Informative Sensing: Theory and Applications

by

Hyun Sung Chang

B.S., Seoul National University (1997)
M.S., Seoul National University (1999)

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2012

Author .....
Department of Electrical Engineering and Computer Science
May 23, 2012

Certified by .....
William T. Freeman
Professor of Electrical Engineering and Computer Science
Thesis Supervisor

Accepted by .....
Leslie A. Kolodziejski
Chair, Department Committee on Graduate Students

Informative Sensing: Theory and Applications

by

Hyun Sung Chang

Submitted to the Department of Electrical Engineering and Computer Science
on May 23, 2012, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

Compressed sensing is a recent theory for the sampling and reconstruction of sparse signals. Sparse signals only occupy a tiny fraction of the entire signal space and thus have a small amount of information, relative to their dimension. The theory tells us that the information can be captured faithfully with few *random* measurement samples, even far below the Nyquist rate.

Despite the successful story, we question how the theory would change if we had a more precise prior than the simple sparsity model. Hence, we consider the settings where the prior is encoded as a probability density. In a Bayesian perspective, we see the signal recovery as an inference, in which we estimate the unmeasured dimensions of the signal given the incomplete measurements. We claim that good sensors should somehow be designed to minimize the uncertainty of the inference. In this thesis, we primarily use Shannon's entropy to measure the uncertainty and in effect pursue the InfoMax principle, rather than the restricted isometry property, in optimizing the sensors.

By approximate analysis on sparse signals, we found random projections, typical in the compressed sensing literature, to be InfoMax optimal if the sparse coefficients are independent and identically distributed (i.i.d.). If not, however, we could find a different set of projections which, in signal reconstruction, consistently outperformed random or other types of measurements. In particular, if the coefficients are groupwise i.i.d., groupwise random projections with nonuniform sampling rate per group prove asymptotically InfoMax optimal. Such a groupwise i.i.d. pattern roughly appears in natural images when the wavelet basis is partitioned into groups according to the scale. Consequently, we applied the groupwise random projections to the sensing of natural images. We also considered designing an optimal color filter array for single-chip cameras. In this case, the feasible set of projections is highly restricted because multiplexing across pixels is not allowed. Nevertheless, our principle still applies. By minimizing the uncertainty of the unmeasured colors given the measured ones, we could find new color filter arrays which showed better demosaicking performance in comparison with Bayer or other existing color filter arrays.

Thesis Supervisor: William T. Freeman
Title: Professor of Electrical Engineering and Computer Science

Acknowledgments

Foremost, I wish to express my deep gratitude to my advisor, Prof. William Freeman, for his support of my study and for his patience, motivation, wisdom, and immense knowledge. His guidance helped me in every step of research and writing of this thesis.

I am greatly indebted to Prof. Yair Weiss, on my thesis committee, for his crucial advice on this thesis. It was my luck and pleasure to work with him.

I am grateful to Prof. Vivek Goyal, also on my thesis committee, for his insightful questions and comments. I also learned a lot about compressed sensing in his 6.342 class.

I thank all my labmates and friends at MIT for helpful discussions as well as for the fun we had together. I also thank my Korean friends who cheered me up abroad.

I thank Royal Dutch/Shell Group, especially Dr. Sandhya Devi and Dr. Jonathan Kane, for their financial support to the research projects related to this thesis and for their warm invitation to Houston.

I also thank Kwanjeong Educational Foundation for generously supporting me during my initial years at MIT.

I wish to thank my parents for their sincere support and inspiration throughout my life. I owe everything to them. To them I dedicate this thesis.

Special thanks go to my brother and his family, my parents-in-law, sisters-in-law, and a brother-in-law, for their constant care and encouragement.

Last but not least, I wish to thank my wife for standing beside me. Without her love and dedication, it would not have been possible for me to finish this work. Thank you, Hyein.

Contents

1 Introduction **9**

1.1 Compressed Sensing 10

1.2 Learning Compressed Sensing 17

1.3 InfoMax Principle in Sensory Systems 22

1.4 Organization of the Thesis 26

1.5 Appendix to Chapter 1 27

1.5.1 Derivation of Equations (1.1) and (1.2) 27

1.5.2 Proofs 27

1.5.3 Miscellaneous Lemmas 32

2 Informative Sensing **41**

2.1 Introduction 42

2.2 Information Maximization 47

2.2.1 Properness for Compressed Sensing 48

2.2.2 Toy Example 50

2.3 Analysis 52

2.3.1 I.I.D. case 53

2.3.2 Non-I.I.D. case 57

2.4 Numerical Experiments 63

2.5 Discussion 68

2.6 Appendix to Chapter 2 69

2.6.1 Proofs 69

2.6.2 Miscellaneous Lemmas 81

2.6.3 MMSE Estimation Based on Posterior Sampling 85

3 Informative Sensing for Natural Images **87**

3.1 Introduction 87

3.2 Informative Sensing 89

3.3 Informative Sensing for Natural Images 90

3.3.1 Mathematical Review 91

3.3.2 Optimal Profile of Measurement Density 92

3.3.3 Implementation and Examples 93

3.4 Noisy Measurements 95

3.4.1 Noise Effect 98

3.4.2 Optimization 99

3.5 Experimental Results 102

3.6 Discussion 108

3.7 Appendix to Chapter 3 109

3.7.1 Proofs 109

3.7.2 Miscellaneous Lemmas 112

4 Learning Color Filter Arrays **115**

4.1 Introduction 115

4.2 Color Image Sensing: Mathematical Review 118

4.3 Image Prior 118

4.4 CFA Design 119

4.4.1 Learning Bayes Optimal CFA 121

4.4.2 Results 125

4.5 Color Demosaicking 126

4.6 Experimental Results 129

4.7 Discussion 132

5 Conclusions **139**

A Generalized Gaussian Distribution **143**

Chapter 1

Introduction

Consider a classic question on sampling and reconstruction of signals. We are given m samples $\boldsymbol{y} = (y_1, \ldots, y_m)$, each of which is a linear filter output in response to the signal $\boldsymbol{x} = (x_1, \ldots, x_n)$, so that $y_i = \sum_{j=1}^{n} W_{ij} x_j$ for $i = 1, \ldots, m$. Suppose that $m/n = \beta < 1$. Can we reconstruct the input signal $\boldsymbol{x}$ from the incomplete output samples $\boldsymbol{y}$?

One case in which the answer is *yes* is when the signal $\boldsymbol{x}$ is bandlimited, for example in 1D Fourier domain, with a smaller bandwidth than $\pi\beta$. Precisely speaking, we require that $n, m \to \infty$, while $m/n = \beta$, for the bandlimitedness.[1] If this is the case, W_{ij} can be

$$W_{ij} = \mathrm{sinc}\left(\frac{i - \beta j}{\beta}\right) \tag{1.1}$$

where the sinc function is defined by $\mathrm{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$. The reconstruction can simply be obtained with a similar linear process, $x_j = \sum_{i=1}^{m} W'_{ij} y_i$ for $j = 1, \ldots, n$, where

$$W'_{ij} = \mathrm{sinc}(i - \beta j). \tag{1.2}$$

The underlying principle here is the so-called Nyquist sampling theorem [111] (see Section 1.5.1 for a simple derivation of Equations 1.1 and 1.2).

What can we say if the signal $\boldsymbol{x}$ is not bandlimited? This does not eliminate the possibility of exact recovery; bandlimitedness is a sufficient condition, but not a necessary condi-

[1] According to the celebrated uncertainty principle, a signal cannot be bounded both in time domain and in Fourier domain (see [111]).

tion, for exact reconstruction. In fact, the answer may depend on other statistical properties of the signal $\boldsymbol{x}$. Two or more signals can occupy the same bandwidth, while having different amounts of information content. Recent studies have argued that the fundamental limit on the sampling rate is determined by the information content (e.g., innovation rate [142], sparsity [51, 32], information dimension [149]), rather than the bandwidth, of the signal. In particular, compressed sensing is a recent theory for the sampling and reconstruction of "sparse" signals at the rate of the sparsity. While the Nyquist sampling theorem defines a minimum number of the samples required to perfectly reproduce an "arbitrary" bandlimited signal, we can further reduce the number if the input signals are known to be sparse in a certain basis.

We start this chapter by briefly reviewing the theory of compressed sensing. For self-containedness, this chapter includes all relevant proofs and derivations of the results, but they are not our own contributions. We simply summarize the literature. We recommend readers to see [47] for a detailed and well-organized review.

1.1 Compressed Sensing

The classical theory of compressed sensing deals with sparse signals, whether exactly k-sparse or k-term approximable. We mean, by k-sparse signals, the signals which have at most k nonzero elements in a certain orthonormal basis and, by k-term approximable signals, the signals for which there exists a k-sparse approximation with little error. To motivate the sparsity model, we may consider natural images (e.g., see Figure 1.1) which are sparse in a wavelet basis, although not in the standard pixel basis. We can obtain a good approximation by keeping only k significant wavelet coefficients while truncating the others. This is related to compressibility of the signals. Let $(x_1, \ldots, x_n)$ be the wavelet coefficients, ordered so that $|x_1| \geq |x_2| \geq \cdots \geq |x_n|$. Figure 1.2 displays such $|x_i|$'s, of Lena, versus i/n, which is referred to as *order statistics*. We observe that $|x_i|$'s decay quite rapidly. If $|x_i| \leq Ci^{-s}$ for some $C > 0$ and $s > 1$, the signal $\boldsymbol{x}$ is said to be *compressible* (see [30]). We will denote the ℓ_p-error of the best k-term approximation by $\sigma_k(\boldsymbol{x})_p$, i.e., $\sigma_k(\boldsymbol{x})_p \triangleq \min_{\widehat{\boldsymbol{x}} \in \mathcal{S}_k} \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_p$, where $\mathcal{S}_k$ represents the set consisting of all k-sparse vectors

Figure 1.1: Lena images. Left: the original, Right: a *sparse* approximation. For both, two types of representations are given. Top: pixel-domain, Bottom: wavelet-domain. On the bottom, large magnitudes are represented by warm colors, while small magnitudes are represented by cold colors. The wavelet coefficients of the original Lena show sparsity: most are near to zero. When only the largest 7% of the wavelet coefficients are kept, the resulting image (right) is very faithful to the original (left) whether in wavelet-domain or in pixel-domain.

in $\mathbb{R}^n$. The best k-term approximation is simply obtained by setting x_i's to zero for all $i > k$. If $\boldsymbol{x}$ is compressible, $\sigma_k(\boldsymbol{x})_p$ is bounded by [49]

$$\sigma_k(\boldsymbol{x})_p = \left(\sum_{i=k+1}^{n} |x_i|^p \right)^{1/p} \leq C \left(\sum_{i=k+1}^{\infty} i^{-ps} \right)^{1/p}$$
$$\leq C \left(\int_k^{\infty} z^{-ps} dz \right)^{1/p} = \frac{C}{(ps-1)^{1/p}} k^{1/p-s}. \quad (1.3)$$

The most basic result of compressed sensing is that, with probability 1, any single $m \times n$ "random" matrix $\boldsymbol{W}$ makes all $\boldsymbol{x} \in \mathcal{S}_k$ exactly recoverable from $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ if

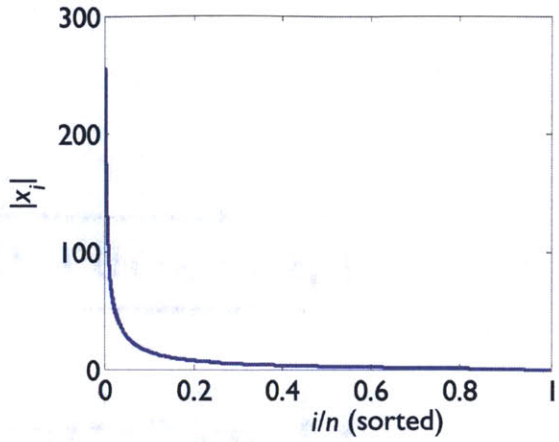


Figure 1.2: Order statistics of the wavelet coefficients of Lena.

$m \geq 2k$. The proof is based on the fact that every k-sparse signal $\boldsymbol{x}$ will have the unique projection $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ if $\boldsymbol{W}$ is random and has sufficiently many ($\geq 2k$) rows (see Proposition 1.11 in Section 1.5.3 for details). The recovery involves nonlinear operations and generally requires extremely high computational cost. However, if $m \geq O(k \log n)$, the exact reconstruction is possible with random matrices by a simple convex optimization (e.g., ℓ_1-minimization) or by a greedy optimization (e.g., orthogonal matching pursuit). We postpone the proof[2] to Corollary 1.5 and here will give a brief intuition behind the ℓ_1-minimization . Given a measurement $\boldsymbol{y}$, what we actually want to solve is

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z}\|_0 \quad \text{subject to } \boldsymbol{z} \in \mathcal{B}(\boldsymbol{y}) \tag{1.4}$$

where $\|\boldsymbol{z}\|_0$ denotes the number of nonzero elements in $\boldsymbol{z}$ (commonly called ℓ_0-*pseudonorm* [52]) and where $\mathcal{B}(\boldsymbol{y})$ represents the set of $\boldsymbol{z}$'s which are consistent with the measurement $\boldsymbol{y}$. At this time, $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{z} : \boldsymbol{W}\boldsymbol{z} = \boldsymbol{y}\}$. Problem (1.4) is NP-hard. A typical relaxation to this original problem is to employ ℓ_1-norm, which has convexity, in place of ℓ_0-pseudonorm. Specifically,

$$\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z}\|_1 \quad \text{subject to } \boldsymbol{z} \in \mathcal{B}(\boldsymbol{y}). \tag{1.5}$$

[2]Our proof will assume recovery based on ℓ_1-minimization. Refer to [153] for a proof with orthogonal matching pursuit.

Problem (1.5) is computationally feasible if $\mathcal{B}(\boldsymbol{y})$ is convex. In our case, it simply becomes linear programming [31]. To see why ℓ_1-minimization promotes sparsity during recovery, refer to Figure 1.3, where the solution to the ℓ_1-minimization problem exactly coincides with the solution to the ℓ_p-minimization problem for any $p < 1$ and notably is sparse.[3]

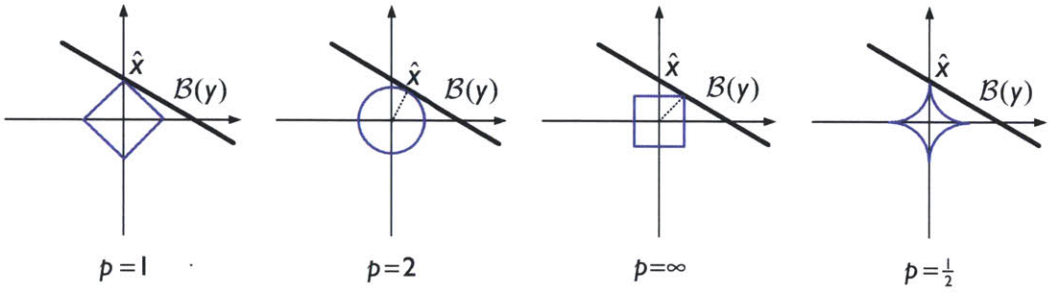


Figure 1.3: The solution $\widehat{\boldsymbol{x}}$ in $\mathbb{R}^2$ to the problem $\widehat{\boldsymbol{x}} = \arg\min_{\boldsymbol{z}} \|\boldsymbol{z}\|_p$ subject to $\boldsymbol{z} \in \mathcal{B}(\boldsymbol{y})$, where $p = 1, 2, \infty$ and $1/2$. Note that the interior of $\|\boldsymbol{z}\|_p = c$ is convex if and only if $p \geq 1$. Also note that ℓ_1-minimization finds a sparse solution as with ℓ_p-minimization for any $p < 1$.

Then, what makes random matrices so magical? The theory of compressed sensing explains the magic through the restricted isometry property (RIP). In linear algebra, the RIP characterizes matrices which are nearly orthonormal, at least when operating on sparse vectors. It keeps any two sparse vectors almost as distant in the rowspace (or measurement subspace) as in the original space. Formally, it is defined as below:

Definition 1.1. A matrix $\boldsymbol{A}$ satisfies the *restricted isometry property* (RIP) of order k if there exists a $\delta_k \in (0, 1)$ such that

$$(1 - \delta_k)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{A}\boldsymbol{x}\|_2^2 \leq (1 + \delta_k)\|\boldsymbol{x}\|_2^2 \tag{1.6}$$

holds for any k-sparse vector $\boldsymbol{x}$.

For any two distinct $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{S}_k$, let $\boldsymbol{e} = \boldsymbol{x} - \boldsymbol{x}'$. Evidently, $\boldsymbol{e} \in \mathcal{S}_{2k}$ and $\|\boldsymbol{e}\|_2 > 0$. If a matrix $\boldsymbol{W}$ satisfies the RIP of order $2k$, then $\|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{W}\boldsymbol{x}'\|_2 = \|\boldsymbol{W}\boldsymbol{e}\|_2 \geq$

[3] In 1990's, before the birth of the theory of compressed sensing, the use of ℓ_1-minimization already received attention from the signal processing community (e.g., for sparse representation of signals [37] and edge-preserving image processing [87]) and also within the statistics literature (as a method for variable selection in regression, known as Lasso [134]).

$\sqrt{1-\delta_{2k}}\|\boldsymbol{e}\|_2 > 0$, which implies that the projections $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ and $\boldsymbol{y}' = \boldsymbol{W}\boldsymbol{x}'$ are also distinct, so the exact recovery is possible. Here, any positive δ_{2k} works as long as it is strictly smaller than one. In the real-world, however, few signals are exactly k-sparse (recall Figure 1.1). Gracefully, the theory is extended to compressible signals as well but requires, for robustness, that δ_{2k} be sufficiently small (typically, $\delta_{2k} < \sqrt{2}-1 \approx 0.41$). To lower δ_{2k}, we need to increase m. It is known that any $m \times n$ matrix that satisfies the RIP of order k with constant $\delta_k < 0.5$ should have $m \geq Ck\log(n/k)$ for some $C > 0$ (see [47] or [46]).

