourierschen Funktionen*. Rendiconto del Circolo Matematico di Palermo, Volume 32, 193-217, 1911.

[Dorfman '43] Robert Dorfman, *The Detection of Defective Members of Large Populations*. The Annals of Mathematical Statistics, Volume 14, Issue 4, 436-440, 1943.

[Gabor '46] Dennis Gabor *Theory of communication*, J. Inst. Elec. Engr. 93, 429-457, 1946.

[Gershgorin '31] S. Gerschgorin *Über die Abgrenzung der Eigenwerte einer Matrix*, Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk 6, 749-754, 1931.

[Heisenberg '27] Werner Heisenberg *Über den anschaulichen Inhalt der quantentheoretischen Kinematic und Mechanik*, Zeit. Phyisik, 43, 172-198, 1927.

[Kennard '27] Earle H. Kennard *Zur Quantenmechanik einfacher Bewegungstypen*, Zeit. Physik, 44, 326-352, 1927.

[Levy & Fullagar '81] S. Levy & P. Fullagar, *Reconstruction of a Sparse Spike Train From a Portion of Its Spectrum and Applications to High-Resolution Deconvolution.* Geophysics, Volume 46, Number 9, 1235-1243, 1981.

[Santosa & Symes '86] Fadil Santosa & William W. Symes, *Linear Inversion of Band-Limited Reflection Seismograms.* SIAM Journal on Scientific and Statistical Computing, Volume 7, Issue 4, 1307-1330-1986.

[Shannon '48] Claude E. Shannon, *"A Mathematical Theory of Communication".* Bell System Technical Journal, Volume 27, 379-423, 1948.

[Slepian '62] David Slepian, *The One-Sided Barrier Problem for Gaussian Noise.* Bell System Technical Journal, Volume 41, 463-501, 1962.

[Taylor, Banks & McCoy '79] H. L. Taylor, S. C. Banks & J. F. McCoy, *Deconvolution with the ℓ_1-norm.* Geophysics, Volume 44, Number 1, 39-53, 1979.

[Walker & Ulrych '83] C. Walker & T. Ulrich, *Autoregressive Recovery of the Acoustic Impedance.* Geophysics, Volume 48, Number 10, 1338-1350, 1983.

[Weyl '28] Hermann Weyl, *Gruppentheorie und Quantenmechanik*, S. Hirzel, 1928. Revised English edition: The Theory of Groups and Quantum Mechanics, Methuen, London, 1931; reprinted by Dover, New York, 1950.

[Wiener '56] Norbert Wiener, I am a Mathematician, MIT Press, Cambridge, 1956.

[von Neumann '53] John von Neumann, *A Certain Zero-Sum Two-Person Game Equivalent to the Optimal Assignment Problem*, in Contributions to the Theory of Games II, H.W. Kahn and A.W. Tucker (editors), 5-12. Princeton University Press, Princeton, NJ, 1953.

Papers:

[Adcock, Hansen & Roman '15] Ben Adcock, Anders Hansen & Bogdan Roman, *The Quest for Optimal Sampling: Computationally Efficient, Structure-Exploiting Measurements for Compressive Sensing* in Compressed Sensing and Its Applications: MATHEON Workshop 2013, Birkhäuser, 2015.

[Adcock, Hansen, Roman & Teschke '13] Ben Adcock, Anders Hansen, Bogdan Roman & Gerd Teschke, *Generalized Sampling: Stable Reconstructions, Inverse Problems and Compressed Sensing Over the Continuum.* Advances in Imaging and Electron Physics, Volume 182, 187-279, 2013.

[Ailon & Chazelle '06] Nir Ailon & Bernard Chazelle, *Approximate Nearest Neighbors and the Fast Johnson-Lindenstrauss Transform.* In *Proceedings of the 38th ACM Symposium on Theory of Computing (STOC)*, 557-563, 2006.

[Ailon & Chazelle '09] Nir Ailon & Bernard Chazelle, *The Fast Johnson-Lindenstrauss Transform and Approximate Nearest Neighbors.* SIAM Journal on Computing, Volume 39, Issue 1, 302-322, 2009.

[Ailon & Liberty '11] Nir Ailon & E. Liberty, *Almost Optimal Unrestricted Fast Johnson-Lindenstrauss transform.* In Proceedings of the 22nd Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), San Francisco, USA, 22-2, 2011.

[Alon '03] Noga Alon, *Problems and Results in Extremal Combinatorics I*, Discrete Mathematics, Volume 273(1-3), 31,53, 2003.

[Alltop '80] W. O. Alltop, *Complex Sequences With Low Periodic Correletions*, IEEE Transaction of Information Theory, Vol. 26, Issue 3, 350-354, 1980.

[Andersson & Strömberg '14] Joel Andersson & Jan-Olov Strömberg, *On the Theorem of Uniform Recovery of Random Sampling Matrices*, IEEE Transactions on Information Theory, Volume 60, Number 3, 1700-1710, 2014.

[Applebaum, Howard, Searle & Calderbank '09] Lorne Applebaum, Stephen D. Howard, Stephen Searle & Robert Calderbank, *Chirp Sensing Codes: Deterministic Compressed Sensing Measurements for Fast Recovery,*?Applied and Computational Harmonic Analysis, Volume 26, Issue 2, 283-290, 2009.

[Bah '12] Bubacarr Bah, *Restricted Isometry Constants in Compressed Sensing.* Ph.D. Thesis, University of Edinburgh, 2012.

[Bah & Tanner '10] Bubacarr Bah & Jared Tanner, *Improved Bounds on Restricted Isometry Constants for Gaussian Matrices*, SIAM Journal Matrix Analysis and Applications, Volume 31, Number 5, 2882-2898, 2010.

[Bandeira et al. '13] Afonso S. Bandeira, Matthew Fickus, Dustin G. Mixon & Percy Wong, *The Road to Deterministic Matrices with the Restricted Isometry Property*, Journal of Fourier Analysis and Applications, Volume 19, Issue 6, 1123-1149, 2013.

[Bandeira, Dobriban, Mixon & Sawin '13] Afonso S. Bandeira, Egdar Dobriban, Dustin G. Mixon & William F. Sawin, *Certifying the Restricted Isometry Property is Hard*, IEEE Transactions of Information Theorym Volume 59, Number 6, 3448-3450, 2013.

[Baraniuk '07] Richard Baraniuk, *Compressive Sensing.* IEEE Signal Processing Magazine, Volume 24, Issue 4, 118-121, 2007.

[Baraniuk et al. '08] R. G. Baraniuk, M. Davenport, R.A. DeVore & M. Wakin, *A Simples Proof of the Restricted Isometry Property for Random Matrices.* Constructive Approximations, Vol. 28, Issue 3, 253-263, 2008.

[Baraniuk et al.] R. G. Baraniuk, M. Davenport, R.A. DeVore & M. Wakin, *The Johnson-Lindenstrauss Lemma Meets Compressed Sensing.* Preprint.

[Bechar '09] Ikhlef Bechar, *A Bernstein-Type Inequality for Stochastic Processes of Quadratic Forms of Gaussian Variables*, Preprint 2009. Available at `https://arxiv.org/abs/0909.3595`.

[Bickel, Ritov & Tsybakov '09] Peter Bickel, Ya'acov Ritov & Alexandre Tsybakov, *Simultaneous Analysis of Lasso and Dantzig Selector.* Annals of Statistics, Volume 37, Issue 4, 1705-1732, 2009.

[Bioucas-Dias & Figueiredo '07] José M. Bioucas-Dias & Mário A. T. Figueiredo, *A New TwIST: Two-Step Iterative Shrinkage/Thresholding Algorithms for Image Restoration*, IEEE Transactions on Image Processing, Volume 16, Issue 12, 2992-3004, 2007.

[Blanchard, Cartis & Tanner '11] Jeffrey D. Blanchard, Coralia Cartis & Jared Tanner, *Compressed sensing: How Sharp is the Restricted Isometry Property?*, SIAM Review, Volume 53, Issue 1, 105-125, 2011.

[Blanchard & Tanner '15] Jeffrey D. Blanchard & Jared Tanner, *Performance Comparisons of Greedy Algorithms in Compressed Sensing*, Volume 22, Issue 2,2 54-282, 2015.

[Blumensath & Davies '08] Thomas Blumensath & Mike Davies, *Gradient Pursuits.* IEEE Transactions on Signal Processing, Volume 56, Issue 6, 2370-2382, 2008.

[Blumensath & Davies II '08] Thomas Blumensath & Mike Davies, *Iterative Thresholding for Sparse Approximations.* Journal of Fourier Analysis and its Applications, Volume 14, Issue 5, 629-654, 2008.

[Blumensath & Davies '09] Thomas Blumensath & Mike Davies, *Iterative Hard Thresholding for Compressed Sensing.* Applied and Computational Harmonic Analysis, Volume 27, Issue 3, 265-274, 2009.

[Bodmann & Kutyniok '09] Bernhard G. Bodmann & Gitta Kutyniok, *Erasure-Proof Transmissions: Fusion Frames meet Coding Theory*, Proc. SPIE 7446, Wavelets XIII, Volume 74460, 2009.

[Borup, Gribonval & Nielsen '08] L. Borup, R. Gribonval & M. Nielsen, *Beyond Coherence: Recovering Structured Time-Frequency Representations* . Appl. Comput. Harmon. Anal., Vol. 14, 120?128, 2008.

[Bouchot, Foucart & Hitczenko '16] Jean-Luc Bouchot, Simon Foucart & Pawel Hitczenko, *Hard Thresholding Pursuit Algorithms: Number of Iterations*, Applied and Computational Harmonic Analysis Volume 41, Issue 2, 412-435. 2016.

[Boufonos '12] Petros Boufonos, *Universal Rate-Efficient Scalar Quantization*, IEEE Transactions on Information Theory, Volume 58 Issue 3, 1861-1872, 2012.

[Boufounos, Kutyniok & Rauhut '11] Petros Boufounos, Gitta Kutyniok & Holger Rauhut, *Sparse Recovery From Combined Fusion Frame Measurements.* IEEE Transactions on Information Theory, Vol. 57, Issue 6, 3864-3876, 2011.

[Bourgain, Dilworth, Ford, Konyagin & Kutzarova '11] Jean Bourgain, Stephen Dilworth, Kevin Ford, Sergei Konyagin & Denka Kutzarova, *Explicit Constructions of RIP Matrices and Related Problems*, Duke Mathematical Journal, Volume 159, Number 1, 145-185, 2011.

[Box '79] George E. P. Box, *Robustness in the Strategy of Scientific Model Building*, in Launer, R. L.; Wilkinson, G. N., Robustness in Statistics, Academic Press, 201-236, 1979.

[Butzer & Stens '92] Paul L. Butzer & Rudolf L. Stens, *Sampling Theory for not Necessarily Band-Limited Functions - A Historical Overview.* SIAM Review, Volume 34, Number 1, 40-53, 1992.

[Cai, Wang & Xu I '10] T. Cai, L. Wang $ G. Xu, *Shifting Inequality and Recovery of Sparse Vectors*, IEEE Transactions Signal Processing, Vol.58, Issue 9, 4388-4394, 2010.

[Cai, Wang & Xu II '10] T. Cai, L. Wang $ G. Xu, *New Bounds for Restricted Isometry Constants.* IEEE Transactions on Information Theory, Volume 56, Issue 9, 4388-4394, 2010.

[Cai, Wang & Xu III '10] T. Cai, L. Wang $ G. Xu, *Stable Recovery of Sparse Signals and an Oracle Inequality*, IEEE Transactions on Information Theory, Vol. 56, Issue 7, 3516-3522, 2010.

[Cai & Wang '11] T. Cai & Lie Wang, *Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise*, IEEE Transactions on Information Theory, Vol. 57, Issue 7, 4680-4688, 2011.

[Cai & Zhang '13] T. Cai & Anru Zhang, *Sharp RIP Bound for Sparse Signal and Low-Rank Matrix Recovery*, Applied and Computational Harmonic Analysis, Volume 35, Issue 1, 74-93, 2013.

[Casazza, Redmond & Tremain '08] P.G. Casazza, D. Redmond & J.C.Tremain, *Real Equiangular Frames*, Conference on Information Sciences and Systems, Princeton, NJ, 2008.

[Candès '06] Emanuel J. Candès,, *Compressive Sampling.* In Proceedings of the International Congress of Mathematicians

[Candès '08] Emanuel J. Candès, *The Restricted Isometry Property and Its Implications for Compressed Sensing*, Comptes Rendus Mathematique, Volume 346, Issue 9-10, 589-592, 2008.

[Candès, Eldar, Strohmer & Voroninski '13] Emmanuel Candès, Yonina Eldar, Thomas Strohmer & Vladislav Voroniski, *Phase Retrieval via Matrix Completion*, SIAM Journal of Imaging Science, Volume 6, Issue 1, 199-225. 2013.

[Candès & Plan '11] Emmanuel Candès & Yaniv Plan, *A Probabilistic and RIPless Theory of Compressed Sensing*. IEEE Transactions on Information Theory, Volume 57, Issue 11, 7235-7254, 2011.

[Candès & Randall '08] Emmanuel J. Candès & Paige A. Randall, *Highly Robust Error Correction by Convex Programming*, IEEE Transactions on Information Theory, Volume 54, Issue 7, 2829-2840, 2008.

[Candès & Recht '09] Emmanuel J. Candès & Benjamin Recht, *Exact Matrix Completion via Convex Optimization*, Foundations of Computational Mathematics, Volume 9, 717-772, 2009.

[Candès & Romberg '07] Emanuel J. Candès & Justin K. Romberg, *Sparsity and Incoherence in Compressive Sampling*, Inverse Problems, Volume 23, Number 3, 969-985, 2007.

[Candès, Romberg & Tao I '06] Emanuel J. Candès, Justin K. Romberg & Terence Tao, *Robust Uncertainty Principles: Exact Signal Reconstruction From Highly Incomplete Frequency Information*, IEEE Transactions on Information Theory, Volume 47, Issue 2, 489-509, 2006.

[Candès, Romberg & Tao II '06] Emanuel J. Candès, Justin K. Romberg & Terence Tao *Stable Signal Recovery From Incomplete and Inaccurate Measurements*, Communications on Pure and Applied Mathematics, Volume 59, Issue 8, 1207-1223, 2006.

[Candès & Tao I '06] Emanuel J. Candès & Terence Tao *Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?*, IEEE Transactions on Information Theory, Vol.52, Issue 12, 5406-5425, 2006.

[Candès & Tao II '06] Emanuel J. Candès & Terence Tao *Decoding by Linear Programming*, IEEE Transactions Information Theory, Volume 51, Issue 12, 4203-4215, 2006.

[Candès & Tao III, 06] Emanuel J. Candès & Terence Tao *The Dantzig Selector: Statistical Estimation When p is Much Larger Than n*, The Annals of Statistics, 2313-2351, 2006.

[Candès & Tao '10] Emanuel J. Candès & Terence Tao *The Power of Convex Relaxation: Near-Optimal Matrix Completion*, IEEE Transactions Information Theory, Volume 56, Number 5, 2053-2080, 2010.

[Candès & Wakin '08] Emanuel Candès & Michael B. Wakin, *An Introduction to Compressive Sampling*, IEEE Signal Processing Magazine, Vol. 25, Issue 2, 21-30, 2008.

[Casazza '00] Peter G. Casazza, *The Art of Frame Theory*, Taiwanese Journal of Mathematics, Vol. 4, No. 2, 129-201, 2000.

[Chambolle & Pock '11] Antonin Chambolle & Thomas Pock, *A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging*, Journal of Mathematical Imaging and Vision, Volume 40, 120-145, 2011.

[Chandrasekaran, Recht, Parrilo & Willsky '12] V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, *The Convex Geometry of Linear Inverse Problems*. Foundations of Computational Mathematics, Volume 12, Issue 6, 805-849, 2012.