Here are a couple of main theorems on the performance bound of the recovery by ℓ_1-minimization when the measurement matrix has the RIP, each for the noiseless and noisy setting:

Lemma 1.2. Suppose that $\boldsymbol{W}$ satisfies the RIP of $2k$ with $\delta_{2k} < \sqrt{2}-1$ and that we obtain the measurement $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$. Then, the solution $\widehat{\boldsymbol{x}}$ to Problem (1.5) obeys

$$\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 \leq C_0 \frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} \tag{1.7}$$

where

$$C_0 = 2\frac{1-(1-\sqrt{2})\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}}. \tag{1.8}$$

Proof. This can be regarded as a special case (with $\epsilon = 0$) of Lemma 1.3, for which the proof will be given in Section 1.5.2.1. □

Lemma 1.3. Suppose that $\boldsymbol{W}$ satisfies the RIP of $2k$ with $\delta_{2k} < \sqrt{2}-1$ and that we obtain the "noisy" measurement $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta}$, where $\|\boldsymbol{\eta}\|_2 \leq \epsilon$. Then, the solution $\widehat{\boldsymbol{x}}$ to Problem (1.5) obeys

$$\|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|_2 \leq C_0 \frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + C_1\epsilon \tag{1.9}$$

where

$$C_0 = 2\frac{1-(1-\sqrt{2})\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}}, \quad C_1 = 4\frac{\sqrt{1+\delta_{2k}}}{1-(1+\sqrt{2})\delta_{2k}}. \tag{1.10}$$

Proof. See Section 1.5.2.1. □

In the above lemmas, the compressibility of signals makes sure that the bound decreases with k. Perhaps the most surprising result in compressed sensing is that it typically uses totally random projections, while some deterministic matrices satisfying the RIP are also known [50, 80, 81, 25, 151, 62]. It is based on the following lemma:

Lemma 1.4. Let $\boldsymbol{A}$ be an $m \times n$ matrix whose entries are i.i.d. samples from $\mathcal{N}(0, \frac{1}{m})$. For any $\delta > 0$, if $m \geq O\left(\frac{k}{\delta^2}\log(n/k)\right)$, the matrix $\boldsymbol{A}$ satisfies, with probability at least $1 - 2e^{-\delta^2 m/72}$, the RIP of order k with $\delta_k < \delta$.

Proof. See Section 1.5.2.2. □

Although Lemma 1.4 concerns only Gaussian random matrices, similar claims can also be made to Bernoulli, or more generally to any sub-Gaussian random matrices (see [141]). Recall that any $m \times n$ matrix satisfying the RIP of order k with constant $\delta_k < 0.5$ should have $m \geq O(k\log(n/k))$, which implies that random matrices satisfy the RIP with the minimal number of rows (i.e., measurements) up to a constant factor, if $\delta < 0.5$.

Corollary 1.5. Suppose an $m \times n$ random matrix $\boldsymbol{W}$ with $m \geq O(\frac{k}{\delta^2}\log(n/k))$, where $\delta = \sqrt{2}-1$. If $\boldsymbol{x} \in \mathcal{S}_k$, ℓ_1-minimization exactly recovers $\boldsymbol{x}$ from $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ with probability at least $1 - 2e^{-\delta^2 m/72}$. This probability goes to one as n goes to infinity.

Proof. This is a consequence of Lemma 1.4 and Lemma 1.2. According to Lemma 1.4, the matrix $\boldsymbol{W}$ will have the RIP of $2k$ with $\delta_{2k} < \sqrt{2}-1$ with probability at least $1-2e^{-\delta^2 m/72}$, and Lemma 1.2 ensures that $\widehat{\boldsymbol{x}} = \boldsymbol{x}$ because $\sigma_k(\boldsymbol{x})_1 = 0$ for any $\boldsymbol{x} \in \mathcal{S}_k$. □

While the RIP provides guarantees for exact recovery of k-sparse signals, testing whether a given matrix satisfies the RIP has a combinatorial complexity, since one must essentially consider $\binom{n}{2k}$ submatrices. In many cases it is preferable to use properties of $\boldsymbol{W}$ that are

more easily computable to provide similar recovery guarantees. The coherence of a matrix is one such property.

Definition 1.6. The coherence of a matrix $\boldsymbol{A}$, $\mu(\boldsymbol{A})$, is the largest absolute inner product between any two normalized columns $\boldsymbol{a}_i$, $\boldsymbol{a}_j$ of $\boldsymbol{A}$, i.e.,

$$\mu(\boldsymbol{A}) \triangleq \max_{i \neq j} \frac{|\boldsymbol{a}_i^T \boldsymbol{a}_j|}{\|\boldsymbol{a}_i\|_2 \cdot \|\boldsymbol{a}_j\|_2}. \tag{1.11}$$

Suppose that $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ for $\boldsymbol{x} \in \mathcal{S}_k$. If $\mu(\boldsymbol{W}) < \frac{1}{2k-1}$, the following fundamental results are guaranteed to hold [52, 66, 135]:

1. The vector $\boldsymbol{x}$ is the unique solution to Problem (1.4) with $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{z} : \boldsymbol{y} = \boldsymbol{W}\boldsymbol{z}\}$;
2. The vector $\boldsymbol{x}$ is the unique solution to Problem (1.5) with $\mathcal{B}(\boldsymbol{y}) = \{\boldsymbol{z} : \boldsymbol{y} = \boldsymbol{W}\boldsymbol{z}\}$;
3. Orthogonal matching pursuit, a greedy algorithm that finds nonzero elements in $\widehat{\boldsymbol{x}}$ one at a time so as to minimize the residual error $\|\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{x}}\|_2$, finally yields $\widehat{\boldsymbol{x}} = \boldsymbol{x}$.

The coherence can also be related to the RIP (e.g, see Lemma 1.7 below, [47]) and sometimes is used to form the RIP*less* theory [27] or as an optimization criterion [57] for compressed sensing.

Lemma 1.7. If $\boldsymbol{A}$ has unit-norm columns and coherence $\mu = \mu(\boldsymbol{A})$, then $\boldsymbol{A}$ satisfies the RIP of order k with $\delta_k = (k-1)\mu$ for all $k < 1 + 1/\mu$.

Proof. See Section 1.5.2.3. □

Remark: Compressed sensing provides a nice framework for sampling and reconstruction of sparse signals, at the rate of their information content encoded by the sparsity. The rationale behind its success is a better modeling of signals. If sparsity better describes the input signal than bandlimitedness in Fourier domain, random measurements can accomplish a sub-Nyquist rate. We doubt, however, that sparsity model is best in describing the signals we wish to measure (e.g., natural images shown in Figure 1.1). What if we had more precise information? The probability density function (pdf) will provide the richest prior information for random signals if somehow known. Although we may not exactly

know the true pdf, we may still be able to find the maximum entropy distribution subject to some testable information, which best summarizes the current state of knowledge [82].

While sharing the same level of sparsity (or order statistics), two or more signals can have different pdfs (see Section 2.1). Can we optimize linear measurements for the respective signals? This question has motivated the present study.

1.2 Learning Compressed Sensing

People have begun to consider learning optimal linear projections from a training set typical of the signals [146, 34, 126]. Let $\{\boldsymbol{x}_i\}_{i=1}^N$ denote the training set which consists of N example signals. We assume that N is large enough to capture the signal distribution sufficiently well. Given a projection $\boldsymbol{y}_i$ of $\boldsymbol{x}_i$ (i.e., $\boldsymbol{y}_i = \boldsymbol{W}\boldsymbol{x}_i + \boldsymbol{\eta}$ where $\boldsymbol{\eta}$ denotes the measurement noise), a natural attempt is to maximize the probability of correct recovery of the original signal $\boldsymbol{x}_i$. This may be formally written as

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}:\boldsymbol{W}\boldsymbol{W}^T=\boldsymbol{I}}{\arg\max}\ \mathbb{E}_{\boldsymbol{\eta}}\left[\sum_i \log \Pr(\boldsymbol{x}_i|\boldsymbol{y}_i;\boldsymbol{W})\right], \tag{1.12}$$

where $\mathbb{E}_{\boldsymbol{\eta}}[\cdot]$ denotes the expectation with respect to the measurement noise $\boldsymbol{\eta}$. For simplicity, the measurement noise is assumed to have a Gaussian density, i.e., $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Weiss et al. [146] named the row vectors of $\boldsymbol{W}^*$ *uncertain components* of data and the procedure of finding $\boldsymbol{W}^*$ *uncertain component analysis* (UCA).[4]

Since the sum of the posterior over all datapoints in the training set is normalized to 1, the value of $\Pr(\boldsymbol{x}_i|\boldsymbol{y}_i;\boldsymbol{W})$ depends on how many datapoints are likely to produce $\boldsymbol{y}_i$ as their noisy projection. From the generative model, the likelihood is given by

$$p(\boldsymbol{y}_i|\boldsymbol{x}_j) = \frac{1}{(2\pi\sigma^2)^{m/2}} e^{-\|\boldsymbol{y}_i - \boldsymbol{W}\boldsymbol{x}_j\|^2/2\sigma^2}. \tag{1.13}$$

Therefore, an optimal matrix $\boldsymbol{W}^*$ must be chosen to maximally separate noisy projections

[4]The name has originated from the fact that optimal projections should capture what is still uncertain about the signals, given the training set. In a later part of this section, we will see that the criterion in (1.12) is essentially equivalent to maximizing the entropy (or uncertainty) of $\boldsymbol{W}\boldsymbol{x}$ if $\sigma \to 0$, which may provide a clear sense of the name.

of the datapoints. Further analytic characterization of $\boldsymbol{W}^*$ is given in the following lemma [146]:

Lemma 1.8. Let $\boldsymbol{W}^*$ be the UCA matrix. Then, $\boldsymbol{W}^*$ satisfies the following fixed-point equations, relating the data assignment probabilities q_{ij} and the projection $\boldsymbol{W}$.

$$q_{ij} = \Pr(\boldsymbol{x}_j|\boldsymbol{y}_i; \boldsymbol{W}) \tag{1.14}$$

$$\boldsymbol{W}^* = \text{top } m \text{ eigenvectors of} \sum_{i,j} q_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T. \tag{1.15}$$

Proof. See Section 1.5.2.4. □

We may be able to find $\boldsymbol{W}^*$ analytically in some special cases such as below (the proofs are given in Section 1.5.2.5–1.5.2.7):

Example 1.1. As $\sigma \to \infty$, UCA approaches principal component analysis (PCA).

Example 1.2. If the data $\{\boldsymbol{x}_i\}$ lie in an m-dimensional subspace, then the UCA vectors and the top m PCA vectors span the same subspace.

Example 1.3. If the data $\{\boldsymbol{x}_i\}$ are k-sparse in any basis and if $m \geq 2k$, then for $\sigma \to 0$, a random matrix becomes one of the UCA matrices.

In many other cases, however, it may be difficult to find $\boldsymbol{W}^*$ analytically; then we may use numerical algorithms based on the gradient of the utility function with respect to $\boldsymbol{W}$:

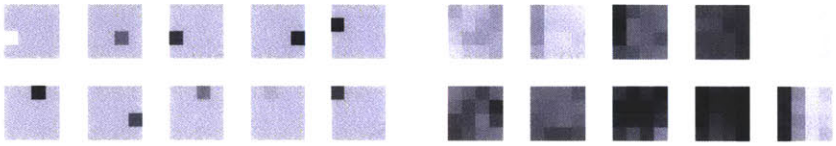
$$\nabla \boldsymbol{W} \propto \boldsymbol{W} \left(\sum_{ij} q_{ij}(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T \right) \tag{1.16}$$

with q_{ij} as given in Equation (1.14).

While the fixed-point equations show that under certain conditions, PCA and UCA give the same projections, they also highlight the difference. The PCA tries to maximize the variance of the projections, which can be thought of as maximizing the *average* distance between the projections of any two signals. The UCA maximizes a *weighted average* distance between the projections of any two signals, weighted by the probability of assignment for each observation. The weighted average gives high weight to pairs of signals

whose projections are similar (determined by the noise level σ). This makes sense in terms of robust reconstruction. For a given noise level σ, two signals whose projected distance is 10σ make little confusion in recovery and are nearly as good as two signals whose projected distance is 100σ.

To illustrate the behavior of the UCA, we refer to Figures 1.5 and 1.6, where experimental results are obtained with two types of signals shown in Figure 1.4:



1-sparse signal samples Natural image patch samples

Figure 1.4: Samples of two types of signals, each in a 4×4 patch.

1. 1-sparse signals in $\mathbb{R}^{16}$: Each 4×4 patch has one nonzero pixel. If the ith element is nonzero, the value is an integer uniformly distributed in the range $[-16, 16]$ plus ϵ_i, where ϵ_i's are small positive numbers decreasing with the index i simply to break symmetries.

2. 4×4 patches sampled from Berkeley dataset of natural images [101].

We first estimated uncertain components for 1-sparse signals for different noise values σ^2 and different numbers of projections m. Recall that for noiseless measurements, if $m \geq 2$, random projections are optimal for such signals because every datapoint has the unique projection with probability one (see Proposition 1.11 in Section 1.5.3). As expected by Example 1.3, when σ^2 is very small, any random matrix, of which the entries are randomly drawn, is a fixed-point of the iteration (1.16). But when σ^2 is large, UCA learns projections that are still incoherent to the sparse basis but nonrandom. To visualize the learned UCA projections, we plot in Figure 1.5 the projections of the sparse signals into two dimensions using random projections (top left) and the UCA projections (top right). Since all signals are 1-sparse in the high dimensional space, the signal set defines a discrete set of rays in

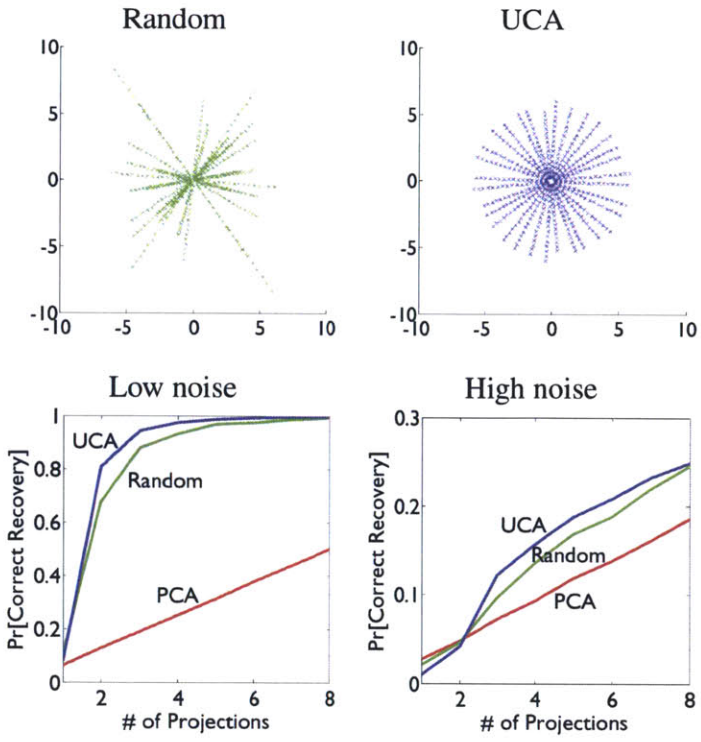


Figure 1.5: UCA results on the 1-sparse signals. Top: projection of the full dataset from sixteen dimensions onto two dimensions using random and UCA projections. Bottom: comparison of percentage of correct decodings as a function of the number of projections, for different noise levels.

high dimensions, all passing through the origin. In both the random projections and the UCA projections, one can still observe the projected rays, but UCA finds a projection in which these rays are (approximately) emanating at regular angles. Thus UCA is finding a projection in which the number of signals with similar projections is smaller than in a random projection. Figure 1.5 also compares the decoding performance of the different projections. As expected, UCA performs slightly better than random projections, and both UCA and random perform much better than PCA.

In the second experiment, we estimated uncertain components for a set of 1,000 4×4 image patches randomly sampled from natural images. Again, to visualize the UCA projection versus a random projection, we show projections of the image signals into two dimensions using random projections (Figure 1.6, top left) and the UCA projections (top right). Note that the random projections are dispersed significantly less than the UCA

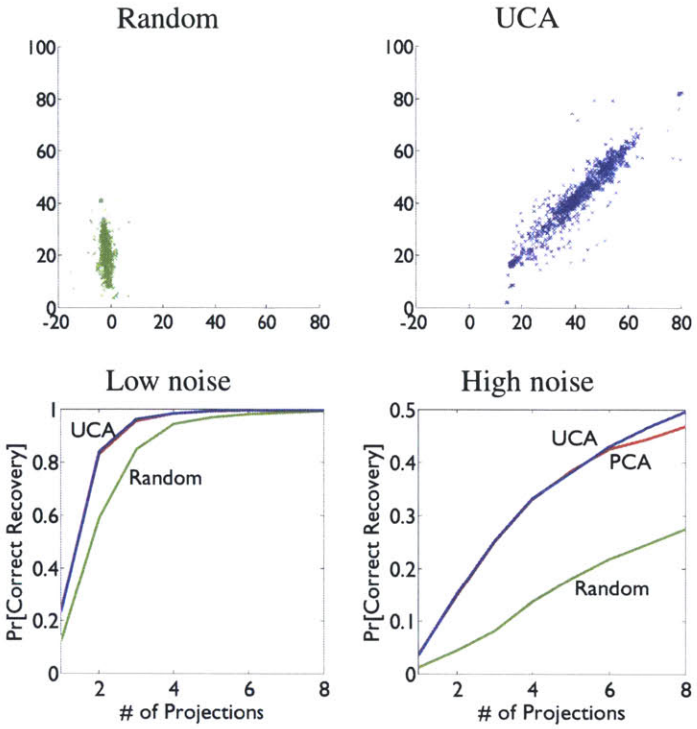


Figure 1.6: UCA results on natural image patches. Top: projection of the full dataset from sixteen dimensions onto two dimensions using random and UCA projections. Bottom: comparison of percentage of correct decodings as a function of the number of projections, for different noise levels. The UCA and PCA results are almost identical, in most ranges, so the red line is occluded by the blue line.

projections. For this dataset, we found that UCA learns projections that are nearly identical to PCA. Figure 1.6 compares the decoding performance of the different projections. In this case, UCA performs almost identically to PCA and much better than random projections.

Remark: UCA provides a principled optimization method for compressed sensing. It is closely related to the InfoMax principle. Briefly speaking, the UCA utility function converges to $-h(\boldsymbol{x}|\boldsymbol{y})$ and becomes maximum when the matrix $\boldsymbol{W}$ maximizes the mutual information $I(\boldsymbol{x};\boldsymbol{y})$ because $I(\boldsymbol{x};\boldsymbol{y}) = h(\boldsymbol{x}) - h(\boldsymbol{x}|\boldsymbol{y})$ where $h(\boldsymbol{x})$ is a constant independent of $\boldsymbol{W}$. We will review the InfoMax principle and related context in the next section. Minimizing data uncertainty given the projection has also been proposed in *sequential* design of compressed sensing [34, 127], in which the projections are chosen one by one so as to minimize the remaining data uncertainty given the outcome of the previous projections.

A major drawback of learning approaches is that they are not free from the curse of dimensionality. The signal dimension n is usually restricted to be small, especially if we learn a non-adaptive, *one-shot* measurement matrix $\boldsymbol{W}$. In case where we design each row of $\boldsymbol{W}$ sequentially, observing previous outcomes, learning can be computationally somewhat easier; some moderate dimension also becomes feasible (e.g., see [126]). But the sequentially designed matrix is adaptive to a specific input signal and thus cannot be reused for others in the same class. The measurement process may also take a long time, proportional to m.

1.3 InfoMax Principle in Sensory Systems

In the first NIPS meeting, Linsker [94] suggested the InfoMax principle for the design of a linear sensory system. According to this principle, the goal of the sensory system is to maximize the mutual information between the sensors and the world (see also [8, 12, 7]). For the input $\boldsymbol{x} \in \mathbb{R}^n$ and the output $\boldsymbol{y} \in \mathbb{R}^m$, let us write the linear sensory system as $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta}$, where $\boldsymbol{W}$ is an $m \times n$ measurement matrix and $\boldsymbol{\eta}$ denotes the sensor noise. The input-output mutual information is defined as

$$I(\boldsymbol{x}; \boldsymbol{y}) = h(\boldsymbol{y}) - h(\boldsymbol{y}|\boldsymbol{x}) \tag{1.17}$$

where $h(\boldsymbol{y})$ is the entropy of the output and where $h(\boldsymbol{y}|\boldsymbol{x})$ denotes the remaining entropy of the output given the input signal and thus merely the entropy of the sensor noise. Because the noise entropy does not depend on the measurement matrix, $h(\boldsymbol{y})$ may directly be used, instead of $I(\boldsymbol{x}; \boldsymbol{y})$, as the InfoMax criterion. Here, we focus on the noiseless case.

The entropy of the linear measurement $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ can be made arbitrarily large, simply by taking $\boldsymbol{W} = c\boldsymbol{W}_o$ with $c \to \infty$ for any fixed $\boldsymbol{W}_o$. To preclude such a trivial manipulation, we usually impose a restriction on the measurement matrix $\boldsymbol{W}$ to satisfy $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{I}$, as in Section 1.2, or to satisfy $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = \gamma$. The latter, called *total power budget constraint*, is a little more common. In the noiseless case, the maximum of $h(\boldsymbol{W}\boldsymbol{x})$, subject to the power budget constraint $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = \gamma$, can be achieved only by a tight frame

matrix satisfying $\boldsymbol{W}\boldsymbol{W}^T = (\gamma/m)\boldsymbol{I}$ (see Lemma 2.2 in Section 2.2). Here, γ/m is an uninteresting scale factor merely introduced to compensate for the case $\gamma \neq m$. Therefore, the linear InfoMax problem is formally stated as

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}_{m\times n}:\boldsymbol{W}\boldsymbol{W}^T=\boldsymbol{I}}{\arg\max} h(\boldsymbol{W}\boldsymbol{x}). \tag{1.18}$$

The solution depends on the prior probability of $\boldsymbol{x}$ and the number of projections m. In particular, in the complete case ($m = n$), $h(\boldsymbol{W}\boldsymbol{x}) = h(\boldsymbol{x})$ for any tight frame matrix, so the InfoMax problem is meaningless. We are therefore interested in the strictly undercomplete case ($m < n$).

The Gaussian case was analyzed by Linsker.

Lemma 1.9. (Linsker [94]) If the signal $\boldsymbol{x}$ is jointly Gaussian, then the solution to the linear InfoMax problem (Equation 1.18) is given by the m principal components of $\boldsymbol{x}$.

Proof. See Section 1.5.2.8. □

Since then, there has been tremendous amount of subsequent work in terms of finding algorithms to estimate the mutual information in a linear system as well as relationships between InfoMax and other learning criteria (e.g. [61, 90]). In 1995, Bell and Sejnowski [18] considered to apply a pointwise *nonlinearity* g after the linear projection so that $\boldsymbol{y} = g(\boldsymbol{W}\boldsymbol{x})$. The range of g is assumed to be $[0, 1]$ so no additional restrictions on $\boldsymbol{W}$ are needed. Formally, the nonlinear InfoMax problem is

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}_{m\times n}}{\arg\max}\, h(g(\boldsymbol{W}\boldsymbol{x})) \tag{1.19}$$

where $g(y_1, y_2, \ldots, y_m) = (g(y_1), g(y_2), \ldots, g(y_m))$ and the range of g is $[0, 1]$.

As in the linear InfoMax case, the solution to the nonlinear InfoMax problem depends on the distribution of the input $\boldsymbol{x}$. In the nonlinear case, it also depends on the form of the nonlinearity g. Bell and Sejnowski showed that given an appropriate g, nonlinear InfoMax performs independent component analysis (ICA).

Lemma 1.10. Suppose that $\boldsymbol{x}$ is distributed according to the ICA generative model: $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$ where $\boldsymbol{D}$ is an invertible square matrix and α_i's are i.i.d. Suppose also that g is equal to the cumulative density function (cdf) of α_i. Then the solution to the nonlinear InfoMax problem (Equation 1.19) is given by choosing m ICA filters, i.e. m rows of $\boldsymbol{D}^{-1}$.

Proof. See Section 1.5.2.9. □

Bell and Sejnowski's discussion of the relationship between ICA and InfoMax mostly focused on the complete case $n = m$, and usually ICA algorithms are not used as dimension reduction techniques. As Lemma 1.10 claims, however, the equivalence between ICA and nonlinear InfoMax actually holds for arbitrary m. At the same time, the lemma highlights the crucial dependence between g and the cdf of the sources.

Bell and Sejnowski [19] also showed that applying the InfoMax principle to whitened natural image patches gives Gabor-like filters, similar to those found in primary visual cortex (see Figure 1.7, left). If we apply their nonlinear InfoMax algorithm to exactly k-sparse signals (e.g., with $k = 3$), it finds the basis sparisfying the coefficients, as shown in Figure 1.7 (right).

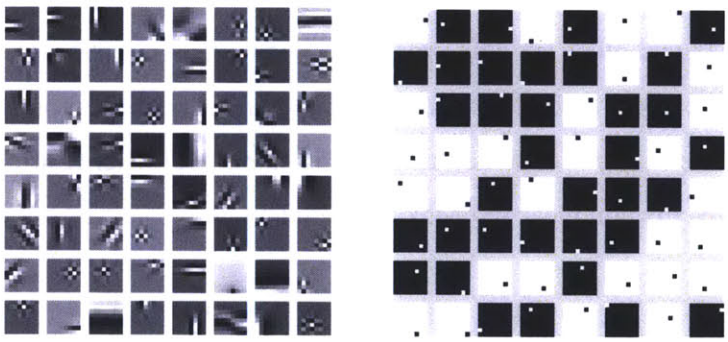


Figure 1.7: The nonlinear InfoMax filters for two types of signals. Left: whitened natural image patches. Right: exactly k-sparse signals with $k = 3$. To learn the filters, we used one million samples per case. For the nonlinearity g, the cdf of the generalized Gaussian with the shape parameter $r = 0.5$ has commonly been used (see Appendix A for the generalized Gaussian distribution).

Remark: As Table 1.1 summarizes, the InfoMax problem has a couple of well-known solutions in specific settings; however, the general question of InfoMax optimal projections

Table 1.1: A couple of well-known solutions to the InfoMax problem

Signal	Network	Solution
jointly Gaussian	linear $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$	PCA [94]
non-Gaussian source separation model	nonlinear $\boldsymbol{y} = g(\boldsymbol{W}\boldsymbol{x})^{\dagger}$	ICA [18]

$^{\dagger}g$ matches the cdf of each source. See Lemma 1.10 for the details.

remains unsolved.

In this thesis, we are particularly interested in the linear InfoMax problem with non-Gaussian signals. To highlight the difference from the previous settings, consider a white signal $\boldsymbol{x}$, with $Cov(\boldsymbol{x}) = \boldsymbol{I}$. The covariance of $\boldsymbol{y}$ remains as the identity for any tight frame matrix $\boldsymbol{W}$ satisfying $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{I}$. Given any fixed covariance, Gaussian has the maximum entropy [43], so all we need to solve the linear InfoMax problem (1.18) is to find a linear projection that makes the projection as Gaussian as possible. This is in stark contrast to a large amount of research on projection pursuit methods (e.g. [84, 107]) that seek projections that are as *non-Gaussian* as possible. Those non-Gaussianity seeking projections must be *least* informative about the white signal. Given the close connection between projection pursuit methods and ICA [79], this raises an intriguing possibility that ICA may find the least informative projections, while it was formulated to find the most informative ones in the setting as in Lemma 1.10. Another interesting observation is that random projections are likely to achieve the maximal Gaussianity because of the central limit theorem, which suggests a novel connection between random projections and compressed sensing, without the use of RIP.

In the undercomplete linear InfoMax setting, the solution will become essentially equivalent to UCA, which desirably finds the projection that minimizes the remaining data uncertainty given the measurement (see our remark in page 21; see also Section 2.2 for more rigorous arguments).

1.4 Organization of the Thesis

The remaining part of this thesis comprises three self-contained chapters, which form a single line of study, plus a conclusion. We will briefly summarize the main results of the three chapters.

Chapter 2: We apply the InfoMax principle to sparse signals $\boldsymbol{x} = (x_1, \ldots, x_n)$, where x_i's are assumed to be mutually independent. We show that the InfoMax principle provides a similar result as the theory of compressed sensing – that we should use the random projections – in case that x_i's are symmetric, i.e., following the same distribution. However, if x_i's are not symmetric, we observe that random projections can be substantially far from being InfoMax optimal. We develop a set of mathematical tools to (approximately) optimize the InfoMax criterion. The InfoMax-based projections consistently outperform random projections as well as the PCA projection in signal recovery.

In subsequent chapters, we consider several applications.

Chapter 3: We show that groupwise random projections are asymptotically InfoMax optimal for groupwise i.i.d. signals. As an application, we model natural images as such a class of signals and deal with how to implement the groupwise random projections with reference to well-known image statistics. We consider the measurement noise as well. We derive the optimal power distribution among sensors, which generalizes Linsker's result [94] – the optimal power distribution for Gaussian inputs – to natural images which are not Gaussian.

Chapter 4: We regard the way single-chip digital cameras handle color as a special case of compressed sensing. For cost reduction, single-chip cameras use a color filter array (CFA) over the sensors and measure one wavelength, instead of three, per pixel. We show how to learn a CFA (and thus a measurement matrix) in a way that it minimizes the uncertainty of the missing color spectra given the measured ones. Both Shannon's conditional entropy and minimum mean-squared error (MMSE) are considered for the uncertainty measure. When the conditional entropy is adopted, this exactly implements the InfoMax principle under CFA physical constraints. By experiments, we demonstrate that our learned CFAs can give significant improvements in performance over existing CFAs.

1.5 Appendix to Chapter 1

1.5.1 Derivation of Equations (1.1) and (1.2)

Given discrete samples $\{z_k\}$, a continuous-time signal $z(t)$ that satisfies $z(kT_z) = k_z$, for a sampling period T_z, with the smallest bandwidth is given by $z(t) = \sum_k z_k \operatorname{sinc}(t/T_z - k)$.

Under the bandlimited assumption, if we let $s(t)$ be such an underlying continuous-time signal, $s(t)$ can be represented as $s(t) = \sum_j x_j \operatorname{sinc}(t/T_x - j)$ or as $s(t) = \sum_i y_i \operatorname{sinc}(t/T_y - i)$. Therefore, we have

$$y_i = s(iT_y) = \sum_j x_j \operatorname{sinc}(iT_y/T_x - j) \tag{1.20}$$

$$x_j = s(jT_x) = \sum_i y_i \operatorname{sinc}(jT_x/T_y - i). \tag{1.21}$$

Note that the ratio of the sampling rates is given by $m/n = T_x/T_y = \beta$. Plugging this into (1.20) and (1.21) yields Equations (1.1) and (1.2), respectively.

1.5.2 Proofs

1.5.2.1 Proof of Lemma 1.3

Note that $\|\boldsymbol{x}\|_1 \geq \|\widehat{\boldsymbol{x}}\|_1$ because $\widehat{\boldsymbol{x}}$ is a minimizer of ℓ_1-norm, while $\boldsymbol{x}, \widehat{\boldsymbol{x}} \in \mathcal{B}(\boldsymbol{y})$. Therefore, we may use Lemma 1.12. Applying Cauchy-Schwarz inequality and the RIP of $\boldsymbol{W}$ in a sequence, we can bound the quantity $|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W}\boldsymbol{e}|$ by

$$|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W}\boldsymbol{e}| \leq \|\boldsymbol{W}\boldsymbol{e}_\Lambda\|_2 \|\boldsymbol{W}\boldsymbol{e}\|_2 \leq \sqrt{1+\delta_{2k}} \|\boldsymbol{e}_\Lambda\|_2 \|\boldsymbol{W}\boldsymbol{e}\|_2. \tag{1.22}$$

Because $\|\boldsymbol{W}\boldsymbol{e}\|_2$ is also bounded by

$$\|\boldsymbol{W}\boldsymbol{e}\|_2 = \|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{W}\widehat{\boldsymbol{x}}\|_2 \leq \|\boldsymbol{W}\boldsymbol{x} - \boldsymbol{y}\|_2 + \|\boldsymbol{y} - \boldsymbol{W}\widehat{\boldsymbol{x}}\|_2 \leq 2\epsilon, \tag{1.23}$$

we obtain $|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W}\boldsymbol{e}| \leq 2\epsilon\sqrt{1+\delta_{2k}} \|\boldsymbol{e}_\Lambda\|_2$. Then, Lemma 1.12 immediately gives us the inequality (1.9).

1.5.2.2 Proof of Lemma 1.4

Let Λ be the set of column indices arbitrarily chosen so that $|\Lambda| = k$ and construct $\boldsymbol{A}_\Lambda$ by keeping only the columns indexed by Λ. Then, $\boldsymbol{A}_\Lambda$ will be an $m \times k$ matrix with i.i.d. entries sampled from $\mathcal{N}(0, \frac{1}{m})$. Random matrix theory [141] tells us that, for any $t \geq 0$, with probability at least $1 - 2e^{-t^2/2}$,

$$|\sigma_i(\boldsymbol{A}_\Lambda) - 1| \leq \sqrt{k/m} + t/\sqrt{m}, \quad \forall i, \tag{1.24}$$

where $\sigma_i(\boldsymbol{A}_\Lambda)$ denotes the singular values of the matrix $\boldsymbol{A}_\Lambda$ (cf. Marčenko-Pastur density [139]). Using Lemma 1.13, we see that the above inequality implies that, for $\epsilon = \sqrt{k/m} + t/\sqrt{m}$,

$$\max_i |\lambda_i(\boldsymbol{A}_\Lambda^T \boldsymbol{A}_\Lambda) - 1| \leq 3\max(\epsilon, \epsilon^2) \tag{1.25}$$

where $\lambda_i(\boldsymbol{A}_\Lambda^T \boldsymbol{A}_\Lambda)$ denotes the eigenvalues of $\boldsymbol{A}_\Lambda^T \boldsymbol{A}_\Lambda$. We take a union bound over all Λ's. Since there are $\binom{n}{k} \leq (\frac{en}{k})^k$ ways,[5] in total, to choose Λ, we can say that

$$\max_{i,\Lambda} |\lambda_i(\boldsymbol{A}_\Lambda^T \boldsymbol{A}_\Lambda) - 1| \leq 3\max(\epsilon, \epsilon^2) \tag{1.26}$$

with probability at least $1 - \binom{n}{k} 2e^{-t^2/2} \geq 1 - 2e^{k\log(en/k) - t^2/2}$. Note that the inequality (1.26) implies the RIP of order k with $\delta_k = 3\max(\epsilon, \epsilon^2)$.