[Chartrand '07] Rick Chartrand, *Exact Reconstruction of Sparse Signals via Nonconvex Minimization*, IEEE Signal Processing Letters, Volume 14, Number 10, 707-710, 2007.

[Chen, Billings & Luo '89] Shaobing Scott Chen, S. Billings & W. Luo, *Orthogonal Least Squares Methods and Their Application to Nonlinear System Identification*. International Journal of Control, Volume 50, Issue 5, 1873-1896, 1989.

[Chen & Donoho '94] Shaobing Scott Chen & David Donoho, *Basis Pursuit*. Technical Report, Department of Statistics, Stanford University.

[Chen, Donoho & Saunders '01] Shaobing Scott Chen, David L. Donoho & Michael A. Saunders, *Atomic Decomposition by Basis Pursuit*, SIAM Journal on Scientific Computing, 20 (1), 33-61, 2001.

[Chen '95] Shaobing Scott Chen, *Basis Pursuit*, Ph.D. Thesis, Stanford, 1995.

[Cohen, Dahmen & DeVore '09] Albert Cohen, Wolfgang Dahmen & Ronald DeVore, *Compressed Sensing and Best k-term Approximation*, Journal of the AMS, Vol. 22, Number 1, 211-231, 2009.

[Cook] Stephen Cook, *The P Versus NP Problem.* Available online: `http://www.claymath.org/sites/default/files/pvsnp.pdf`.

[Dai & Milenkovic '09] W. Dai & O. Milenkovic, *Subspace Pursuit for Compressive Sensing Signal Reconstruction.* IEEE Transactions on Information Theory, Volume 55, Issue 5, 2230-2249, 2009.

[Dasgupta, Kumar & Sarlós '10] Anirban Dasgupta, Ravi Kumar & Tamás Sarlós, *A Sparse Johnson-Lindenstrauss Transform.* In *Proceedings of the 42nd ACM Symposium on Theory of Computing (STOC)*, 205-214, 2010.

[Daubechies, Defrise & De Mol '04] Ingrid Daubechies, Michel Defrise & Christine De Mol, *An Iterative Thresholding Algorithm for Linear Inverse Problems With a Sparsity Constraint.* Communications in Pure and Applied Mathematics, Volume 63, Issue 1, 1-38, 2010.

[Daubechies, DeVore, Fornasier & Güntürk '10] Ingrid Daubechies, Ronald DeVore, Maximo Fornasier & Sinan Güntürk, *Iteratively Re-Weighted Least Squares Minimization for Sparse Recovery*, Communications in Pure and Applied Mathematics, Volume 63, Issue 1, 1-38, 2010.

[Daubechies, Grossman & Meyer '85] Ingrid Daubechies, A. Grossman, & Yves Meyer, *Painless Nonorthogonal Expansions*, Journal of Mathematical Physics, Volume 127, 1271-1283, 1985.

[Davenport & Romberg '16] Mark Davenport & Justin Romberg, *An Overview of Low-Rank Matrix Recovery From Incomplete Observations*, IEEE Journal of Selected Topics in Signal Processing, Volume 10, Number 4, 608-622, 2016.

[Davidson & Szarek '01] K. Davidson & S. Szarek, *Local Operator Theory, Random Matrices and Banach Spaces.* Handbook of the Geometry of Banach Spaces, 1-317, 2001.

[Davies & Gribonval '09] M. Davies & R. Gribonval, *Restricted Isometry Constants where ℓ^p Sparse Recovery can Fail for $0 < p \leq 1$*, IEEE Transaction on Information Theory, Vol. 55, Issue 5, 2203-2214, 2009.

[Davies, Mallat & Avellaneda '97] Geoffrey Davies, Stephane Mallat & G. Davis, S. Mallat, and M. Avellaneda. Greedy adaptive approximation. Journal Constructive Approximation, Volume 13, 57-98, 1997.

[Davies, Mallat & Zhang '94] Geoffrey Davies, Stephane Mallat & Zhifeng Zhang, *Adaptive Time-Frequency Decompositions.* Optical Engineering, Volume 33, Issue 7, 2183-2191, 1994.

[Davies & Simon '84] Edward B. Davies & Barry Simon, *Ultracontractivity and the Heat Kernel for Schrödinger Operators and Dirichlet Laplacians*, Journal of Functional Analysis, Volume 59, Issue 2, 335-395, 1984.

[DeVore '07] Ronald DeVore, *Deterministic constructions of compressed sensing matrices*, Journal of Complexity, Volume 23, Issues 4?6, 918-925, 2007.

[DeVore '06] Ronald DeVore, *Optimal Computation*, Proceedings of the International Congress of Mathematicians, European Mathematical Society, Madrid, Spain, 2006.

[DeVore & Temlyakov '96] Ronald DeVore & Vladimir Temlyakov, *Some Remarks on Greedy Algorithms*, Advances in Computational Mathematics, Volume 5, 173-187, 1996.

[Dominici '08] Diego Dominici, *Variations on a Theme by James Stirling*, Note di Matematica Volume 28, Issue 1, 2008.

[Donoho '06] David L. Donoho, *Compressed Sensing*, IEEE Transactions on Information Theory, Volume 52, Issue 4, 1289-1306, 2006.

[Donoho '10] David L. Donoho, *Scanning the Technology*, Proceedings of the IEEE, VOlume 98, Number 6, 910=912, 2010.

[Donoho & Elad ' 03] David. L. Donoho & Michael Elad, *Optimally Sparse Representation in General (Nonorthogonal) Dictionaries via l1 Minimization*, Proc. Nat. Acad. Sci. USA, Vol. 100, 2197-2202, 2003.

[Donoho & Huo '01] David L. Donoho & Xiaoming Huo, *Uncertainty Principles and Ideal Atomic Decomposition*, IEEE Transactions on Information Theory, Volume 47, Issue 7, 2845-2862, 2001.

[Donoho, Johnstone, Stern & Hoch '92] David L. Donoho, Iain M. Johnstone, Alan S. Stern & Jeffrey C. Hoch, *Maximum entropy and the nearly black object (with discussion)*. Journal of the Royal Statistical Society B, Volume 54, Number 1, 41-81, 1992.

[Donoho & Kutyniok '13] David L. Donoho & Gitta Kutyniok *Microlocal Analysis of the Geometric Separation Problem*, Communications on Pure and Applied Mathematics, Vol. 66, Issue 1, pages 1-47, 2013.

[Donoho & Stark '89] David L. Donoho & Philip B. Stark, *Uncertainty Principles and Signal Recovery*, SIAM J. Appl. Math., Volume 49, Issue 3, 906-931, June 1989. DOI:10.1137/0149053

[Donoho & Tsaig, '08] David Donoho & Yaakov Tsaig, *Fast Solution of ℓ_1-norm Minimization Problems When the Solution May Be Sparse*, IEEE Transactions on Information Theory, Volume 54, Issue 11, 4789-4812, 2008.

[Donoho, Tsaig, Drori & Starck '12] David Donoho, Yaakov Tsaig, Iddo Drori & Jean-Luc Starck, *Sparse Solution of Underdetermined Systems of Linear Equations by Stagewise Orthogonal Matching Pursuit*, IEEE Transactions on Information Theory, Volume 58, Issue 2, 1094-1121, 2012.

[Duarte & Eldar '11] Marco F. Duarte & Yonina C. Eldar, *Structured Compressive Sensing: Theory and Applications*. IEEE Transactions on Signal Processing, Volume 59, Issue 9, 4053-4085, 2011.

[Dudley '75] Richard Dudley, *The Gaussian Process and How to Approach It*, Proceedings of the International Congress of Mathematicians, Vancouver, 1974.

[Duffin & Schaeffer '52] Duffin, R., Schaeffer, A. *A class of Nonharmonic Fourier series*, Trans. Am. Math. Soc. Vol. 72, 341-366, 1952.

[Efron, Hastie, Johnstone & Tibshirani '04] Bradley Efron, Trevor Hastie, Iain Johnstone & Robert Tibshirani, *Least Angle Regression*, Annals of Statistics, Volume 32, Issue 2, 407-499, 2004.

[Elad & Bruckstein '02] Michael Elad & Alfred Bruckstein, *A Generalized Uncertainty Principle and Sparse Representation in Pairs of Bases*, IEEE Transaction on Information Theory, Vol. 48, Issue 9, 2558-2567, 2002.

[Fazel '02] Maryam Fazel, *Matrix rank Minimization with Applications*, Ph.D. Thesis, Stanford University, 2002.

[Fazel, Hindi & Boyd '01] Maryam Fazel, Haitham Hindi & Stephen Boyd, *A Rank Minimization Heuristic with Application to Minimum Order System Approximation*, in Proceedings of the American COntrol Conference, IEEE, 4734-4739, 2001.

[Fefferman '83] Charles L. Fefferman, *The Uncertainty Principle*, Bull. Amer. Math. Soc, Vol. 9, 129-206, 1983.

[Feller '45] William Feller, *The Fundamental Limit Theorems in Probability*, Bulletin of the American Mathematical Society, Volume 51, Number 11, 800-832, 1945.

[Feng & Zhang '07] Xinlong Feng & Zhinan Zhang, *The Rank of a Random Matrix*, Applied Mathematics and Computation, Volume 185, Issue 1, 689-694, 2007.

[Ferreira & Higgins '11] Paulo J. S. G. Ferreira & Rowland Higgins, *The Establishment of Sampling as a Scientific Principle: A Striking Case of Multiple Discovery.* Notices of the AMS, Volume 58, Number 10, 1446-1450, 2011.

[Fickus & Mixon '15] Matthew Fickus & Dustin G. Mixon, *Tables of the Existence of Equiangular Tight Frames*, Preprint 2015. Available at `http://arxiv.org/abs/1504.00253`

[Folland & Sitaram '97] Gerald B. Folland & Alladi Sitaram, *The Uncertainty Principle: A Mathematical Survey*, Journal of Fourier Analysis and Applications, Volume 3, Issue 3, 207-238, May 1997.

[Fortnow & Homer '03] Lance Fortnow & Steve Homer, *A Short History of Computational Complexity* Bulletin of the European Association for Theoretical Computer Science, Volume 80, 1-26, 2003.

[Foucart '11] Simon Foucart, *Hard Thresholding Pursuit: An Algorithm for Compressive Sensing.* SIAM Journal of Numerical Analysis, Volume 49, Issue 6, 2543-2563, 2011.

[Foucart II '10] Simon Foucart, *A Note on Guaranteed Sparse Recovery Via ℓ_1-Minimization*, Applied and Computational Harmonic Analysis, Volume 29, Issue 1, 97-103, 2010.

[Foucart '10] Simon Foucart, *Sparse Recovery Algorithms: Sufficient Conditions in Terms of Restricted Isometry Constants*, In Approximation THeory XIII: San Antonio 2010, 65-77, ed. by M. Neamtu, L. Schumaker. Springer Proceedings in Mathematics, Volume 13, Springer, New York, 2012.

[Foucart & Gribonval '10] Simon Foucart $ Remi Gribonval, *Real vs. Complex Null Space Properties for Sparse Vector Recovery.* Comptes Rendus de L'académie des Sciences Mathématiques, Volume 348, Issue 15-16, 863-865, 2010.

[Foucart & Lai '10] Simon Foucart & Ming-Jun Lai, *Sparsest Solutions of Underdetermined Linear Systems via ℓ_q-minimization for* $0 < q1$, Applied and Computational Harmonic Analysis, Volume 26, Issue 3, 395-407, 2009.

[Foucart, Pajor, Rauhut & Ullrich '10] S. Foucart, A. Pajor, H. Rauhut& T. Ullrich, *The Gelfand Widths onf ℓ_p-balls for* $0 < p \leq 1$, Journal of Complexity, Vol. 26, Issue 6, 629-640, 2010.

[Fowler '00] David Fowler, *The Factorial Function: Stirling's Formula*, The Mathematical Gazette, Volume 84, Number 499, 42-50, 2000.

[Friedman & Stuetzle '81] J. Friedman & W. Stuetzle, *Projection Pursuit Regression.* Journal of the American Statistical Association,Volume 76, 817-823, 1981.

[Gantz & Reinsel '10] John Gantz & David Reinsel, *The Digital Universe - Are You Ready?.* IDC Report, Sponsored by EMC Corporation. Available at `https://www.emc.com/collateral/analyst-reports/idc-digital-universe-are-you-ready.pdf`.

[Gantz & Reinsel '12] John Gantz & David Reinsel, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and BIggest Growth in the Far East.* IDC Report, Sponsored by EMC Corporation. Available at `https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf`.

[Gao, Peng, Yue & Zhao '15] Yi Gao, Jigen Peng, Shigang Yue & Yuan Zhao, *On the Null Space Property of ℓ_q-Minimization for $0 < q \leq 1$ in Compressed Sensing*, Journal of Function Spaces, Volume 10, 1-10, 2015.

[Garnaev & Gluskin '84] A. Y. Garnaev & E. D. Gluskin, *On Widths of the Euclidean Ball.* Soviet Mathematics - Doklady, Vol. 30, 200-203, 1984.

[Ge, Jiang & Ye '11] Dongdong Ge, Xiaoye Jiang & Yinyu Ye, *A Note on the Complexity of L_p Minimization*, Mathematical Programming, Volume 129, Issue 2, 285-299, 2011.

[van de Geer & Buhlmann '09] Sara A. van de Geer & Peter Bühlmann, *On the Conditions Used to Prove Oracle Results for the LASSO*, Electronic Journal of Statistics, Volume 3, 1360-1392, 2009.

[Geman '80] Stuart Geman, *A Limit Theorem for the Norm of Random Matrices*, Annals of Probability, Volume 8, 252-261, 1980.

[Gilbert, Iwen & Strauss '08] Anna C. Gilbert, Mark A. Iwen, Martin J. Strauss, *Group Testing and Sparse Signal Recovery.* Proceedings of the 42nd Asilomar Conference on Signals, Systems and Computers, 1059-1063, 2008.

[Gluskin '84] E. D. Gluskin, *Norms of Random Matrices and Widths of Finite-Dimensional Sets.* Mathematics of the USSR-Sbornik, Volume 48, Number 1, 173-182, 1984.

[Gondzio '12] Jacek Gondzio, *Interior Point Methods 25 Years Later*, European Journal of Operational Research, Volume 218, 587-601, 2012.

[Gordon '85] Yehoram Gordon, *Some Inequalities for Gaussian Processes and Applications.* Israel Journal of Mathematics, Volume 50, Number 4, 1985.

[Gordon '88] Yehoram Gordon, *On Milman's Inequality and Random Subspaces Which Escape Through a Mesh in $\mathbb{R}^n$.* In Geometric Aspects of Functional Analysis (1986/87), Volume 1317 of Lecture Notes in Mathematics, 84-106. Springer, Berlin, 1988.

[Graham & Sloane '80] Ronald Graham & Neil Sloane, *Lower Bounds for Constant Weight Codes.* IEEE Transactions on Information Theory, Vol. 26, Issue 1, 37-43, 1980.

[Gribonval & Nielsen '03] Remi Gribonval & Morten Nielsen, *Sparse Representation in the Union of Basis*, IEEE Transactions on Information Theory, Volume 49, Issue 12, 3320-3325, 2003.

[Gribonval & Nielsen '07] Remi Gribonval & Morten Nielsen, *Highly Sparse Representations From Dictionaries are Unique and Independent of the Sparseness Measure*, Applied and Computational Harmonic Analysis, Volume 22, Issue 3, 335-355, 2007.

[Gross '75] Leonard Gross, *Logarithmic Sobolev Inequalities.* American Journal of Mathematics, Volume 97, Number 4, 1061-1083, 1975.

[Güntürk '00] C. Sinai Güntürk, *Harmonic Analysis of Two Problems in Signal Quantization and Compression*, Ph. D. Thesis, Princeton University, 2000.