Letting $t = \sqrt{2k\log(en/k)} + \delta\sqrt{m}/6$, we can conclude with probability at least $1 - 2e^{-\delta^2 m/72}$ that the matrix $\boldsymbol{A}$ satisfies the RIP of order k with $\delta_k = 3\max(\epsilon, \epsilon^2)$. Finally, if we take

$$m \geq \frac{36}{\delta^2}\left(\sqrt{k} + \sqrt{2k\log(en/k)}\right)^2 = O\left(\frac{k}{\delta^2}\log(n/k)\right), \tag{1.27}$$

[5] In [42], the authors make use of the inequality $k! \geq (k/e)^k$, derived from Stirling's approximation $k! \approx \sqrt{2\pi k}(k/e)^k$, to obtain the upperbound $\binom{n}{k} = \frac{n(n-1)\cdots(n-k+1)}{k(k-1)\cdots 1} \leq \frac{n^k}{k!} \leq (\frac{en}{k})^k$.

ϵ becomes smaller than or equal to $\delta/3$ by the following:

$$\epsilon = \sqrt{k/m} + t/\sqrt{m} = (\sqrt{k} + \sqrt{2k\log(en/k)})/\sqrt{m} + \delta/6 \leq \delta/3. \tag{1.28}$$

This implies that $\delta_k \leq \delta$, completing the proof.

1.5.2.3 Proof of Lemma 1.7

Let Λ be the set of column indices arbitrarily chosen so that $|\Lambda| = k$ and construct a Gram matrix $\boldsymbol{G} = \boldsymbol{A}_\Lambda^T \boldsymbol{A}_\Lambda$. From the given conditions, $G_{ii} = 1$, for all i, and $|G_{ij}| \leq \mu$, for all $i \neq j$. Therefore, $r_i \triangleq \sum_{j \neq i} |G_{ij}| \leq (k-1)\mu$. According to the Gershgorin circle theorem (Lemma 1.14), all the eigenvalues of $\boldsymbol{G}$ should lie within $[1 \pm (k-1)\mu]$. To make it useful, we restrict $(k-1)\mu < 1$ or $k < 1 + 1/\mu$. Then,

$$\|\boldsymbol{A}_\Lambda \boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{G} \boldsymbol{x}} \in [\sqrt{1-(k-1)\mu}\|\boldsymbol{x}\|_2, \sqrt{1+(k-1)\mu}\|\boldsymbol{x}\|_2]. \tag{1.29}$$

Because this holds for every Λ such that $|\Lambda| = k$, the matrix $\boldsymbol{A}$ satisfies the RIP of k with $\delta_k = (k-1)\mu$.

1.5.2.4 Proof of Lemma 1.8

The posterior probability $\Pr(\boldsymbol{x}_i|\boldsymbol{y}_i; \boldsymbol{W})$ can be explicitly written as

$$\Pr(\boldsymbol{x}_i|\boldsymbol{y}_i; \boldsymbol{W}) = \frac{\Pr(\boldsymbol{x}_i) p(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{W})}{p(\boldsymbol{y}_i; \boldsymbol{W})} \tag{1.30}$$

where the numerator is independent of $\boldsymbol{W}$ since, in (1.13), $p(\boldsymbol{y}_i|\boldsymbol{x}_i; \boldsymbol{W}) = \frac{1}{(2\pi\sigma^2)^{m/2}} e^{-\frac{\|\boldsymbol{\eta}\|^2}{2\sigma^2}}$. Thus, we can alternatively rewrite the criterion as

$$\boldsymbol{W}^* = \arg\max_{\boldsymbol{W}} \mathbb{E}_{\boldsymbol{\eta}} \left[\sum_i -\log p(\boldsymbol{y}_i; \boldsymbol{W}) \right]. \tag{1.31}$$

The marginal log likelihood can be rewritten using the familiar, "free energy" functional (e.g. [109]):

$$-\sum_i \log p(\boldsymbol{y}_i; \boldsymbol{W}) = \min_{\boldsymbol{q}:\sum_j q_{ij}=1} -\sum_{ij} q_{ij} \log p(\boldsymbol{x}_j, \boldsymbol{y}_i; \boldsymbol{W}) + \sum_{ij} q_{ij} \log q_{ij}, \tag{1.32}$$

so that

$$\boldsymbol{W}^* = \arg\max_{\boldsymbol{W}} \min_{\boldsymbol{q}} F(\boldsymbol{W}, \boldsymbol{q}) \tag{1.33}$$

with

$$F(\boldsymbol{W}, \boldsymbol{q}) = \frac{1}{2\sigma^2} \sum_{ij} q_{ij} \|\boldsymbol{W}(\boldsymbol{x}_i - \boldsymbol{x}_j)\|^2 + \sum_{ij} q_{ij} \log q_{ij}. \tag{1.34}$$

The fixed-point equations are simply saying that at the optimal $\boldsymbol{W}^*$, minimizing $F(\boldsymbol{W}, \boldsymbol{q})$ with respect to $\boldsymbol{q}$ (Equation 1.14) and then maximizing with respect to $\boldsymbol{W}$ (Equation 1.15) should leave us at the same $\boldsymbol{W}$.

1.5.2.5 Details on Example 1.1

As $\sigma \to \infty$, the likelihood (1.13) approaches a constant for all $\boldsymbol{x}_j$'s, and assuming that $\boldsymbol{x}_j$'s are all equally likely, the posteriors q_{ij} will also be uniform. Thus the fixed-point equation (1.15) implies that the row vectors of $\boldsymbol{W}^*$ are simply the eigenvectors of $(\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$ and these are the principal components of the data.

1.5.2.6 Details on Example 1.2

Define a new dataset whose elements are the difference vectors $\boldsymbol{d}_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)$. By (1.15), the row vectors of $\boldsymbol{W}^*$ are the principal components of the dataset $\{\boldsymbol{d}_{ij}\}$ where each difference vector is weighted by $\sqrt{q_{ij}}$. Since $\boldsymbol{x}_i, \boldsymbol{x}_j$ both lie in an m-dimensional subspace, so does $\sqrt{q_{ij}}\boldsymbol{d}_{ij}$, regardless whatever value q_{ij}'s are. Hence, the rowspace of $\boldsymbol{W}^*$ will be this m-dimensional subspace. On the other hand, if the data lie in an m-dimensional basis, the top m principal components will also be an orthonormal basis of this m-dimensional

subspace.

1.5.2.7 Details on Example 1.3

Every k-sparse vector $\boldsymbol{x}$ has the unique projection $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ if $\boldsymbol{W}$ is random and has sufficiently many ($\geq 2k$) rows (see Section 1.1; see also Proposition 1.11 in Section 1.5.3). This means that the empirical posterior probability $\Pr(\boldsymbol{x}_i|\boldsymbol{y}_i; \boldsymbol{W})$ will approach one as $\sigma \to 0$ for all datapoints $\boldsymbol{x}_i$, maximizing the UCA utility function in (1.12).

1.5.2.8 Proof of Lemma 1.9

Since $\boldsymbol{x}$ is jointly Gaussian, $\boldsymbol{y}$ will also be Gaussian. If we denote, by $\boldsymbol{\Sigma}$, the covariance of $\boldsymbol{x}$, the entropy of $\boldsymbol{W}\boldsymbol{x}$ is given by $h(\boldsymbol{W}\boldsymbol{x}) = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\log\det(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)$. Let us relax the orthonormal constraint of the matrix $\boldsymbol{W}$. Instead, we assume a little *weaker* power budget constraint that $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = m$. To maximize $h(\boldsymbol{W}\boldsymbol{x})$ with a Lagrange multiplier ξ,

$$\frac{\partial h}{\partial \boldsymbol{W}} = (\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-1}\boldsymbol{W}\boldsymbol{\Sigma} - \xi\boldsymbol{W} = \boldsymbol{0}. \tag{1.35}$$

Post-multiplying $\boldsymbol{W}^T$ to each side of (1.35), we obtain $\boldsymbol{W}\boldsymbol{W}^T = \frac{1}{\xi}\boldsymbol{I}$, and ξ must be equal to one by the power budget constraint.

Let $\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ by its singular value decomposition (SVD), where $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and where $\boldsymbol{\Lambda}$ is a diagonal matrix. The equation (1.35) then becomes

$$\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{\Sigma} - \boldsymbol{W} = \boldsymbol{0}, \tag{1.36}$$

and pre-multiplying $\boldsymbol{\Lambda}\boldsymbol{U}^T$ to each side, we obtain

$$\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{\Sigma} - \boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{W} = \boldsymbol{0}, \tag{1.37}$$

from which we know that the row vectors of $\boldsymbol{U}^T\boldsymbol{W}$ should be the eigenvectors of $\boldsymbol{\Sigma}$, with the eigenvalues being the diagonal entries of $\boldsymbol{\Lambda}$. Because $\boldsymbol{U}$ is simply a rotation matrix, $\boldsymbol{W}$ is still a PCA projection (up to the rotation). Among all combinations of m principal

components, the "major" m principal components should be chosen to globally maximize $h(\boldsymbol{W}\boldsymbol{x})$.

Finally, it is obvious that the PCA projection is the maximizer of h under the original condition that $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{I}$ (tight frame) since it is the maximizer of h even for the weaker power budget constraint.

1.5.2.9 Proof of Lemma 1.10

The joint entropy $h(\boldsymbol{y})$ expands as $h(\boldsymbol{y}) = \sum_{i=1}^{m} h(y_i) - I(y_1; \ldots; y_m)$ where $I(y_1; \ldots; y_m)$ is the multi-information of $\boldsymbol{y}$. Thus to maximize $h(\boldsymbol{y})$, we need to maximize the individual marginal entropies as well as to minimize the multi-information. If we set $\boldsymbol{W}$ to be m ICA filters, $\boldsymbol{W}\boldsymbol{x}$ simply becomes a concatenation of m sources. The multi-information is therefore zero. Since g is the cdf of the sources, $g(\alpha_i)$ will be uniformly distributed, so it will maximize the marginal entropies as well [33].

1.5.3 Miscellaneous Lemmas

This section provides a set of lemmas which have been referred to in the body of this chapter or in the proofs of previous lemmas. In particular, Proposition 1.11 was referred to in pages 12, 19, 31. Lemma 1.12 was used in the proof of Lemma 1.3; Lemma 1.13 in the proof of Lemma 1.4; Lemma 1.14 in the proof of Lemma 1.7; Finally, Lemmas 1.15–1.17 will be used in the proof of Lemma 1.12.

Proposition 1.11. Let $\boldsymbol{W}$ be an $m \times n$ random matrix. Define $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$. With probability one, if $m \geq 2k$, then any k sparse signal has a unique projection.

Proof. Suppose, by contradiction, that there exists another k sparse vector $\boldsymbol{x}'$ such that $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}\boldsymbol{x}'$. Let J be a set of $2k$ indices that includes all the indices on which either $\boldsymbol{x}$ or $\boldsymbol{x}'$ is nonzero. Note that since both $\boldsymbol{x}$ and $\boldsymbol{x}'$ are k sparse, the union set of nonzero indices cannot be of size greater than $2k$. Define $\boldsymbol{W}_J$ to be an $m \times |J|$ submatrix of $\boldsymbol{W}$ obtained by taking all columns in J and all rows. By definition of matrix multiplication, $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}_J\boldsymbol{x}_J$ and $\boldsymbol{W}\boldsymbol{x}' = \boldsymbol{W}_J\boldsymbol{x}'_J$ (since the zero elements can be ignored in the matrix multiplication). This means that $\boldsymbol{W}_J\boldsymbol{x}'_J = \boldsymbol{W}_J\boldsymbol{x}_J$ with $\boldsymbol{x}_J \neq \boldsymbol{x}'_J$, which implies that the $|J|$

columns of $\boldsymbol{W}$ are linearly dependent. But since these columns of $\boldsymbol{W}$ are $|J|$ random m dimensional vectors and $|J| \leq m$, this happens with probability zero. $\square$

Lemma 1.12. Suppose that $\boldsymbol{W}$ satisfies the RIP of $2k$ with $\delta_{2k} < \sqrt{2}-1$. Given $\boldsymbol{x}, \widehat{\boldsymbol{x}} \in \mathbb{R}^n$, define $\boldsymbol{e} = \boldsymbol{x} - \widehat{\boldsymbol{x}}$. Let Λ_0 denote the index set corresponding to the k entries of $\boldsymbol{x}$ with largest magnitude and Λ_1 the index set corresponding to the k entries of $\boldsymbol{e}_{\Lambda_0^c}$ with largest magnitude. Set $\Lambda = \Lambda_0 \cup \Lambda_1$. If $\|\widehat{\boldsymbol{x}}\|_1 \leq \|\boldsymbol{x}\|_1$, then

$$\|\boldsymbol{e}\|_2 \leq C_0 \frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + C_2 \frac{|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}|}{\|\boldsymbol{e}_\Lambda\|_2}. \tag{1.38}$$

where

$$C_0 = 2\frac{1-(1-\sqrt{2})\delta_{2k}}{1-(1+\sqrt{2})\delta_{2k}}, \quad C_2 = \frac{2}{1-(1+\sqrt{2})\delta_{2k}}. \tag{1.39}$$

Proof. Note that Λ_1 is the index set corresponding to the k largest entries of $\boldsymbol{e}_{\Lambda_0^c}$ (in absolute value). We also define Λ_2 as the index set corresponding to the next k largest entries, and so on. Then, we observe that for $j \geq 2$,

$$\|\boldsymbol{e}_{\Lambda_j}\|_\infty \leq \frac{\|\boldsymbol{e}_{\Lambda_{j-1}}\|_1}{k} \tag{1.40}$$

because the Λ_j's sort $\boldsymbol{e}_{\Lambda_0^c}$ to have decreasing magnitude.

We first use the triangle inequality to bound $\|\boldsymbol{e}\|_2$:

$$\|\boldsymbol{e}\|_2 = \|\boldsymbol{e}_\Lambda + \boldsymbol{e}_{\Lambda^c}\|_2 \leq \|\boldsymbol{e}_\Lambda\|_2 + \|\boldsymbol{e}_{\Lambda^c}\|_2. \tag{1.41}$$

Then, $\|\boldsymbol{e}_{\Lambda^c}\|_2$ is bounded by

$$\|\boldsymbol{e}_{\Lambda^c}\|_2 = \left\| \sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j} \right\|_2 \leq \sum_{j\geq 2} \|\boldsymbol{e}_{\Lambda_j}\|_2 \overset{(a)}{\leq} \sqrt{k} \sum_{j\geq 2} \|\boldsymbol{e}_{\Lambda_j}\|_\infty \overset{(b)}{\leq} \frac{1}{\sqrt{k}} \sum_{j\geq 1} \|\boldsymbol{e}_{\Lambda_j}\|_1 = \frac{\|\boldsymbol{e}_{\Lambda_0^c}\|_1}{\sqrt{k}} \tag{1.42}$$

where the inequality (a) is due to Lemma 1.15 and the inequality (b) due to (1.40). We now

wish to bound $\|\boldsymbol{e}_{\Lambda_0^c}\|_1$. Since $\|\boldsymbol{x}\|_1 \geq \|\widehat{\boldsymbol{x}}\|_1$, by applying the triangle inequality, we obtain

$$\|\boldsymbol{x}\|_1 \geq \|\boldsymbol{x} - \boldsymbol{e}\|_1 = \|\boldsymbol{x}_{\Lambda_0} - \boldsymbol{e}_{\Lambda_0}\|_1 + \|\boldsymbol{x}_{\Lambda_0^c} - \boldsymbol{e}_{\Lambda_0^c}\|_1 \tag{1.43}$$

$$\geq \|\boldsymbol{x}_{\Lambda_0}\|_1 - \|\boldsymbol{e}_{\Lambda_0}\|_1 + \|\boldsymbol{e}_{\Lambda_0^c}\|_1 - \|\boldsymbol{x}_{\Lambda_0^c}\|_1. \tag{1.44}$$

Rearranging and applying again the triangle inequality,

$$\|\boldsymbol{e}_{\Lambda_0^c}\|_1 \leq \|\boldsymbol{x}\|_1 - \|\boldsymbol{x}_{\Lambda_0}\|_1 + \|\boldsymbol{e}_{\Lambda_0}\|_1 + \|\boldsymbol{x}_{\Lambda_0^c}\|_1 \tag{1.45}$$

$$\leq \|\boldsymbol{x} - \boldsymbol{x}_{\Lambda_0}\|_1 + \|\boldsymbol{e}_{\Lambda_0}\|_1 + \|\boldsymbol{x}_{\Lambda_0^c}\|_1. \tag{1.46}$$

Recalling that $\sigma_k(\boldsymbol{x})_1 = \|\boldsymbol{x}_{\Lambda_0^c}\|_1 = \|\boldsymbol{x} - \boldsymbol{x}_{\Lambda_0}\|_1$,

$$\|\boldsymbol{e}_{\Lambda_0^c}\|_1 \leq \|\boldsymbol{e}_{\Lambda_0}\|_1 + 2\sigma_k(\boldsymbol{x})_1. \tag{1.47}$$

Combining this with (1.42), we obtain

$$\|\boldsymbol{e}_{\Lambda^c}\|_2 \leq \frac{\|\boldsymbol{e}_{\Lambda_0}\|_1 + 2\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} \leq \|\boldsymbol{e}_{\Lambda_0}\|_2 + 2\frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}}. \tag{1.48}$$

where the last inequality follows from Lemma 1.15. By observing that $\|\boldsymbol{e}_{\Lambda_0}\|_2 \leq \|\boldsymbol{e}_{\Lambda}\|_2$, this combines with (1.41) to yield

$$\|\boldsymbol{e}\|_2 \leq 2\|\boldsymbol{e}_{\Lambda}\|_2 + 2\frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}}. \tag{1.49}$$