[Haboba et al. '12] Javier Haboba, Mauro Mangia, Fabio Pareschi, Riccardo Rovatti & Gianluca Setti, *A Pragmatic Look at Some Compressive Sensing Architectures With Saturation and Quantization*, IEEE Journal on Emerging and Selected Topics in Circuits and Systems, Vol.2, Number 3, 443-459, 2012.

[Hanson & Wright '71] D. L. Hanson & F. T. Wright, *A Bound on Tail Probabilities for Quadratic Forms in Independent Random Variables*, The Annals of Mathematical Statistics, Volume 42, Number 3, 1079-1083, 1971.

[Hartmanis & Stearns '65] Juris Hartmanis & RIchard E. Stearns, *On the Computational Complexity of Algorithms.* Transactions of the American Mathematical Society, Volume 117, 285-306, 1965.

[Higgins '85] John Rowland Higgins, *Five Short Stories About the Cardinal Series.* Bulletin of the American Mathematical Society. Volume 12, Number 1, 45-89, 1985.

[Hinrichs & Vybiral '11] Aicke Hinrichs & Jan Vybíral, *Johnson-Lindenstrauss Lemma for Circulant Matrices.* Random Structures & Algorithms, Volume 39, Issue 3, 391?398, 2011.

[Holmes & Paulsen ' 04] Roderick B. Holmes & Vern I. Paulsen, *Optimal Frames for Erasures*, Linear Algebra and Its Applications, Vol. 377, 31?51, 2004.

[Holtz '08] Olga Holtz, *Compressive Sensing: A Paradigm Shift in Signal Processing*, Preprint 2008. Available at `https://arxiv.org/abs/0812.3137`.

[Huang & Rao '11] Yanping Huang & Rajesh P. N. Rao, *Predictive Coding.* Wiley Interdisciplinary Reviews. Cognitive Science. Volume 2, Issue 5, 580-593, 2011.

[Jacques & Vandergheynst '11] Laurent Jacques & Pierre Vandergheynst, *Compressed Sensing: When Sparsity Meets Sampling*, in Optical and Digital Image Processing: Fundamentals and Applications (eds G. Cristobal, P. Schelkens and H. Thienpont), Wiley, Weinheim, Germany, 2011.

[James, Radchenko & Lv '09] Gareth M. James, Peter Radchenko & Jinchi Lv, *DASSO: Connections Between the Dantzig Selector and Lasso.* Journal of the Royal Statistical Society: Series B, Volume 71, 127-142. 2009.

[Jameson '15] Graham J. O. Jameson, *A Simple Proof of Stirling's Formula for the Gamma Function*, The Mathematical Gazette, Volume 99, Issue 544, 68-74, 2015.

[Jerri '77] Abdul J. Jerri, *The Shannon Sampling Theorem: Its Various Extensions and Applications, A Tutorial Review*, Proceedings of the IEEE, Volume 65, Number 11, 1565-1596, 1977.

[Johnson & Lindenstrauss '84] W. B. Johnson & J. Linderstrauss, *Extensions of Lipschitz Mappings Into a Hilbert Space.* In *Conference in Modern Analysis and Probability (New Haven, Conn. 1982).* Contemporary Mathematics, Vol.26, 189-206, AMS, 1984.

[Kabanava, Kueng, Rauhut & Terstiege '16] Maryia Kabanava, Richard Kueng, Holger Rauhut & Ulrich Terstiege, *Stable Low-Rank Matrix Recovery Via Null Space Properties*, Information and Inference, Volume 5, Issue 4 ????? COMPLETAR ISSO

[Kahane '60] Jean-Pierre Kahane, *Proprietes Locales des Fonctions a Series de Fourier Aleatoires*, Studia Mathematica, Volume 19, 1-25, 1960.

[Kahane'86] Jean-Pierre Kahane, *Une Inegalité du Type de Slepian et Gordon Sur Les Processus Gaussiens*, Israel Journal of Mathematics, Volume 55, 109-110, 1986.

[Karp '72] Richard M. Karp, *Reducibility Among Combinatorial Problems.* In R.E. Miller and J.W. Thatcher (editors). Complexity of Computer Computations. Proc. of a Symp. on the Complexity of Computer Computations, 85-103, 1972.

[Kashin '77] B. S. Kashin, *Diameters of Some Finite-Dimensional Sets in Classes of Smooth Functions.* Izvestiya Rossiiskoi Akademii Nauk, Seriya Matematicheskaya, Vol. 11, no.2, 317-333, 1977.

[Keshavan, Montanari & Oh '10] Raghunandan Keshavan, Andrea Montanari & Sewoong Oh, *Matrix Completion From a Few Entries.* IEEE Transactions on Information Theory, Volume 56, Number 6, 2980-2998, 2010.

[Kim, Koh, Lustig, Boyd & Gorinevsky '08] Seung-Jean Kim, Kwangmoo Koh, Michael Lustig, Stephen Boyd & Dimitry Gorinevsky, *An Interior-Point Method for Large-Scale-Regularized Least Squares*, IEEE Journal of Selected Topics in Signal Processing, Volume 1, Issue 4, 606-617, 2008.

[Kim & Shevlyakov '08] Kiseon Kim & Georgy Shevlyakov, *Why Gaussianity?*. IEEE Signal Processing Magazine, Volume 25, Number 2, 2008.

[Kitsos & Tavoularis '09] Christos P. Kitsos & Nikolaos K. Tavoularis, *Logarithmic Sobolev Inequalities for Information Measures* . IEEE Transactions on Information Theory, Volume 55, Number 6, 2554-2561, 2009.

[Krahmer '09] Felix Krahmer, *Novel Schemes for Sigma-Delta Modulation: From Improved Exponential Accuracy to Low-Complexity Design*. Ph.D. Thesis, New York University, 2009.

[Krahmer & Ward '12] Felix Krahmer & Rachel Ward, *New and Improved Johnson-Linderstrauss Embeddings via the Restricted Isometry Property*. SIAM Journal of Mathematical Analysis, Vol.43, Issue 3, 1269-1281, 2011.

[Kolmogorov '36] Andrei Kolmogorov, *Uber die beste Annaherung von Funktionen einer gegebenen Funktionenklasse*, Annals of Mathematics, Volume 2, Issue 37, 107-110, 1936.

[Lai & Liu '11] Ming-Jun Lai & Yang Liu, *The Null Space Property for Sparse Recovery From Multiple Measurement Vectors*, Applied and Computational Harmonic Analysis, Volume 30, Issue 3, 402-406, 2011.

[Landau & Pollack '61] H. J. Landau & H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty-II*, Bell Syst. Tech. J, vol. 40, 65-84, 1961.

[Landau & Pollack '62] H. J. Landau & H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty-III:: The Dimension of the Space of Essentially Time- and Band-Limited Signals*, Bell Syst. Tech. J, vol. 41, 1295-1336, 1962.

[Larsen & Nelson '14] Kasper Larsen & Jelani Nelson, *The Johnson-Lindenstrauss Lemma is Optimal for Linear Dimensionality Reduction*, Preprint 2014. Available at `http://arxiv.org/abs/1411.2404`

[Lidskii '50] Victor B. Lidskii, *On the Characteristic Numbers of the Sum and Product of Symmetric Matrices*, Doklady Akad. Nauk SSSR (N.S.) Volume 75, 769-772, 1950.

[Linial, London & Rabinovich '15] Nathan Linial, Eran London & Yuri Rabinovich,*The Geometry of Graphs and Some of Its Algorithmic Applications* , Combinatorica, Volume 15, Issue 2, 215-245, 1995.

[Lorenz, Pfetsch & Tillmann '15] Dirk A. Lorenz, & Marc E. Pfetsch, *Solving Basis Pursuit: Heuristic Optimality Check and Solver Comparison*, ACM Transactions on Mathematical Software, Volume 41, Issue 2, 1-28, 2015. (An overview of updated computational results (as of April 2016) can be found at `http://www.mathematik.tu-darmstadt.de/~tillmann/docs/SolvingBPupdateApr2016.pdf`.)

[Lustig, Donoho & Pauly '07] Michael Lustig, David Donoho & John M. Pauly, *Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging*. Magnetic Resonance in Medicine, Volume 58, Issue 6, 1182-1195, 2007.

[Lv & Fan '09] Jinchi Lv & Yingying Fan, *A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares*, The Annals of Statistics, Volume 37, Number 6A, 3498-3528, 2009.

[Kutyniok '12] Gitta Kutyniok, *Theory and Applications of Compressed Sensing*, Unpublished notes. Available at `http://arxiv.org/pdf/1203.3815v2.pdf`

[Maleki & Donoho '10] Arian Maleki & David Donoho, *Optimally Tuned Iterative Reconstruction Algorithms for Compressed Sensing*, IEEE Journal of Selected Topics in Signal Processing, Volume 4, Issue 2, 330-341, 2010.

[Mallat & Zhang '93] Stephane Mallat & Zhifeng Zhang, *Matching Pursuit With Time-Frequency Dictionaries*, IEEE Transactions on Signal Processing, Vol. 41, Number 12, 3397-3415, 1993.

[Marcenko & Pastur '67] Vladimir Marcenko & Leonid Pastur, *Distributions of Eigenvalues for SOme Sets of Random Matrices*, Mathematics of the USSR-Sbornik, Volume 1, Number 4, 1967.

[Matousek '08] Jiri Matousek, *On Variants of the Johnson-Lindenstrauss Lemma*, Random Structures & Algorithms archive, Vol. 33, Issue 2, 142-156, 2008.

[Meijering '02] Erik Meijering, *A Chronology of Interpolation: From Ancient Astronomy to Modern Signal and Image Processing.* Proceedings of the IEEE, Volume 90, Issue 3, 319-342, 2002.

[Meinshausen, Rocha & Yu '07] N. Meinshausen, G. Rocha & B. Yu, *Discussion: A Tale of Three Cousins: Lasso, L2Boosting and Dantzig*, The Annals of Statistics, Volume 35, Number 6, 2373-2384, 2007.

[Mendelson, Pajor & Rudelson '05] S. Mendelson, A. Pajor & M. Rudelson, *The Geometry of Random* $\{-1,1\}$*-polytopes.* Discrete Mathematics, Vol. 34, Issue 3, 365-379. 2005.

[Mendelson, Pajor & Tomczak-Jaegermann '08] S. Mendelson, A. Pajor & N. Tomczak-Jaegermannm, *Uniform Uncertainty Principle for Bernoulli and Subgaussian Ensembles.* Constructive Approximation, Vol. 28, Issue 3, 277-289, 2008.

[Milman & Pajor '03] Vitali Milman & Alain Pajor, *Regularization of Star Bodies by Random Hyperplane Cut Off.* Studia Mathematica, Vol. 159, Issue 2, 247-261, 2003.

[Mishali & Eldar '11] Moshe Mishali & Yonina Eldar, *Sub-Nyquist Sampling: Bridging Theory and Practice.* IEEE Signal Processing Magazine, Volume 28, Issue 6, 98-124, 2011.

[Mixon '15] Dustin Mixon, *Explicit Matrices with the Restricted Isometry Preperty: Breaking the Square-Root Blottleneck* in Compressive Sensing and Its Applications, 389-417, Birkhäuser, 2015.

[Mo & Li '11] Qun Mo & Sung Li, *New Bounds on The Restricted Isometry Constant* δ_{2k}, Applied and Computational Harmonic Analysis, Volume 31, Issue 3, 460-468, 2011.

[Mo & Shen '12] Qun Mo & Yi Shen, *A Remark on the Restricted Isometry Property in Orthogonal Matching Pursuit*, IEEE Transactions on Information Theory, Volume 58, Issue 6, 3654-3656, 2012.

[Natarajan '96] Balas K. Natarajan, *Sparse Approximate Solutions to Linear Systems.* SIAM Journal on Computing, Volume 24, Issue 2, 227-234, 1996.

[Needell '09] Deanna Needell, *Topics in Compressed Sensing.*, Ph. D. Thesis, University of California, 2009.

[Needell & Tropp '08] Diane Needell & Joel Tropp, *CoSaMP: Iterative Signal Recovery From Incomplete and Inaccurate Samples*, Applied and Computational Harmonic Analysis, Volume 26, Issue 3, 201-321, 2008.

[Nelson '??] Jelani Nelson, *Johnson-Lindenstrauss Notes* , Unpublished notes. Available at `http://web.mit.edu/minilek/www/jl_notes.pdf`

[Noam & Avi '94] N. Noam & W. Avi, *Hardness vs Randomness.* Journal of Computer and Systems Sciences, Vol. 49, Issue 2, 149-167, 1994.

[Oliveira '13] Roberto I. Oliveira, *The Lower Tail of Random Quadratic Forms, With Applications to Ordinary Least Squares and Restricted Eigenvalue Properties*, Preprint 2013. Available at `http://arxiv.org/abs/1312.2903`

[Olshausen & Field '04] Bruno A. Olshausen & David J. Field, *Sparse Coding of Sensory Inputs.* Current Opinion in Neurobiology, Volume 14, 481-487, 2004.

[Pati, Rezaiifar & Krishnaprasad '93] Y. C. Pati, R. Rezaiifar & P. S. Krishnaprasad, *Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition*, Proceedings of the 27th Annual Asilomar Conference on Signals Systems and Computers, Nov. 1-3, 1993.

[Pearson '20] Karl Pearson, *Notes on the History of Correlation.* Biometrika, Volume 11, 145-158, 1920.

[Pope '09] Graeme Pope, *Compressive Sensing: A Summary of Reconstruction Algorithms*, Master's Thesis, Department of Computer Science, Eidgenössische Technische Hochschule, Zurich, 2009.

[Rao & Gorodnistsky '97] , Irina F. Gorodnistsky & Bhaskar D. Rao, *Sparse Signal Reconstruction From Limited Data Using FOCUSS: A Re-Weighted Norm Minimization Algorithm*, IEEE Transactions on Signal Processing, Vol. 4, Issue 3, 600-616, 1997.

[Recht, Fazel & Parrilo '10] Benjamin Recht, Maryam Fazel & Pablo A. Parrilo, *Guaranteed Minimum-Rank Solutions of Linear Matrix Equations via Nuclear Norm Minimization* SIAM Review, Volume 52, Issue 3, 471-501, 2010.

[Recht, Xu & Hassibi '08] Benjamin Recht, Weiyu Xu & Babak Hassibi, *Necessary and Sufficient Conditions for Success of the Nuclear Norm Heuristic for Rank Minimization*, in Proceedings of 47th IEEE Conference on Decision and Control, 3065-3070, 2008.

[Rivasplata '12] Omar Rivasplata, *Subgaussian Random Variables: An Expository Note.* Unpublished notes. Available at `http://www.stat.cmu.edu/~arinaldo/36788/subgaussians.pdf`.

[Rohde & Tsybakov '11] Angelika Rohde & Alexandre Tsybakov, *Estimation of High-Dimensional Low-Rank Matrices*, The Annals of Statistics, Volume 39, Number 2, 887-930, 2011.

[Romberg '08] Justin K. Romberg, *Imaging via Compressive Sampling.* IEEE Signal Processing Magazine, Volume 25, Issue 2, 14-20, 2008.

[Rudelson & Vershynin '08] Mark Rudelson & Roman Vershynin, *On Sparse Reconstruction from Fourier and Gaussian Measurements.* Communications in Pure and Applied Mathematics, Volume 61, Issue 8, 1025-1045, 2008.

[Rudelson & Vershynin '10] Mark Rudelson & Roman Vershynin, *Non-asymptotic Theory of Random Matrices: Extreme Singular Values*, Proceedings of the International Congress of Mathematicians, Hyderabad, India, 2010

[Rudelson '14] Mark Rudelson, *Recent Developments in Non-asymptotic Theory of Random Matrices* in Modern Aspects of Random Matrix Theory by Van Vu(ed.), 83-120, AMS, 2014. Also available at `http://arxiv.org/pdf/1301.2382v2.pdf`.