We now turn to establishing a bound for $\|\boldsymbol{e}_{\Lambda}\|_2$. By the RIP of $\boldsymbol{W}$,

$$(1 - \delta_{2k})\|\boldsymbol{e}_{\Lambda}\|_2^2 \leq \|\boldsymbol{W}\boldsymbol{e}_{\Lambda}\|_2^2. \tag{1.50}$$

since $\boldsymbol{e}_{\Lambda} \in \mathcal{S}_{2k}$. Using the equality that $\boldsymbol{W}\boldsymbol{e}_{\Lambda} = \boldsymbol{W}\boldsymbol{e} - \sum_{j\geq 2} \boldsymbol{W}\boldsymbol{e}_{\Lambda_j}$, we can rewrite (1.50) as

$$(1 - \delta_{2k})\|\boldsymbol{e}_{\Lambda}\|_2^2 \leq \boldsymbol{e}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{e}_{\Lambda} - \sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{e}_{\Lambda}. \tag{1.51}$$

In order to bound the second term of (1.51), we use lemma 1.17, which implies that

$$|\boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda_i}| \leq \delta_{2k} \|\boldsymbol{e}_{\Lambda_i}\|_2 \|\boldsymbol{e}_{\Lambda_j}\|_2 \tag{1.52}$$

for any $i \neq j$. Furthermore, Lemma 1.16 gives $\|\boldsymbol{e}_{\Lambda_0}\|_2 + \|\boldsymbol{e}_{\Lambda_1}\|_2 \leq \sqrt{2}\|\boldsymbol{e}_{\Lambda}\|_2$. Thus, we obtain

$$\left|\sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda}\right| = \left|\sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda_0} + \sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda_1}\right| \tag{1.53}$$

$$\leq \sum_{j\geq 2} \left|\boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda_0}\right| + \sum_{j\geq 2} \left|\boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda_1}\right| \tag{1.54}$$

$$\leq \sum_{j\geq 2} \delta_{2k} \|\boldsymbol{e}_{\Lambda_0}\|_2 \|\boldsymbol{e}_{\Lambda_j}\|_2 + \sum_{j\geq 2} \delta_{2k} \|\boldsymbol{e}_{\Lambda_1}\|_2 \|\boldsymbol{e}_{\Lambda_j}\|_2 \tag{1.55}$$

$$\leq \sqrt{2}\delta_{2k} \|\boldsymbol{e}_{\Lambda}\|_2 \sum_{j\geq 2} \|\boldsymbol{e}_{\Lambda_j}\|_2. \tag{1.56}$$

In (1.42), we have bounded $\sum_{j\geq 2} \|\boldsymbol{e}_{\Lambda_j}\|_2$ by $\sum_{j\geq 2} \|\boldsymbol{e}_{\Lambda_j}\|_2 \leq \|\boldsymbol{e}_{\Lambda_0^c}\|_1/\sqrt{k}$, so

$$\left|\sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda}\right| \leq \sqrt{2}\delta_{2k} \|\boldsymbol{e}_{\Lambda}\|_2 \frac{\|\boldsymbol{e}_{\Lambda_0^c}\|_1}{\sqrt{k}}. \tag{1.57}$$

Combining (1.57) and (1.51), we obtain

$$(1-\delta_{2k})\|\boldsymbol{e}_{\Lambda}\|_2^2 \leq \left|\boldsymbol{e}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda} - \sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda}\right| \tag{1.58}$$

$$\leq \left|\boldsymbol{e}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda}\right| + \left|\sum_{j\geq 2} \boldsymbol{e}_{\Lambda_j}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda}\right| \tag{1.59}$$

$$\leq \left|\boldsymbol{e}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda}\right| + \sqrt{2}\delta_{2k} \|\boldsymbol{e}_{\Lambda}\|_2 \frac{\|\boldsymbol{e}_{\Lambda_0^c}\|_1}{\sqrt{k}}. \tag{1.60}$$

Dividing both sides of (1.60) by $(1-\delta_{2k})\|\boldsymbol{e}_{\Lambda}\|_2$, we can bound $\|\boldsymbol{e}_{\Lambda}\|_2$ by

$$\|\boldsymbol{e}_{\Lambda}\|_2 \leq \underbrace{\frac{\sqrt{2}\delta_{2k}}{(1-\delta_{2k})}}_{\triangleq \alpha} \frac{\|\boldsymbol{e}_{\Lambda_0^c}\|_1}{\sqrt{k}} + \underbrace{\frac{1}{(1-\delta_{2k})}}_{\triangleq \beta} \frac{|\boldsymbol{e}^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}_{\Lambda}|}{\|\boldsymbol{e}_{\Lambda}\|_2}. \tag{1.61}$$

Combining this with (1.47) and applying Lemma 1.15, we obtain

$$\|\boldsymbol{e}_\Lambda\|_2 \leq \alpha \frac{\|\boldsymbol{e}_{\Lambda_0^c}\|_1}{\sqrt{k}} + \beta \frac{|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}|}{\|\boldsymbol{e}_\Lambda\|_2} \tag{1.62}$$

$$\leq \alpha \frac{\|\boldsymbol{e}_{\Lambda_0}\|_1 + 2\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + \beta \frac{|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}|}{\|\boldsymbol{e}_\Lambda\|_2} \tag{1.63}$$

$$\leq \alpha \|\boldsymbol{e}_{\Lambda_0}\|_2 + 2\alpha \frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + \beta \frac{|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}|}{\|\boldsymbol{e}_\Lambda\|_2}. \tag{1.64}$$

Since $\|\boldsymbol{e}_{\Lambda_0}\|_2 \leq \|\boldsymbol{e}_\Lambda\|_2$,

$$(1-\alpha)\|\boldsymbol{e}_\Lambda\|_2 \leq 2\alpha \frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + \beta \frac{|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}|}{\|\boldsymbol{e}_\Lambda\|_2}. \tag{1.65}$$

The assumption that $\delta_{2k} < \sqrt{2} - 1$ ensures that $\alpha < 1$. Dividing by $(1-\alpha)$ and combining with (1.49) results in

$$\|\boldsymbol{e}\|_2 \leq \frac{2(1+\alpha)}{1-\alpha} \frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + \frac{2\beta}{1-\alpha} \frac{|\boldsymbol{e}_\Lambda^T \boldsymbol{W}^T \boldsymbol{W} \boldsymbol{e}|}{\|\boldsymbol{e}_\Lambda\|_2}. \tag{1.66}$$

Plugging in for $\alpha = \frac{\sqrt{2}\delta_{2k}}{(1-\delta_{2k})}$ and $\beta = \frac{1}{(1-\delta_{2k})}$ yields the desired constants. □

Lemma 1.13 (Lemma 5.36 of [141]). Let $\sigma_i(\cdot)$ and $\lambda_i(\cdot)$ denote the singular values and the eigenvalues of the input argument matrix, respectively. Suppose that, for some $\epsilon > 0$, a matrix $\boldsymbol{A}$ satisfies

$$|\lambda_i(\boldsymbol{A}^T\boldsymbol{A}) - 1| \leq \max(\epsilon, \epsilon^2), \quad \forall i. \tag{1.67}$$

Then,

$$|\sigma_i(\boldsymbol{A}) - 1| \leq \epsilon, \quad \forall i. \tag{1.68}$$

Conversely, if $\boldsymbol{A}$ satisfies the inequality (1.68) for some $\epsilon > 0$, then $|\lambda_i(\boldsymbol{A}^T\boldsymbol{A}) - 1| \leq 3\max(\epsilon, \epsilon^2)$, for all i.

Proof. Note that $\sigma_i(\boldsymbol{A}) = \sqrt{\lambda_i(\boldsymbol{A}^T\boldsymbol{A})}$. Therefore,

$$|\sigma_i(\boldsymbol{A}) - 1| \leq \epsilon \;\Leftrightarrow\; \left|\sqrt{\lambda_i(\boldsymbol{A}^T\boldsymbol{A})} - 1\right| \leq \epsilon. \tag{1.69}$$

The lemma simply follows from the elementary inequality

$$\max(|z-1|, |z-1|^2) \leq |z^2 - 1| \leq 3\max(|z-1|, |z-1|^2), \quad \forall z \geq 0. \tag{1.70}$$

□

Lemma 1.14 (Gershgorin circle theorem [65])**.** Let $\boldsymbol{A}$ be an $n \times n$ matrix, with entries A_{ij}. For $i \in \{1, \ldots, n\}$, let $r_i = \sum_{j \neq i} |A_{ij}|$. Let $D_i(A_{ii}, r_i)$ be the closed disc centered at A_{ii} with radius r_i. Then, every eigenvalue of $\boldsymbol{A}$ lies within at least one of the discs $D_i(A_{ii}, r_i)$.

Proof. Let λ be an eigenvalue of $\boldsymbol{A}$ and let $\boldsymbol{v} = (v_1, \ldots, v_n)$ be a corresponding eigenvector. Define $i^* = \arg\max_i |v_i|$. Necessarily, $|v_{i^*}| > 0$; otherwise, $\boldsymbol{v} = \boldsymbol{0}$. Because $\boldsymbol{v}$ satisfies $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$,

$$\sum_j A_{ij} v_j = \lambda v_i, \quad i = 1, \ldots, n. \tag{1.71}$$

We choose $i = i^*$ and split the sum to obtain

$$\sum_{j \neq i^*} A_{i^* j} v_j = \lambda v_{i^*} - A_{i^* i^*} v_{i^*}. \tag{1.72}$$

Dividing both sides by v_{i^*} and taking the absolute value, we obtain

$$|\lambda - A_{i^* i^*}| = \left|\frac{\sum_{j \neq i^*} A_{i^* j} v_j}{v_{i^*}}\right| \leq \sum_{j \neq i^*} |A_{i^* j}| \left|\frac{v_j}{v_{i^*}}\right| \leq \sum_{j \neq i^*} |A_{i^* j}| = r_{i^*}. \tag{1.73}$$

Therefore, $\lambda \in D_{i^*}(A_{i^* i^*}, r_{i^*})$. □

Lemma 1.15. Suppose $\boldsymbol{x} \in \mathcal{S}_k$. Then,

$$\frac{\|\boldsymbol{x}\|_1}{\sqrt{k}} \leq \|\boldsymbol{x}\|_2 \leq \sqrt{k}\|\boldsymbol{x}\|_\infty. \tag{1.74}$$

Proof. Take Λ, with $|\Lambda| = k$, so that it includes the support of $\boldsymbol{x}$. Define $\boldsymbol{z}$ as the k-dimensional vector comprising the elements, indexed by Λ, of $\boldsymbol{x}$. By the norm inequalities (e.g., [131]),

$$\frac{\|\boldsymbol{z}\|_1}{\sqrt{k}} \leq \|\boldsymbol{z}\|_2 \leq \sqrt{k}\|\boldsymbol{z}\|_\infty. \tag{1.75}$$

Because, for any p,

$$\|\boldsymbol{x}\|_p = \left(\sum_{i\in\Lambda\cup\Lambda^c} x_i^p\right)^{1/p} = \left(\sum_{i\in\Lambda} x_i^p\right)^{1/p} = \|\boldsymbol{z}\|_p, \tag{1.76}$$

we can replace $\|\boldsymbol{z}\|_p$, in (1.75), with $\|\boldsymbol{x}\|_p$. Then, we obtain the lemma. □

Lemma 1.16. Suppose that $\boldsymbol{u}$, $\boldsymbol{v}$ are orthogonal vectors. Then,

$$\|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2 \leq \sqrt{2}\|\boldsymbol{u}+\boldsymbol{v}\|_2. \tag{1.77}$$

Proof. We begin with the inequality $(\|\boldsymbol{u}\|_2 - \|\boldsymbol{v}\|_2)^2 \geq 0$, which expands as

$$2\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2 \leq \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 = \|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2 + 4\boldsymbol{u}^T\boldsymbol{v} \tag{1.78}$$

because $\boldsymbol{u}^T\boldsymbol{v} = 0$. Adding $\|\boldsymbol{u}\|_2^2 + \|\boldsymbol{v}\|_2^2$ to, and subsequently factorizing, both sides of (1.78), we obtain

$$(\|\boldsymbol{u}\|_2 + \|\boldsymbol{v}\|_2)^2 \leq 2\|\boldsymbol{u}+\boldsymbol{v}\|_2^2. \tag{1.79}$$

The inequality (1.77) immediately follows by taking the square root of both sides of (1.79).

□

Lemma 1.17. If $\boldsymbol{W}$ satisfies the RIP of $2k$, then for any two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{S}_k$ with disjoint support,

$$|\boldsymbol{u}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{v}| \leq \delta_{2k}\|\boldsymbol{u}\|_2\|\boldsymbol{v}\|_2. \tag{1.80}$$

Proof. If we let $\widetilde{\boldsymbol{u}} = \boldsymbol{u}/\|\boldsymbol{u}\|_2$ and $\widetilde{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|_2$, then $\widetilde{\boldsymbol{u}} \pm \widetilde{\boldsymbol{v}} \in \mathcal{S}_{2k}$ and $\|\widetilde{\boldsymbol{u}} \pm \widetilde{\boldsymbol{v}}\|_2^2 = \|\widetilde{\boldsymbol{u}}\|_2^2 + \|\widetilde{\boldsymbol{v}}\|_2^2 = 2$. Using the RIP, we have

$$2(1 - \delta_{2k}) \leq \|\boldsymbol{W}(\widetilde{\boldsymbol{u}} \pm \widetilde{\boldsymbol{v}})\|_2^2 \leq 2(1 + \delta_{2k}). \tag{1.81}$$

Applying the polarization identity, we have

$$|\widetilde{\boldsymbol{u}}^T \boldsymbol{W}^T \boldsymbol{W} \widetilde{\boldsymbol{v}}| = \frac{1}{4} \left| \|\boldsymbol{W}\widetilde{\boldsymbol{u}} + \boldsymbol{W}\widetilde{\boldsymbol{v}}\|_2^2 - \|\boldsymbol{W}\widetilde{\boldsymbol{u}} - \boldsymbol{W}\widetilde{\boldsymbol{v}}\|_2^2 \right| \leq \delta_{2k}. \tag{1.82}$$

Substituting back $\widetilde{\boldsymbol{u}} = \boldsymbol{u}/\|\boldsymbol{u}\|_2$ and $\widetilde{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|_2$ in (1.82), we finally obtain (1.80). □

Chapter 2

Informative Sensing

Compressed sensing is a set of mathematical results showing that sparse signals can be reconstructed from incomplete linear measurement samples, substantially below the Nyquist rate. Interestingly, random measurements allow good reconstruction almost regardless of the basis in which the signals are sparse. The universality of random measurements, however, means that they are not particularly tuned to a specific distribution of signals. Surely, if we knew something about the statistics of the signals on which we want to do compressed sensing, we should be able to design a projection that is optimal for the class of signals.

To this end, we revisit the classical InfoMax criterion. We seek an undercomplete linear projection $\boldsymbol{W}$ that maximizes the mutual information between the input $\boldsymbol{x}$ and output $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$. For very special cases, the solution for the optimal $\boldsymbol{W}$ is known and may coincide with principal components (when $\boldsymbol{x}$ is Gaussian) or independent components (when $\boldsymbol{y}$ goes through a pointwise sigmoidal nonlinearity). But the solution for a general input distribution and a linear network is still unknown.

In this chapter, we focus on the input signals that have a sparse representation $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$ in an orthonormal basis $\boldsymbol{D}$, as typical in the compressed sensing literature. By analysis, we find that in i.i.d. settings of $\boldsymbol{\alpha}$, random projections are asymptotically InfoMax optimal, making an interesting connection to the classical theory of compressed sensing. But, in non-i.i.d. settings, we find a rather novel set of projections by optimizing the InfoMax criterion, which turn out to consistently outperform random or PCA projections in signal reconstruction experiments.

2.1 Introduction

We consider a classic question: can we determine an n-dimensional vector $\boldsymbol{x}$ given m linear equations, $\boldsymbol{W}\boldsymbol{x} = \boldsymbol{y}$, when $m < n$? As is well known, this is an ill-posed problem for which there exist an infinite number of solutions. Obviously, however, if we are given side information that $\boldsymbol{x}$ lies in a particular low-dimensional linear subspace, say of dimension k, then $m = k$ is sufficient to exactly reconstruct the original vector. What if $\boldsymbol{x}$ lies in a fractional but *nonlinear* manifold? Can we still get away with fewer than n measurement samples?

The theory of compressed sensing (CS) deals with the above question, particularly for k-sparse signals, which can be represented with a small number, at most k, of nonzero elements in some basis [51, 32]. A basic result is that, with probability 1, any $m \times n$ "random" matrix $\boldsymbol{W}$ suffices to make the mapping from $\boldsymbol{x}$ to $\boldsymbol{W}\boldsymbol{x}$ invertible on the entire set of k-sparse signals if $m \geq 2k$ (see Proposition 1.11 or [116]). Furthermore, the recovery can be performed by a simple convex optimization [31] or by a greedy optimization procedure [136], if m increases to the order of $k \log(n/k)$.

These results have generated huge excitement in both theoretical and practical communities. On the theoretical side, the performance of CS with the restricted isometry property (RIP) [32] has rigorously been analyzed when the signals are not exactly k-sparse but rather *compressible* (i.e., can be well approximated with a small number of nonzero elements) [51, 32, 119, 40, 114] as well as when the measurements are quantized [22, 157] or contaminated with noise [75, 143, 34, 63]. Several attempts to deploy CS for analog signals have also been developed [137, 88, 60]. On the practical side, applications of CS have been explored in building "single-pixel" cameras [54], medical imaging [99, 128] and geophysical data analysis [92, 110], etc.