[Saab, Chartrand & Yilmaz '08] Rayan Saab, Rick Chartrand & Ozgür Yilmaz, *Stable Sparse Approximations via Nonconvex Optimization*, in IEEE International Conference on Acoustics, Speech, and Signal Processing, 2008.

[Sarwate '98] Dilip V. Sarwate, *Meeting the Welch Bound with Equality* in Sequences and Their Applications: Proceedings of SETA '98 by J.-P. Allouche et al. (eds.), 79-102, Springer, 1999.

[Schechtman '03] Gideon Schechtman, *Concentration, Results and Applications* in Handbook of The Geometry of Banach Spaces Volume 2 by Joram Lindenstrauss & William. B. Johnson, 1603-1635, North Holland, 2003.

[Schlotter & Wehmeier '13] Sven Schlotter & Kai F. Wehmeier, *Gingerbread Nuts and Pebbles: Frege and the Neo-Kantians - Two Recently Discovered Documents* British Journal for the History of Philosophy, Volume 21, Number 3, 591-609, 2013.

[Schnass & Vandergheynst '07] Karin Schnass & Pierre Vandergheynst, *Average Performance Analysis for Thresholding*, IEEE Signal Processing Letters, Volume 14, Issue 11, 828-831, 2007.

[Scott & Grassl '10] A. J. Scott & M. Grassl *Symmetric Informationally Complete Positive-Operator-Valued Measures: A New Computer Study*, Journal Mathematical Physics Vol. 51, 2010.

[Seneta '13] Eugene Seneta, *A Tricentenary History of the Law of Large Numbers*, Bernoulli, Volume 19, Number 4, 1088-1121, 2013.

[Shi, Han & Zheng '15] ZivQiang Shi, JivQing Han & Tie Ran Zheng, *Soft Margin Based Low-Rank Audio Signal Classification*, Neural Processing Letters, Volume 42, Issue 2, 291-299, 2015.

[Silverstein '85] Jack Silverstein, *The Smallest Eigenvalue of a Large-Dimensional Wishart Matrix*, Annals of Probability, Volume 13, 1364-1368, 1985

[Slepian & Pollack '61] D. Slepian & H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis and uncertainty-I*, Bell Syst. Tech. J, vol. 40, 43-64, 1961.

[So & Ye '07] Anthony Man-Cho So & Yinyu Ye *Theory of semidefinite programming for Sensor Network Localization*, Mathematical Programming, Volume 109, Issue 2, 367-384, 2007.

[Strassen '69] Volker Strassen, *Gaussian Elimination is not Optimal.* Numerical Mathematics, Volume 13, 354-356, 1969.

[Strohmer '12] Thomas Strohmer, *Measure What Should be Measured: Progress and Challenges in Compressive Sensing*, IEEE Signal Processing Letters, Volume 19, Issue 12, 887-893, 2012.

[Strohmer & Heath '03] Thomas Strohmer & R. W. Heath, *Grassmannian Frames With Applications to Coding and Communication*, Appl. Comput. Harmon. Anal. Vol. 14, 257?275, 2003.

[Su, Bogdan & Candès '15] Weijie Su, Malgorzata Bogdan, & Emmanuel J. Candès, *False Discoveries Occur Early on the Lasso Path*, To appear in Annals of Statistics. Available at `https://arxiv.org/abs/1511.01957`.

[Sustik et al. '07] M. A. Sustik, Joel A. Tropp, I. S. Dhillon, R. W. Heath, *On the Existence of Equiangular Tight Frames*, Linear Algebra and its Applications. Vol. 426, Issue 2, 619-635, 2007.

[Székely & Rizzo '07] Gábor J. Székely & Maria L. Rizzo, *The Uncertainty Principle of Game Theory*, The American Mathematical Monthly, Volume 114, Number 8, 688-702, 2007.

[Talagrand '94] Michel Talagrand, *The Supremum of Some Canonical Processes*, American Journal of Mathematics, Volume 116, Number 2, 283-325, 1994.

[Tibshirani '96] Robert Tibshirani, *Regression Shrinkage and Selection via the Lasso*, Journal of the Royal Statistical Society. Series B, VOlume 58, Issue 1, 267-288, 1996.

[Tikhomirov '66] Vladimir M. Tikhomirov, *A remark on L.A. Tumarkin's article "On diameters of infinite-dimensional compact sets"*. Vestnik Mosk. Univ., ser. mat. rnekh. astron., Number 3, Issue 73, 1966.

[Tikhomirov '83] Vladimir M. Tikhomirov, *Widths and entropy*, Uspekhi Mat. Nauk, Vol. 38, Issue 4, 91-99, 1983.

[Tikhomirov '03] Vladimir M. Tikhomirov, *Some thoughts on the mathematical work of Boris Sergeevich Kashin*. Proceedings of the Steklov Institute of Mathematics, Volume 280, Issue 1, 1-4, 2013.

[Tillmann & Pfetsch '14] Marc E. Pfetsch & Andreas Tillmann, *The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing*, IEEE Transactions on Information Theory, Volume 60, Issue 2, 1248-1259, 2014. DOI:10.1109/TIT.2013.2290112.

[Tropp '04] Joel A. Tropp, *Greed is Good: Algorithmic Results for Sparse Approximation*. IEEE Transactions on Information Theory, Vol. 50, Issue 10, 2231-2242, 2004.

[Tropp '05] Joel A. Tropp, *Average-Case Analysis of Greedy Pursuit*, Proceedings SPIE, Volume 5914, Wavelets XI, 2005.

[Tropp & Gilbert '07] Joel Tropp & Anna C. Gilbert, *Signal Recovery From Random Measurements via Orthogonal Matching Pursuit*. IEEE Transactions on Information Theory, Volume 53, Issue 12, 4655-4666, 2007.

[Tropp, Laska, Duarte, Romberg & Baraniuk '10] Joel Tropp, Jason Laska, Mario Duarte, Justin Romberg & Richard Baraniuk, *Beyond Nyquist: Efficient Sampling os Sparse Bandlimited Signals*. IEEE Transactions on Information Theory, Volume 56, Issue 1, 520-544, 2010.

[Tsirelson, Ibragimov & Sudakov '76] Boris Tsirelson, Ildar A. Ibragimov & Vladimir N. Sudakov, *Norm of Gaussian Sample Function*. In Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory, Volume 550 of Lecture Notes in Mathematics, pp. 20-41. Springer-Verlag, Berlin, 1976.

[Tweddle '84] Ian Tweddle, *Approximating n! Historical Origins and Error Analysis*, American Journal of Physics, Volume 52, 487-488, 1984.

[Unser '00] Michael Unser, *Sampling - 50 Years After Shannon*. Proceedings of the IEEE, Volume 88, Number 4, 569-587, 2000.

[Vandenberghe & Boyd '96] Lieven Vandenberghe & Stephen Boyd, *Semidefinite Programming*, SIAM Review, Volume 38, Number 1, 49-95, 1996.

[Vasanawala, Alley, Hargreaves, Barth, Pauly & Lustig '10] Shreyas S. Vasanawala, Marcus T. Alley, Brian A. Hargreaves, Richard A. Barth, John M. Pauly, & Michael Lustig, *Improved Pediatric MR Imaging with Compressed Sensing*. Radiology, Volume 256, Issue 2, 607-616, 2010.

[Vershynin '12] Roman Vershynin, *Introduction to the Non-asymptotic Analysis of Random Matrices*.Compressed Sensing, 210-268, Cambridge University Press, Cambridge, 2012.

[Vershynin '15] Roman Vershynin, *Estimation in High Dimensions: A Geometric Perspective*. Sampling Theory, a Renaissance, 3-66, Birkhauser, 2015.

[Vitale '00] Richard A. Vitale, *Some Comparisons for Gaussian Processes*, Proceedings of the American Mathematical Society, Volume 128, 3043-3046, 2000.

[Vybiral '11] Jan Vybiral, *A Variant of the Johnson-Lindenstrauss Lemma for Circulant Matrices*, Journal of Functional Analysis, Volume 260, Issue 4, 1096-1105, 2011.

[Vybiral '08] Jan Vybiral, *Widths of Embeddings in Function Spaces*. Journal of Complexity, Vol. 24, Issue 4, 545-570, 2008.

[Wainwright '15] Martin Wainwright, *Statistics Meets Optimization: Randomization and Approximation for High-Dimensional Problems* , Nachdiplom Lecture, ETH Zurich, 2015. Available at `http://www.stat.berkeley.edu/~wainwrig/nachdiplom/`.

[Waldron '09] Shayne Waldron, *On the Construction of Equiangular Frames From Graphs*, Linear Algebra and its Applications, Vol. 431, 2228-2242, 2009.

[Welch '74] L. R. Welch, *Lower Bounds on the Maximum Cross Correlation of Signals.* IEEE Trans. on Info. Theory, Vol. 20, Issue 3, 1974. 397-399. doi:10.1109/TIT.1974.1055219

[Wielandt '55] Helmut Wielandt, *An Extremum Property of Sums of Eigenvalues*, Proceedings of the American Mathematical Society, Volume 6, 106-110, 1955.

[Wright '04] Margaret H. Wright, *The Interior-Point Revolution in Optimization: History, Recent Developments, and Lasting Consequences*, Bulletin of the American Mathematical Society, Volume 42, Number 1, 39-56, 2004.

[Yin, Bai & Krishnaiah '88] Y. Q. Yin, Z. D. Bai & P. R. Krishnaiah, *On the Limit of the Largest Eigenvalue of the Large-Dimensional Sample Covariance Matrix*, Probability Theory and Related Fields, Volume 78, 509-521, 1988.

[Yu & Feng '14] Yi Yu & Yang Feng, *Modified Cross-Validation for Penalized High-Dimensional Linear Regression Models*, Journal of Computational and Graphical Statistics, Volume 23, Issue 4, 1009-1027, 2014.

[Zauner '99] Gerhard Zauner, *Quantendesigns - Grundz¨uge einer nichtkommutativen Designtheorie*, Ph.D. Thesis, Universitat Wien, 1999.

[Zhang '09] Tong Zhang, *Some Sharp Performance Bounds for Least Squares Regression with L_1 Regularization*, The Annals of Statistics, Volume 37, Number 5A, 2109-2144, 2009.

[Zhang '11] Tong Zhang, *Sparse Recovery with Orthogonal Matching Pursuit under RIP.* IEEE Transactions on Information Theory, Volume 57, Issue 9, 6215-6221, 2011.

[Zhou, Kong & Xiu '13] Shenglong Zhou, Lingchen Kong & Naihua Xiu, *New Bounds for RIC in Compressed Sensing*, Journal of the Operations Research Society of China, Volume 1, Issue 2, 227-237, 2013.

Blog Posts and Lectures:

[Tao's Blog - 22/03/2008] Terence Tao Blog's Entry on Dantzig Selector
https://terrytao.wordpress.com/2008/03/22/the-dantzig-selector-statistical-estimation-when-p-is-much-larger-than-n/

[Tao's Blog - 24/04/2008] Terence Tao Blog's Entry on Logarithmic Sobolev Inequality and Perelman's Entropy
https://terrytao.wordpress.com/2008/04/24/285a-lecture-8-ricci-flow-as-a-gradient-flow-log-sobolev-inequalities-and-perelman-entropy/

[Tao's Blog - 09/06/2009] Terence Tao Blog's Entry on Talagrand?s Concentration Inequality
https://terrytao.wordpress.com/2009/06/09/talagrands-concentration-inequality/

[Tao's Blog - 01/03/2010] Terence Tao Blog's Entry on Concentration of Measure
https://terrytao.wordpress.com/2010/01/03/254a-notes-1-concentration-of-measure/

[Tao's Blog - 06/25/2010] Terence Tao Blog's Entry on Uncertainty Principle
https://terrytao.wordpress.com/2010/06/25/the-uncertainty-principle/.

[Tao's Blog - 09/14/2010] Terence Tao Blog's Entry on Universality
https://terrytao.wordpress.com/2010/09/14/a-second-draft-of-a-non-technical-article-on-universality/

[Tao's Blog - 07/02/2010] Terence Tao Blog's Entry on Deterministic RIP Matrices
https://terrytao.wordpress.com/2007/07/02/open-question-deterministic-uup-matrices/.

[Tao's Blog - 01/12/2010] Terence Tao Blog's Entry on Eigenvalues and Sums of Hermitian matrices
https://terrytao.wordpress.com/tag/wielandt-hoffman-inequality/

[Mixon's Blog - 12/02/2013] Dustin Mixon Blog's Entry on Breaking the Quadratic Bottleneck I
https://dustingmixon.wordpress.com/2013/12/02/deterministic-rip-matrices-breaking-the-square-root-bottleneck/

[Mixon's Blog - 12/11/2013] Dustin Mixon Blog's Entry on Breaking the Quadratic Bottleneck II
https://dustingmixon.wordpress.com/2013/12/11/deterministic-rip-matrices-breaking-the-square-root-bottleneck-ii/.

[Mixon's Blog - 01/14/2014] Dustin Mixon Blog's Entry on Breaking the Quadratic Bottleneck III
https://dustingmixon.wordpress.com/2014/01/14/deterministic-rip-matrices-breaking-the-square-root-bottleneck-iii/

[Mixon's Blog - 02/08/2014] Dustin Mixon Blog's Entry on Gordon?s Escape Through a Mesh Theorem
https://dustingmixon.wordpress.com/2014/02/08/gordons-escape-through-a-mesh-theorem/

[Ellenberg - 02/22/10] Jordan Ellenberg Blog's Entry on Compressive Sensing
https://www.wired.com/2010/02/ff_algorithm/

[Bandeira's Blog - 12/04/2013] Afonso Bandeira Blog's Entry on Deterministic Matrices
https://afonsobandeira.wordpress.com/2013/12/04/deterministic-rip-matrices-a-new-cooperative-online-project/

[Wiki - Deterministic RIP Matrices] Math Research Wiki - Deterministic RIP Matrices
http://math-research.wikia.com/wiki/Deterministic_RIP_Matrices

Lectures:

[Ledoux - Lecture I] Michel Ledoux - Lecture on Logarithmic Sobolev Inequalities
http://www.math.univ-toulouse.fr/~ledoux/Logsobwpause.pdf.

[Ledoux - Lecture II] Michel Ledoux - Lecture on The Concentration of Measure Phenomenon
http://perso.math.univ-toulouse.fr/ledoux/files/2015/07/Villani2wp.pdf

Algorithms Websites:

[NESTA] by Jérôme Bobin, Stephen Becker & Emmanuel Candès
http://statweb.stanford.edu/~candes/nesta/.

[ℓ_1-MAGIC] by Emmanuel Candes & Justin Romberg
http://statweb.stanford.edu/~candes/l1magic/.

[YALL1] by Yin Zhang, Wei Deng, Junfeng Yang & Wotao Yin
http://yall1.blogs.rice.edu/.

[L1-LS] Kwangmoo Koh, Seung-Jean Kim & Stephen Boyd
https://stanford.edu/~boyd/l1_ls/.

[Fast ℓ_1] by Allen Yang, Arvind Ganesh, Zihan Zhou, Andrew Wagner, Victor Shia, Shankar Sastry & Yi Ma
https://people.eecs.berkeley.edu/~yang/software/l1benchmark/.

[CVXPY] by Martin Andersen, Joachim Dahl & Lieven Vandenberghe
http://cvxopt.org/examples/mlbook/l1.html.

[GPSR] by Mário Figueiredo, Robert D. Nowak & Stephen J. Wright
http://www.lx.it.pt/~mtf/GPSR/.

"After writing a story I was always empty and both sad and happy."

Ernest Hemingway in
A Moveable Fast, p. 6