In fact, the classical theory of CS says little about what types of linear measurements enable the best recovery for a particular class of non-ideally sparse signals when m is fixed. To cope with the worst-case, which scarcely occurs in a certain class, the RIP basically requires that the measurement be mutually incoherent with the basis in which the signals are assumed to be sparse (see also [27] for the RIPless theory). While recently a set of deter-

ministic matrices satisfying the RIP with relatively small m are found [81, 25, 151, 62], random projections have typically been used [51, 32, 11] because they prove mutually incoherent with almost any basis. However, the universality of random projections does not mean that they are *universally optimal* for every class of sparse signals, as we will demonstrate shortly. Elad [57] has shown that increasing the average incoherence of a measurement matrix using an iterative algorithm, can give a small increase in the recovery performance (see also [55, 150] for relevant subsequent works). For natural images, the standard low-pass filtering, e.g., the measurement based on principal component analysis (PCA), often gives better reconstruction results than random projections in noisy and noiseless settings [74, 127]. Lustig, Donoho, and Pauly [98] noticed that undersampling low-pass signals less than high-pass signals can produce a better performance for real images when using a random Fourier matrix. In a similar context, Romberg [115] first takes a fixed small number of PCA coefficients, which are at low frequencies, to capture a holistic outline of the image data before switching to random projections for filling in the details. The necessity of a better model, beyond simple sparsity or compressibility, has also been recognized in terms of signal recovery [78, 10] as well as in terms of measurement matrix design [146, 34, 127, 64]. More accurate side information (e.g., group sparsity or probabilistic model) may shrink, or better describe, the space on which the signal $\boldsymbol{x}$ can actually sit. It helps to generally improve the reconstruction error given any fixed measurement and also to specialize the measurement matrix.

The following is a quick demonstration that random projections are not universally optimal for every class of sparse signals. First consider $\boldsymbol{x} = (x_1, \ldots, x_n)$ where x_i's are i.i.d. The marginal probability density is given by $p(x_i) = \frac{1}{10}\mathcal{N}(x_i; 0, 1^2) + \frac{9}{10}\mathcal{N}(x_i; 0, 0.1^2)$, where $\mathcal{N}(\cdot; \mu, \sigma^2)$ denotes a Gaussian density with mean μ and variance σ^2 (see, for the density, Figure 2.1, top left). If n is large, the signal has the order statistics shown in Figure 2.1 (top right), with little deviation (cf. Sanov's theorem [43]). This signal is seemingly compressible (refer to Section 1.1 or [26, 35] for the precise definition of compressibility) and the theory of CS applies well. In Figure 2.1 (bottom left), we show the performance of random projections when the signal $\boldsymbol{x}$ is reconstructed either by minimizing $\|\boldsymbol{x}\|_1$ (ℓ_1-regularization) or by minimizing the mean-squared error (MSE) subject to the measurement

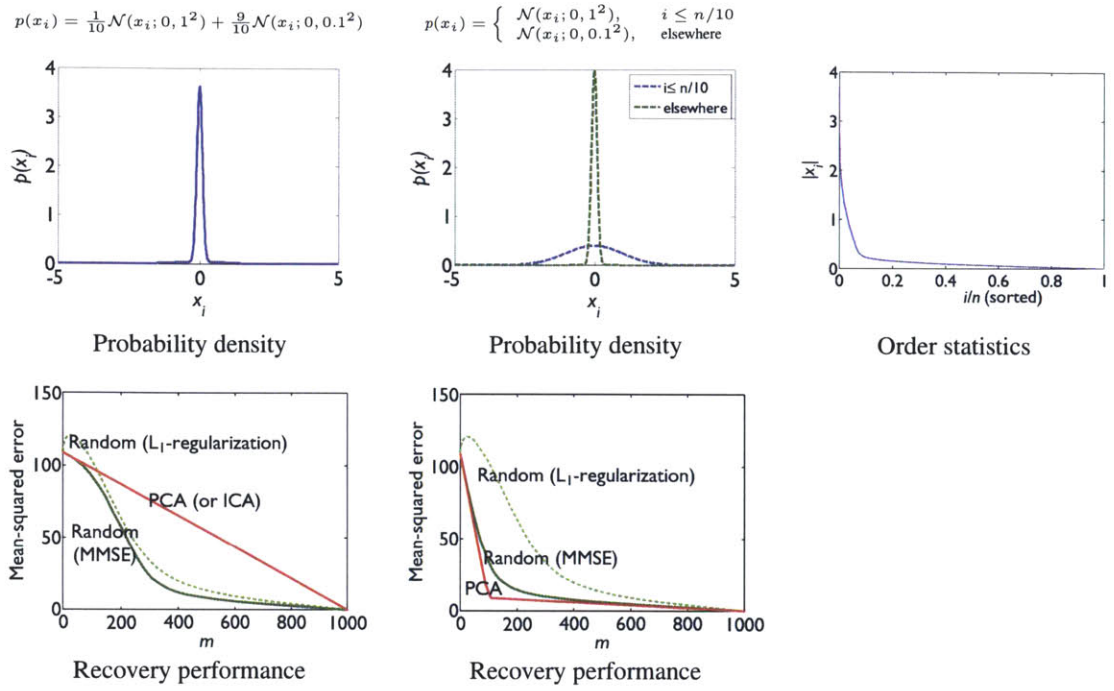


Figure 2.1: Compressed sensing for two classes of signals which have the same order statistics. One is an i.i.d. Gaussian mixture signal and the other is a non-i.i.d. Gaussian signal. Top: probability densities of the two signals and the common order statistics. Bottom: recovery performance of PCA and random projections. For recovery, two alternative methods (MMSE and ℓ_1-regularization) have been employed. When the PCA projections are used, mean-squared errors from the two recovery methods happen to coincide. Note that the relative performances of the PCA versus random projections are completely different in the two cases.

(a specific algorithm to compute the minimum MSE (MMSE) estimate will be given in Section 2.6.3). Random projections enable a good reconstruction with ℓ_1-regularization if m is sufficiently large. A slightly better reconstruction comes with MMSE recovery; more importantly, the MMSE scheme always gives a better reconstruction than (at least as good as) the canonical linear recovery scheme, even when m is very small. The PCA projection, coherent to the sparse basis in this case,[1] cannot enjoy the benefits of such nonlinear recovery schemes at all and shows inferior recovery performance in comparison with random projections. As a matter of fact, the order statistics shown in Figure 2.1 (top right) are not unique to the i.i.d. signal but also shared with a Gaussian signal of the following density

[1]More exactly, this is the projection based on independent component analysis (ICA).

(see, for the density, Figure 2.1, top middle):

$$x_i \sim \begin{cases} \mathcal{N}(0, 1^2), & i \leq n/10 \\ \mathcal{N}(0, 0.1^2), & \text{elsewhere.} \end{cases}$$

The same order statistics mean the same level of sparsity, in the view of the RIP-based theory; ℓ_1-regularization with random projections performs exactly as well as in the previous i.i.d. case. For the Gaussian signal, however, the MMSE estimate, based on the true prior, can greatly reduce the reconstruction error (see Figure 2.1, bottom right), which corroborates the necessity of a precise model during signal recovery. At the sensor side, still random projections may not be a bad choice, but they are far from being optimal for this particular signal. As shown in Figure 2.1 (bottom right), the PCA projection works better in terms of MSE. If the MMSE estimate is used for signal recovery, the PCA projection is 3.79 times better (equivalently 5.79dB) than random projections when $m/n = 0.1$ and 1.88 times (2.73dB) better when $m/n = 0.2$.

Besides the above two, infinitely many classes of signals have the same order statistics. Given only the order statistics, perhaps we must assume the i.i.d. signal model because it is the maximum entropy distribution subject to the given order statistics,[2] which goes along with the worst-case consideration by the RIP-based theory. However, if more information is available so that the signal class is better discriminated, what types of linear measurements are optimal for each class? In general, they are none of random, PCA, or ICA. This question has motivated this work, the rationale for which is somewhat analogous to that for CS: The classical theory of CS adopted a sparsity model, rather than bandlimitedness in Fourier domain, for better description of signals. Likewise, we use a probability density, whether it is true or maximizing entropy subject to all available information, rather than simple sparsity.

Meanwhile, the InfoMax principle has long been established for the design of a linear sensory system in the field of computational neuroscience since proposed by Linsker [94]. According to this principle, the goal of a sensory system is to maximize the mutual

[2] In Bayesian probability, the principle of maximum entropy [82] is a postulate which states that, subject to known constraints (called testable information), the probability distribution which best represents the current state of knowledge is the one with largest entropy.

information between the sensors and the world (see also [8, 12, 7]). In particular, Linsker showed that the PCA projection is the maximally informative linear sensory system for the Gaussian input. There has been a tremendous amount of subsequent work in terms of finding algorithms to estimate the mutual information in a linear system as well as relationships between InfoMax and other learning criteria (e.g. [61, 90]). In [18], Bell and Sejnowski showed that when the InfoMax principle is applied to linear projections followed by a specific sigmoidal nonlinearity, then maximizing the mutual information can be equivalent to finding sensors that are as independent as possible. The self-organizing neural network [17] and relevant component analysis (RCA) [9] are still other well-known examples where the InfoMax criterion was used (to reconcile the stimuli-sensitive responses of neighboring cells and to alleviate the undesired data variability in a Mahalanobis distance, respectively).

In this chapter, we show that the InfoMax principle could also have predicted a similar story as the classical theory of CS. For the inputs which have a sparse representation $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$, where $\boldsymbol{D}$ is an orthonormal basis sparsifying its coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, we seek an undercomplete linear projection that is most informative about the inputs. In the settings where α_i's are i.i.d. and n is asymptotically large, we find that random projections are indeed InfoMax optimal. This is somewhat surprising in that they are very different than the "structured" projections (e.g., [19]) implemented by the same principle (with different settings, of course) in the past. The incoherence of $\boldsymbol{W}$ with respect to $\boldsymbol{D}$ is central in the optimization of the InfoMax criterion, which finds an interesting connection to the classical theory of CS. However, if α_i's are not i.i.d., we observe that random projections can be substantially far from being InfoMax optimal. We particularly explore the case where $\{\alpha_i/\sigma_i\}$ are i.i.d. with nonuniform variances σ_i^2 (see Section 2.3), by assuming the knowledge, beyond simple sparsity, on where the sparsity comes from, the non-Gaussianity or asymmetric variances of the signal elements. Such a model is often realistic. For example, wavelet coefficients of natural images are sparse with the variance falling off as a power law along the increasing spatial frequency [140] . For very special cases, the InfoMax optimal projection may coincide with the PCA projection but generally not. We develop a set of mathematical tools to solve the InfoMax problem and derive a novel measurement scheme that (approximately) optimizes the InfoMax criterion. In our signal recovery experiments,

the InfoMax-based projections consistently outperform random projections as well as the PCA projections.

To summarize, our results suggest the utility of InfoMax for the application of CS. Given the input distribution, InfoMax provides an optimization criterion, unlike the RIP, which is a *sufficient* condition for bounded-error recovery by ℓ_1-regularization. InfoMax does not assume any specific recovery scheme, but the utility can be maximally demonstrated with the best one we can use. For this reason, we mainly use the MMSE estimate rather than ℓ_1-regularization. Our MMSE estimate usually takes a much longer time than ℓ_1-regularization, but the recovery cost is not a focus of this work. Emerging research trends in CS include efficient computation of the MMSE or maximum a posteriori (MAP) estimates (e.g., see [113]). In non-i.i.d. sparse cases where the InfoMax optimal projection deviates from random projections, the InfoMax-based story has its most novelty, in comparison with the classical theory. However, even in i.i.d. cases, where InfoMax favors random projections, the claim regarding recovery (or inference) is different from, *not* contradictory to, the RIP-based one. For example, the InfoMax optimality does not entail a success of ℓ_1-regularization or orthogonal matching pursuit. Instead, it guarantees maximal mutual information between *any* i.i.d. signals and random measurements, or equivalently minimal uncertainty of the unobserved part of the signals (we will show this equivalence in Section 2.2), for *any* m, as long as n is asymptotically large. While no "good-quality" reconstruction may be possible if the signal is not *compressible* or if $m < O(k \log(n/k))$, we still seek the "best" measurement and inference. For example, recall that in Figure 2.1 (bottom left), random measurements constantly gave a better reconstruction than the PCA projection or any no-inference (linear) scheme, even when m was extremely small, if we employed the MMSE estimate.

2.2 Information Maximization

Let $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y} \in \mathbb{R}^m$ be the input and the output of a sensory system, related by $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta}$, where $\boldsymbol{W}$ denotes an $m \times n$ measurement matrix and $\boldsymbol{\eta}$ represents the

sensor noise. The input-output mutual information is defined as

$$I(\boldsymbol{x};\boldsymbol{y}) \triangleq h(\boldsymbol{y}) - h(\boldsymbol{y}|\boldsymbol{x}) \tag{2.1}$$

where $h(\boldsymbol{y})$ is the entropy of the output and where $h(\boldsymbol{y}|\boldsymbol{x})$ denotes the remaining entropy of the output given the input signal and thus merely the entropy of the sensor noise. Because the noise entropy remains constant to whatever the measurement matrix is, $h(\boldsymbol{y})$ may directly be used, instead of $I(\boldsymbol{x};\boldsymbol{y})$, as the InfoMax criterion to maximize. Without any nonlinearity involved, $h(\boldsymbol{y})$ can be made arbitrarily large, simply by taking $\boldsymbol{W} = c\boldsymbol{W}_o$ with $c \to \infty$ for any fixed $\boldsymbol{W}_o$. A practical convention is to restrict the total power of the sensors (or the squared sum of all entries of $\boldsymbol{W}$), which plays a role of precluding such a trivial manipulation.

In this chapter, we focus on noiseless and undercomplete cases without any nonlinearity, i.e., $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ with $m < n$. Under the power budget constraint $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = m$, we have two immediate lemmas regarding the multiplicity of solutions and a necessary condition on optimal matrices:

Lemma 2.1. Under the power budget constraint $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = m$, a matrix $\boldsymbol{W}$ that maximizes $h(\boldsymbol{W}\boldsymbol{x})$ is not unique.

Proof. See Section 2.6.1.1. □

Lemma 2.2. Under the power budget constraint $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = m$, the maximum entropy can always be obtained only with a tight frame matrix satisfying $\boldsymbol{W}\boldsymbol{W}^T = \boldsymbol{I}$.

Proof. See Section 2.6.1.2. □

Based on the above lemmas, the solution of linear InfoMax will only be defined up to a rotation in the sensor space (i.e., rowspace), but we can safely, without any loss of generality, restrict our search to the space of $m \times n$ tight frames.

2.2.1 Properness for Compressed Sensing

The InfoMax principle has been widely accepted in the field of computational neuroscience and machine learning (e.g., [18, 17, 9]) since it was introduced by Linsker. However, it is

not so straightforward whether the principle is appropriate for the application of CS where the goal primarily pertains to the error of the reconstructed signal. A good reason, although not entirely satisfactory, is that, in the setting of our interest, InfoMax minimizes the uncertainty of the signal $\boldsymbol{y}^{\perp}$ which is in the null space of $\boldsymbol{W}$ [124], given the measurement $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$, where the uncertainty is measured by Shannon's entropy. This can be easily shown as below:

For any deterministic function f, $h(\boldsymbol{y}^{\perp}|f(\boldsymbol{y})) \geq h(\boldsymbol{y}^{\perp}|\boldsymbol{y})$, which is provable with the data processing inequality [43]. Any post-processing, whether linear or nonlinear, would not decrease the posterior entropy. The equality holds if f is invertible, and therefore, $h(\boldsymbol{y}^{\perp}|\boldsymbol{W}_o\boldsymbol{x}) = h(\boldsymbol{y}^{\perp}|\boldsymbol{W}\boldsymbol{x})$ with $\boldsymbol{W}$ being a *normalization* of the original sensing matrix $\boldsymbol{W}_o$, i.e., $\boldsymbol{W} = (\boldsymbol{W}_o\boldsymbol{W}_o^T)^{-1/2}\boldsymbol{W}_o$. The normalized matrix $\boldsymbol{W}$ always becomes a tight frame, as linear InfoMax ultimately focuses. We have

$$h(\boldsymbol{y}^{\perp}|\boldsymbol{W}_o\boldsymbol{x}) = h(\boldsymbol{y}^{\perp}|\boldsymbol{W}\boldsymbol{x}) = h(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{y}^{\perp}) - h(\boldsymbol{W}\boldsymbol{x}), \tag{2.2}$$

where $(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{y}^{\perp})$ is simply a linear transformation (or basis change) of $\boldsymbol{x}$, with the Jacobian factor equal to $\det(\boldsymbol{W}\boldsymbol{W}^T)^{1/2} = 1$; therefore, $h(\boldsymbol{W}\boldsymbol{x}, \boldsymbol{y}^{\perp}) = h(\boldsymbol{x})$. Because $h(\boldsymbol{x})$ is a constant, the posterior entropy $h(\boldsymbol{y}^{\perp}|\boldsymbol{W}\boldsymbol{x})$ is minimized by $\boldsymbol{W}$ that maximizes $h(\boldsymbol{W}\boldsymbol{x})$.

On the other hand, MMSE is simply another type of the uncertainty measure of the same signal $\boldsymbol{y}^{\perp}$ given the same measurement $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$. The measurement matrices that achieve InfoMax and the smallest MMSE are not theoretically guaranteed to be the same. However, the projection designed to minimize the uncertainty in one (practically easier) measure often minimizes the uncertainty in another measure. Our empirical findings show close similarity in the behavior of the two uncertainty measures relative to measurement matrices (e.g., see Section 2.2.2 and Section 2.4). In a theoretical aspect, the entropy power forms a lower bound of the MMSE, i.e., $\frac{1}{n-m}\mathsf{mmse}(\boldsymbol{y}^{\perp}|\boldsymbol{y}) \geq \frac{1}{2\pi e}\exp(\frac{1}{n-m}h(\boldsymbol{y}^{\perp}|\boldsymbol{y}))$. In this regard, InfoMax may be interpreted as minimizing a lower bound of the MMSE criterion. Extensive studies to reveal the relationship between information-theoretic measures (entropy, mutual information, etc.) and the MMSE are still active. Some good examples are found in a series of the papers by Guo, Shamai, and Verdú (e.g., see [69, 70]).

In the context of CS, mutual information has been employed [122] to define the Gaussian channel capacity in modeling a *noisy* measurement process. The capacity, $\frac{1}{2}\log(1 + \mathsf{snr})$, is then used to form a lower-bound of the measurement rate required to keep the distortion below a given level. A number of similar studies have followed (e.g., see [144, 5, 1] and references therein). However, the studies are only about identifying the required measurement rates, not about designing measurement matrices, and they assume a noisy measurement process mostly on ideally sparse signals.

Minimizing the posterior entropy (a.k.a. Bayesian experimental design) has also been proposed in regard to CS; some for nonadaptive *one-shot* settings of measurements [146, 64], others for adaptive *sequential* settings [83, 34, 127]. They are certainly relevant to, and indeed have inspired, this work. Perhaps, the most notable feature of our work is that ours is analytic, whereas the prior works are all algorithmic. As such, this study reveals that InfoMax has an aspect strongly favorable to random matrices, which makes a novel connection to the RIP-based theory. The fact that we focus on the noiseless setting while the others on noisy settings is another difference.

2.2.2 Toy Example

To further motivate the utility of InfoMax in CS, let us consider a very simple but interesting example shown in Figure 2.2. The signal x lies in $\mathbb{R}^2$, occupying the space with a mixture density of four Gaussian clusters (see the top left figure). The question is, what is the best scheme if we are allowed to take only a single linear measurement? This is "literally" a compressed sensing. Although the existing theories of CS, as well as this study, focus only on sparse signals in an asymptotically large dimension, there is no conceptual reason we should restrict the scope of CS to such a specific class of signals.

A key role is played by the nonlinearity of the reconstruction. It is well known that the PCA projection gives the optimal reconstruction, in terms of MSE, provided that recovery is linear. But if recovery is allowed to be nonlinear, the optimal projection may significantly differ from the PCA projection. As shown in Figure 2.2, the nonlinearity adapts the reconstruction to the signal density (read our explanations in the figure caption), making

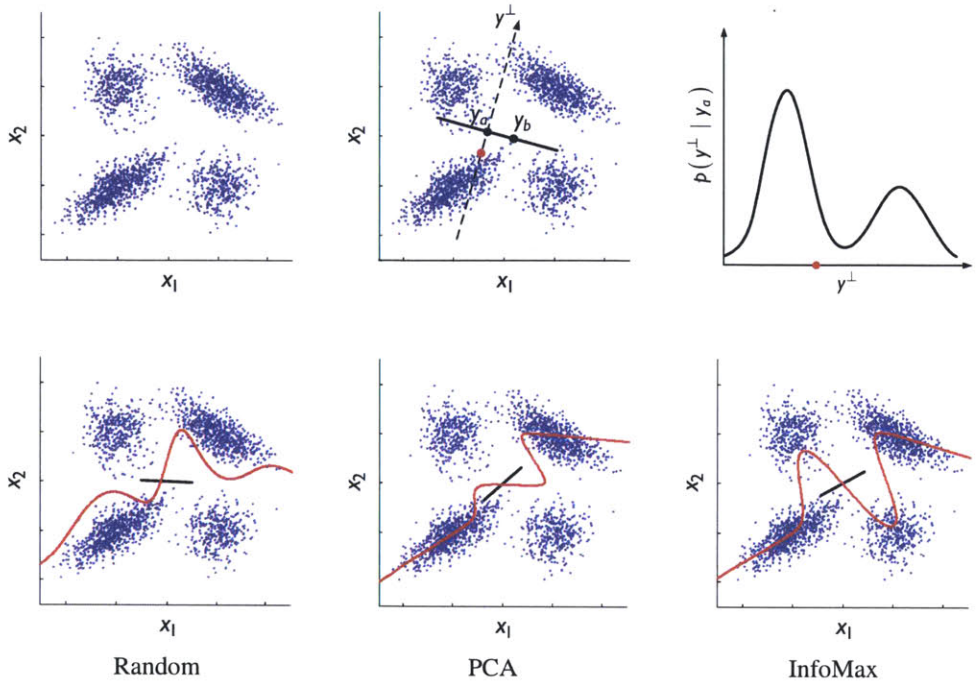


Figure 2.2: Compressed sensing example on the input signal which has a mixture density of four Gaussians $(n, m) = (2, 1)$. Top left: i.i.d. samples (blue dots) from the underlying probability density. Note *non*-sparseness, in any basis, of the signal. Top middle: measurement and reconstruction example. Assume that the measurement gave us a value $\boldsymbol{y}_a$ by projecting the input signal $\boldsymbol{x}$ onto the row vector $\boldsymbol{W}$ (solid black line). Any feasible input compatible with the measurement must lie in the dashed black line, a specific point (red dot) of which can be selected as the reconstruction. Top right: probability density, along the dashed black line, of the signal $\boldsymbol{y}^\perp$ as a mixture of four Gaussians (some clusters may have little effect on this conditional density). Given the density, we may reasonably use the mean (red dot) for the reconstruction. As is well-known, this is the MMSE estimate that minimizes the expected ℓ_2-error. For a different input signal, the measurement could give us another value (e.g., like $\boldsymbol{y}_b$). In the same way as above, we can determine the reconstruction point for every possible measurement value, which will form a continuous curve of red dots as illustrated in the figures on the bottom. Bottom: Different kinds of measurements (corresponding projection vector shown as a solid black line in each case) and the MMSE reconstruction points (red curve). In this toy example, $\boldsymbol{W}$ can be written as $\boldsymbol{W} = (\cos\theta, \sin\theta)$ for some $\theta \in [0, \pi)$. Here, the random measurement is a particular instantiation with random θ drawn from a uniform distribution, while the InfoMax measurement has been exhaustively searched for, by numerically evaluating $h(\boldsymbol{W}\boldsymbol{x})$ for fine-scale discretized θ in $[0, \pi)$. Generally, a better-fit, of the reconstruction, to the original datapoints means a better measurement scheme. In the figure, we can say that InfoMax $>$ PCA $>$ Random (particular instantiation) in terms of goodness of the measurement scheme for this non-sparse signal.

the best effort to fit the original data. The achievable reconstruction performance is inherently bounded to the measurement scheme used. As seen in the figures on the bottom, a random projection may be a bad choice for this type of signal; the PCA projection works somewhat better but is significantly outperformed by the InfoMax projection. The values of $h(\boldsymbol{W}\boldsymbol{x})$ for the three measurement schemes are numerically evaluated and compared in Table 2.1. Indeed, for this signal, the InfoMax scheme turns out to also minimize the MSE.

Table 2.1: Performance comparison among three measurement schemes in example of Figure 2.2, in terms of entropy and MMSE. **cf.** If recovery were restricted to be linear, the smallest MSE would be achieved by the PCA projection. See the last row.

	Random	PCA	InfoMax
$h(\boldsymbol{W}\boldsymbol{x})$	1.189	1.345	**1.449**
MMSE	0.931	0.726	**0.476**
Linear MMSE	1.034	**0.798**	0.818

The signal we have considered is obviously non-sparse in any basis and thus is beyond the scope of the RIP-based theory. Nonetheless, it is clear, from the demonstrated results, that CS still makes sense even for the non-sparse signal. The InfoMax criterion is quite universal; it is conceptually applicable, setting aside technical difficulties in the optimization, wherever the information about the signal is given as a proper form of the prior probability density. Furthermore, the InfoMax projection often minimizes the MMSE criterion, as is true in this toy example.

2.3 Analysis

We are now ready to state the linear InfoMax problem formally, that is,

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}_{m\times n}:\boldsymbol{W}\boldsymbol{W}^T=\boldsymbol{I}}{\arg\max} \; h(\boldsymbol{W}\boldsymbol{x}). \tag{2.3}$$

The solution obviously depends on the prior density $p(\boldsymbol{x})$ which, we assume, is known. In this chapter, we will focus on the inputs that have a *sparse* representation in some orthonor-

mal basis $\boldsymbol{D}$. The sparsification in an orthonormal basis is usually conducted by making the coordinates maximally independent. We assume that the coordinates are indeed independent, which is a common practice to make things easier [114, 14]. For convenience, hereafter, let us write an arbitrary vector in $\mathbb{R}^n$ (e.g., the input $\boldsymbol{x}$, row vectors of $\boldsymbol{W}$) with reference to the basis $\boldsymbol{D}$. Then, $\boldsymbol{x} = (x_1, \ldots, x_n)$ is a vector of independent random variables. Without loss of generality, we assume that $\boldsymbol{x}$ has zero-mean (due to the mean-shift invariant property of the entropy) and the covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ with $\sigma_1 \geq \cdots \geq \sigma_n > 0$. We will denote a variance-normalized (or "whitened") version of $\boldsymbol{x}$ by $\overline{\boldsymbol{x}} \triangleq \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}$, i.e., $\overline{x}_i = x_i/\sigma_i$ for $i = 1, \ldots, n$. For the clarity of exposition, we further assume that $\overline{x}_i$'s are identically distributed (and thus i.i.d. together with the previous independence assumption) according to a symmetric density with negentropy J_x. The negentropy, sometimes called syntropy, is mathematically the Kullback-Leibler (KL) divergence between the true density and a Gaussian with the same first- and second-order statistics.

Specifically, the Gaussian case (i.e., $J_x = 0$ in our input model) was analyzed by Linsker.

Observation 2.3 (Linsker [94])**.** If the input $\boldsymbol{x}$ is jointly Gaussian, then the solution to the linear InfoMax problem (Equation 2.3) is given by the m principal components of $\boldsymbol{x}$.

Proof. Since $\boldsymbol{x}$ is jointly Gaussian, $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ will also be Gaussian so that maximizing the entropy of $\boldsymbol{y}$ is equivalent to maximizing the determinant of the covariance of $\boldsymbol{y}$. This determinant is maximized by the principal components. □

Even if the input is non-Gaussian, the entropy of $\boldsymbol{y}$ generally has a strong dependence on the log determinant of the covariance $Cov(\boldsymbol{y}) = \boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T$: by the formula in Lemma 2.12, $h(\boldsymbol{y}) = \frac{1}{2}\log\det Cov(\boldsymbol{y}) + h(\overline{\boldsymbol{y}})$ where $\overline{\boldsymbol{y}}$ is a whitened version of $\boldsymbol{y}$, i.e., a linear transformation of $\boldsymbol{y}$ whose covariance matrix is the identity.

2.3.1 I.I.D. case

For the i.i.d. case (i.e., $\boldsymbol{\Sigma} = \sigma_x^2\boldsymbol{I}$, for some σ_x, in our input model), the log determinant plays no discriminative role since $Cov(\boldsymbol{y}) = \sigma_x^2\boldsymbol{I}$ for any $m \times n$ tight frame $\boldsymbol{W}$. For this

class of inputs and for a small number of projections, we can prove that random projections are asymptotically InfoMax optimal.

Observation 2.4. For the input $\boldsymbol{x} = (x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are i.i.d., and for $m < O(\sqrt{n})$, let $\boldsymbol{W}$ be the normalization (i.e. $\boldsymbol{W} = (\boldsymbol{H}\boldsymbol{H}^T)^{-1/2}\boldsymbol{H}$) of an $m \times n$ random matrix $\boldsymbol{H}$ with i.i.d. entries sampled from $\mathcal{N}(0, 1/n)$. As $n \to \infty$, $\boldsymbol{W}$ is the solution to the linear InfoMax problem (Equation 2.3).

The proof is based on the following two lemmas:

Lemma 2.5. For white input $\boldsymbol{x}$, if there exists a matrix $\boldsymbol{W}$ such that $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ is jointly Gaussian, then $\boldsymbol{W}$ is the solution to the linear InfoMax problem (Equation 2.3).

Proof. This follows from the fact that the maximum entropy density subject to $Cov(\boldsymbol{y}) = \sigma_x^2 \boldsymbol{I}$ is a multidimensional Gaussian (e.g., see [43]). □

The preceding lemma tells us that, for white input, all we need to solve the InfoMax problem is to find a linear projection that makes the projection Gaussian. If we seek a single projection, we can simply set $\boldsymbol{W} \propto (1, 1, 1, \ldots, 1)$ so that $\boldsymbol{y}$ is the sum of all x_i's and by the central limit theorem, $\boldsymbol{y}$ becomes Gaussian as $n \to \infty$. But when we search for an m-dimensional projection, it is not sufficient that each marginal is Gaussian, but rather we need the vector $\boldsymbol{y}$ to be jointly Gaussian.

Lemma 2.6. For the input $\boldsymbol{x} = (x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are i.i.d., and for $m < O(\sqrt{n})$, let $\boldsymbol{W}$ be the normalization (i.e. $\boldsymbol{W} = (\boldsymbol{H}\boldsymbol{H}^T)^{-1/2}\boldsymbol{H}$) of an $m \times n$ random matrix $\boldsymbol{H}$ with i.i.d. entries sampled from $\mathcal{N}(0, 1/n)$. As $n \to \infty$, $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ almost surely approaches a multidimensional Gaussian.

Proof. See Section 2.6.1.3. □

The previous lemma requires that the number of projections grow slower than the square root of the input dimension. Only for such a "small" number of projections, can we prove that the projection $\boldsymbol{y}$ has Gaussianity. Indeed, for a larger number, a Gaussian projection may not exist unless the input itself is Gaussian (e.g., consider the extremal case $m = n$, where the projection is simply a rotation of the input). Thus proving the InfoMax optimality

using exact Gaussianity requires a small number of projections. If we wish to solve the InfoMax problem for arbitrary m, we need another proof technique. Here we use a central limit behavior approximation based on Jones and Sibson's work [84].

Proposition 2.7 (Central limit behavior approximation, based on [84])**.** Consider the input $\boldsymbol{x} = (x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are i.i.d., according to a zero-mean symmetric density with variance σ_x^2 and negentropy J_x. Given an $m \times n$ tight frame $\boldsymbol{W}$, the negentropy of the multiplexed output $\boldsymbol{W}\boldsymbol{x}$ may be approximated by $J(\boldsymbol{W}\boldsymbol{x}) \approx J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4$, where $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ denote the column vectors of $\boldsymbol{W}$. In terms of entropy, this corresponds to $h(\boldsymbol{W}\boldsymbol{x}) \approx \frac{m}{2} \log(2\pi e \sigma_x^2) - J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4 \triangleq \widehat{h}(\boldsymbol{W}\boldsymbol{x})$.

Proof. See Section 2.6.1.4. □

The negentropy $J(\boldsymbol{W}\boldsymbol{x})$ is a non-Gaussianity measure for the multiplexed output $\boldsymbol{W}\boldsymbol{x}$ in its central limit behavior. The formula $J(\boldsymbol{W}\boldsymbol{x}) \approx J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4$ becomes exact for two extremal settings (i.e., $m < O(\sqrt{n})$ or $m = n$), while it is an approximation in the middle range of m. Derived from Gram-Charlier Type A series [13] in the vicinity of Gaussian, the formula is quite accurate near $J_x \approx 0$. Given the approximation, we need to minimize $\nu(\boldsymbol{W}) \triangleq \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4$ for the InfoMax optimality.

Lemma 2.8. For an $m \times n$ tight frame $\boldsymbol{W}$, the columns of which are denoted by $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$, $\nu(\boldsymbol{W}) = \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4$ is lower-bounded by $\frac{m^4}{n^3} + \frac{m^2(n-m)^2}{n^3(n-1)}$. This bound is achieved if and only if

$$|\boldsymbol{w}_i^T \boldsymbol{w}_j| = \begin{cases} m/n, & \text{if } i = j \\ \sqrt{\frac{m(n-m)}{n^2(n-1)}}, & \text{otherwise.} \end{cases} \tag{2.4}$$

Proof. See Section 2.6.1.5. □

Lemma 2.8 suggests the InfoMax optimality of an equiangular tight frame, if one exists, where each column vector lies on a common spherical surface and forms an equal angle with any other column vectors. An equiangular tight frame is also known to minimize the

mutual coherence [150, 62], defined by

$$\mu(\boldsymbol{W}) \triangleq \max_{i \neq j} \frac{|\boldsymbol{w}_i^T \boldsymbol{w}_j|}{\|\boldsymbol{w}_i\| \cdot \|\boldsymbol{w}_j\|},$$

which suggests that the quantity $\nu(\boldsymbol{W})$ coming up with the (approximate) InfoMax criterion has a strong inclination towards incoherent measurements. This is an interesting coincidence (harmony) with the classical context of CS [28, 57]. The existence of such an equiangular tight frame depends on the dimensions m and n (e.g., see [132]). However, if $m, n \to \infty$, the normalization of an $m \times n$ random Gaussian matrix tends to approach an equiangular tight frame and thus becomes a minimizer of the approximated entropy.

Observation 2.9. For the input $\boldsymbol{x} = (x_1, \ldots, x_n)$, where $x_1, \ldots, x_n$ are i.i.d., let $\boldsymbol{W}$ be a tight frame obtained by normalizing an $m \times n$ random matrix $\boldsymbol{H}$ with i.i.d. entries sampled from $\mathcal{N}(0, 1/n)$. As $m, n \to \infty$, with $m/n \to \beta < 1$, $\boldsymbol{W}$ is the solution to the linear InfoMax problem (Equation 2.3) with the entropy approximation $\widehat{h}$ (Proposition 2.7). Further,

$$\widehat{h}(\boldsymbol{W}\boldsymbol{x}) = \frac{m}{2} \log(2\pi e \sigma_x^2) - m J_x \beta^3. \tag{2.5}$$

As a remark, Equation 2.5 holds also for finite m; if $m \leq O(\sqrt{n})$ and thus if $\beta = 0$, Equation 2.5 gives $\widehat{h}(\boldsymbol{W}\boldsymbol{x}) = \frac{m}{2} \log(2\pi e \sigma_x^2)$, which is exact by Lemma 2.6.

Proof. See Section 2.6.1.6. □

We have seen, for i.i.d. inputs, that the InfoMax optimality requires the projection to be as Gaussian as possible and that random projections are InfoMax optimal, as $n \to \infty$ (see Observations 2.4, 2.9). This result is analogous to the classical theory of CS where random matrices are shown to asymptotically satisfy the RIP with high probability [31]. However, the subsequent story regarding recovery (or inference) is a little different: our result says nothing about the faithful recovery by ℓ_1-regularized least squares or orthogonal matching pursuit, but guarantees a minimal uncertainty for the unobserved part of the signal, for any m and for any J_x (in contrast to the classical CS which is applicable only for sufficiently large m and sufficiently large J_x; refer to Section 2.4 also).

As will be shown shortly, InfoMax makes a substantial departure from classical CS for non-i.i.d. inputs: If the input is not i.i.d., random projections can be far from being optimal.

2.3.2 Non-I.I.D. case

In the previous cases (Gaussian or i.i.d.), either $h(\overline{\boldsymbol{y}})$ or $\frac{1}{2}\log\det Cov(\boldsymbol{y})$ was constant, and the other non-constant term was maximized by PCA or random projections, respectively. Generally, if the input is neither i.i.d. nor Gaussian, there is a trade-off between the two terms. We must accomplish a good balance between PCA and random projections.

To remind readers, the input $\boldsymbol{x} = (x_1, \ldots, x_n)$, written in the sparse basis $\boldsymbol{D}$, has the covariance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ with $\sigma_1 \geq \cdots \geq \sigma_n > 0$. We denote $\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{x}$ by $\overline{\boldsymbol{x}}$; we assume that $\overline{x}_i$'s are i.i.d., according to a symmetric density with negentropy J_x. We want to find an $m \times n$ tight frame that maximizes $h(\boldsymbol{W}\boldsymbol{x})$.

Here, we define $\boldsymbol{Q} \triangleq (\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{\Sigma}^{\frac{1}{2}}$, which helps us to keep the subsequent mathematics clean. The matrix $\boldsymbol{Q}$ is an $m \times n$ tight frame, i.e., satisfies $\boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I}$. Conversely, given $\boldsymbol{Q}$, we can also determine $\boldsymbol{W}$ by $\boldsymbol{W} = (\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ utilizing that $\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T = (\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-1}$. Then, we can write $h(\boldsymbol{W}\boldsymbol{x})$ in terms of $\boldsymbol{Q}$, as

$$\begin{aligned} h(\boldsymbol{W}\boldsymbol{x}) &= h(\boldsymbol{W}\boldsymbol{\Sigma}^{\frac{1}{2}}\overline{\boldsymbol{x}}) \\ &= \frac{1}{2}\log\det(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T) + h((\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{\Sigma}^{\frac{1}{2}}\overline{\boldsymbol{x}}) \\ &= -\frac{1}{2}\log\det(\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T) + h(\boldsymbol{Q}\overline{\boldsymbol{x}}). \end{aligned}$$

Further, we can use Proposition 2.7 to approximate $h(\boldsymbol{Q}\overline{\boldsymbol{x}})$ by $\widehat{h}(\boldsymbol{Q}\overline{\boldsymbol{x}}) = \frac{m}{2}\log(2\pi e) - J_x\nu(\boldsymbol{Q})$ since both $\boldsymbol{Q}$ and $\overline{\boldsymbol{x}}$ satisfy the premise condition. Finally, therefore,

$$\widehat{h}(\boldsymbol{W}\boldsymbol{x}) = -\frac{1}{2}\log\det(\sum_i \frac{1}{\sigma_i^2}\boldsymbol{q}_i\boldsymbol{q}_i^T) + \frac{m}{2}\log(2\pi e) - J_x\sum_{i,j}(\boldsymbol{q}_i^T\boldsymbol{q}_j)^4, \qquad (2.6)$$

where $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ denote the column vectors of $\boldsymbol{Q}$.

While it is very difficult to directly compute an $m \times n$ tight frame $\boldsymbol{Q}$ that maximizes Equation (2.6), we can find an upper-bound of the approximated entropy, as described in

the following lemma:

Lemma 2.10. The approximated entropy $\widehat{h}$ in Equation (2.6) is upper-bounded by $\widehat{h}^*$ which is defined as

$$\widehat{h}^*(\sigma_i^2, J_x) \triangleq \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_i \varepsilon_i \log \sigma_i^2 - J_x \sum_i \varepsilon_i^4, \tag{2.7}$$

where

$$\varepsilon_i \triangleq \text{median}\left(\sqrt[3]{\frac{\log \sigma_i^2 + \xi}{8J_x}},\ 0,\ 1\right), \quad i = 1, \ldots, n, \tag{2.8}$$

with a Lagrange multiplier ξ chosen to satisfy $\sum_i \varepsilon_i = m$. If a matrix $\boldsymbol{Q}$ achieves $\widehat{h}^*$ in Equation (2.6), the squared ℓ_2-norm (or power) of each column vector of $\boldsymbol{Q}$ should be equal to $\{\varepsilon_i\}$.

Proof. See Section 2.6.1.7. □

Lemma 2.10 implies that an $m \times n$ tight frame $\boldsymbol{W}$ which makes $\widehat{h}(\boldsymbol{W}\boldsymbol{x}) = \widehat{h}^*$, if any, will be InfoMax optimal (under the central limit behavior approximation). Further, the matrix $\boldsymbol{Q} = (\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-1/2}\boldsymbol{W}\boldsymbol{\Sigma}^{1/2}$ should have the power distribution $\|\boldsymbol{q}_i\|^2 = \varepsilon_i$ (in Equation 2.8) for all its columns, which can be described as a non-usual type of water-filling scheme (see Figure 2.3 for illustration). The converse is not always true: in some settings of $\{\sigma_i^2, J_x\}$, the upper-bound $\widehat{h}^*$ may not be achievable with any $m \times n$ tight frame matrix.

In the following, we enumerate a few special examples for which we have successfully found a matrix that achieves $\widehat{h}^*$. Subsequently, we will briefly explain how to compute a *near*-optimal matrix for more general settings.

Example 2.1. With no decay in σ_i (i.e., $\sigma_1 = \cdots = \sigma_n = \sigma_x$), the upper-bound $\widehat{h}^*$ is given by $\widehat{h}^* = \frac{m}{2}\log(2\pi e) + \frac{m}{2}\log\sigma_x^2 - mJ_x(\frac{m}{n})^3$ and, if $n \to \infty$ with $m/n \to \beta < 1$, this is asymptotically achievable by the normalization of an $m \times n$ random Gaussian matrix. This is actually a restatement of Observation 2.9, but here based on the achievability of $\widehat{h}^*$.

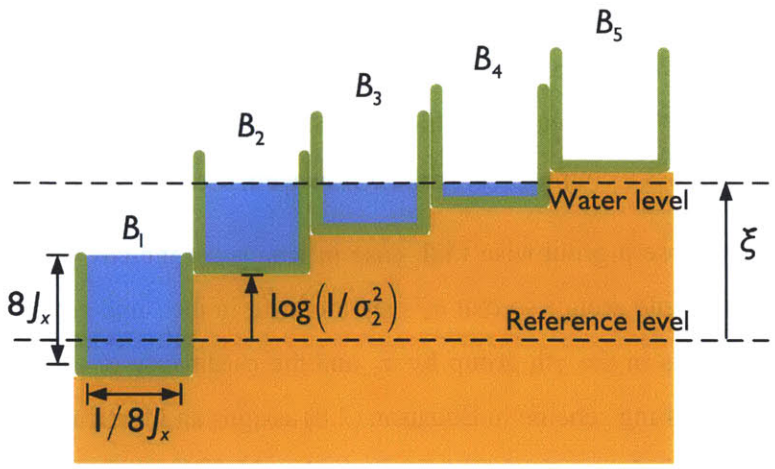


Figure 2.3: Illustration of the water-filling implied by Equation (2.8). Imagine n buckets B_i (bottom area $1/8J_x \times$ height $8J_x$) placed on the ground level $\log(1/\sigma_i^2)$ above a reference. Water is poured then in a way that all the buckets have the same water level, if any, from the reference. In some buckets, water may overflow, due the limited height of the buckets. Since ground levels are uneven, the volume V_i of water in B_i can be different from each other. We measure the cube root $\sqrt[3]{V_i}$ of the water volume in each bucket and stop pouring water as soon as the sum reaches the given budget m. After all, ξ corresponds to the water level from the reference, and ε_i is equal to the cube root of the final volume of water contained in B_i.

Proof. See Section 2.6.1.8. □

Example 2.2. If the decay rate in σ_i is steep, specifically if

$$\sigma_m/\sigma_{m+1} \geq e^{4J_x}, \tag{2.9}$$

the upper-bound $\widehat{h}^*$ is given by $\widehat{h}^* = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^{m}\log\sigma_i^2 - mJ_x$ and this is achievable by the PCA projection. Note that this sufficient condition strongly depends on the negentropy J_x. For example, if $J_x = 0$, the condition holds for whatever decay rate, simply reasserting the InfoMax optimality of the PCA projection for Gaussian. When $J_x > 0$, the condition (2.9) is satisfied, for example, if σ_i's are distributed as

i. $\sigma_i = e^{-\gamma i}$ with $\gamma \geq 4J_x$ (exponential decay);

ii. $\sigma_i = (\frac{1}{i})^\gamma$ with $\gamma \geq \frac{4J_x}{\log 2}$ (power-law decay) if $m \leq \frac{1}{e^{4J_x/\gamma}-1}$.

Proof. See Section 2.6.1.9. □

In the previous two examples, the decay rate in σ_i was either very low or sufficiently high, and random or PCA projections achieved $\widehat{h}^*$ (and thus were InfoMax optimal) in each case. Below we discuss a couple of hybrid cases as well.

Example 2.3. Suppose a groupwise i.i.d. case in which the indices $\{1, 2, \ldots, n\}$ can be partitioned into multiple groups so that σ_i's are identical in the same group. We denote the identical value of σ_i's in the jth group by $\widetilde{\sigma}_j$ and the cardinality of the jth group by n_j. Note that the water-filling scheme in Equation (2.8) assigns an identical value ε_i for all i in the same group. Let $\widetilde{\varepsilon}_j$ denote the ε_i value assigned to the jth group.

Here is our claim: In the asymptotic case where $n_j \to \infty$ for all j, the following groupwise measurement scheme asymptotically achieves $\widehat{h}^*$: To measure each group, we use a tight frame obtained by normalizing an $m_j \times n_j$ random Gaussian matrix, with $m_j = n_j\widetilde{\varepsilon}_j$.

Proof. See Section 2.6.1.10 □

A roughly groupwise i.i.d. pattern appears, for example, in natural images [36] if the wavelet coefficients are grouped by the octave frequency band. Therefore, the nonuniform-density bandwise random measurement, described above, provides an InfoMax optimal scheme for compressed sensing of natural images. Regarding the nonuniform-density sampling, our result gives more analytical reasoning to previous work [98], as well as a principled way (i.e., water-filling) of determining the density. Note that the density should depend on the total number of projections as well as the input statistics (σ_i^2, J_x).

Example 2.4. For some $m_0, m_1 \in \{0\} \cup \mathbb{Z}^+$, suppose that $\frac{\sigma_{m-m_0}}{\sigma_{m+m_1+1}} \geq e^{4J_x}$ and that $\frac{\sigma_{m-m_0+1}}{\sigma_{m+m_1+1}} = \cdots = \frac{\sigma_{m+m_1}}{\sigma_{m+m_1+1}} = e^{4J_x\left(\frac{m_0}{m_0+m_1}\right)^3}$. In this case, the upper-bound $\widehat{h}^*$ is given by $\widehat{h}^* = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^{m}\log\sigma_i^2 - mJ_x + m_0J_x\left(1 - (\frac{m_0}{m_0+m_1})^3\right)$ and this is achievable by the normalization of $(m - m_0)$ PCA plus m_0 random projections over the next $(m_0 + m_1)$ PCA coefficients, if $(m_0 + m_1)$ is asymptotically large.

Proof. See Section 2.6.1.11. □

While the signal setting in Example 2.4 looks very special, it can be considered as an approximation of a bit more general signals which are nonwhite as a whole but remarkably

white within some large-dimensional subspace around the mth coordinate. If we denote, by $\mathcal{P}_m$, the set of all the signal distributions supposed in Example 2.4, we may be able to find $p' \in \mathcal{P}_m$ that is close to the true density p in a proper distance measure $d(\cdot, \cdot)$. See Figure 2.4 for example. For the approximated density p', $(m - m_0)$ PCA plus m_0 random projections over the next $(m_0 + m_1)$ PCA coefficients are asymptotically InfoMax optimal. For the original density p, this set of projections may be said to be "nearly" InfoMax optimal if $d(p', p)$ is sufficiently small. The joint use of the PCA and random projections have

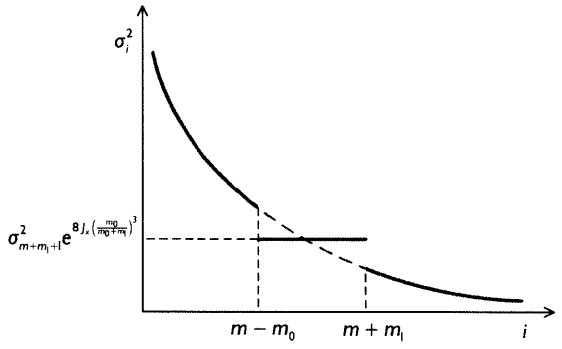


Figure 2.4: A probability density approximation example. Shown here are the variance profiles: original σ_i (dashed) vs. approximation σ_i' (solid). For $m - m_0 < i \leq m + m_1$, $\sigma_i' = \sigma_{m+m_1+1} e^{4J_x(\frac{m_0}{m_0+m_1})^3}$, while it remains the same as σ_i elsewhere. The negentropy J_x remains unchanged.

recently been proposed for CS [115, 121, 129], but mostly based on heuristics. Our result gives the InfoMax principled reasoning to such ideas. Moreover, it tells a novel finding which most other methods had previously failed to predict: we need to take the random part of measurements over rather a restricted PCA subspace, not over the entire Euclidean space, which is often more important than to include a small number of PCA projections.

Let us more formally present the near-optimality of $\boldsymbol{W}'$ on the true density p, where $\boldsymbol{W}'$ is an $m \times n$ tight frame such that $\widehat{h}(\boldsymbol{W}'\boldsymbol{x}') = \widehat{h}^*(p')$ for $\boldsymbol{x}' \sim p' \in \mathcal{P}_m$. For the best approximation p', we may use the I-projection [44], i.e., $p' = \arg\min_{p'' \in \mathcal{P}_m} D_{\mathrm{KL}}(p''\|p)$, with the KL divergence $D_{\mathrm{KL}}(\cdot\|\cdot)$ for the distance measure $d(\cdot, \cdot)$. If $D_{\mathrm{KL}}(p'\|p) \leq \epsilon$, then $|\widehat{h}(\boldsymbol{W}'\boldsymbol{x}) - \widehat{h}(\boldsymbol{W}'\boldsymbol{x}')|$ and $|\widehat{h}^*(p') - \widehat{h}^*(p)|$ can be upper-bounded by $\theta_1(\epsilon)$ and $\theta_2(\epsilon)$, respectively, where $\theta_1(\cdot)$ and $\theta_2(\cdot)$ are some deterministic functions based on the definition of $\widehat{h}$ and $\widehat{h}^*$. Both $\theta_1(\epsilon)$ and $\theta_2(\epsilon)$ yield zero at $\epsilon = 0$ and monotonically increase with ϵ.

By triangle inequality,

$$|\widehat{h}(\boldsymbol{W}'\boldsymbol{x}) - \widehat{h}^*(p)| \leq |\widehat{h}(\boldsymbol{W}'\boldsymbol{x}) - \widehat{h}(\boldsymbol{W}'\boldsymbol{x}')| + |\widehat{h}(\boldsymbol{W}'\boldsymbol{x}') - \widehat{h}^*(p)|$$
$$= \theta_1(\epsilon) + \theta_2(\epsilon). \quad (2.10)$$

If ϵ is sufficiently small, $\boldsymbol{W}'$ makes $\widehat{h}(\boldsymbol{W}'\boldsymbol{x})$ close to the upper-bound $\widehat{h}^*$ with the true density p and thus is nearly InfoMax optimal.

We pose a question here. What if $D_{\text{KL}}(p'\|p) > \epsilon$? Do there exist any other "near"-optimal measurement schemes also applicable to the signals that are not ϵ-approximable to $\mathcal{P}_m$? We will present one in the following. We discard the approximability assumption but instead suppose a weaker assumption, that $\widehat{h}^*$ is nearly achievable. Then, as argued in Lemma 2.10, the matrix $\boldsymbol{Q} = (\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{\Sigma}^{\frac{1}{2}}$ should have the ℓ_2-norm of each column $\boldsymbol{q}_i$ almost equal to ε_i produced by the water-filling scheme (in Equation 2.8). In its definition, the matrix $\boldsymbol{Q}$ should also be a tight frame; therefore, we enforce the matrix $\boldsymbol{Q}$ to satisfy both conditions. Let $\mathcal{Q}_{\boldsymbol{\varepsilon}}$ denote the set of all $m \times n$ matrices that satisfy the two conditions. While the matrices in $\mathcal{Q}_{\boldsymbol{\varepsilon}}$ may not be unique modulo a rotation, all of them have a common property (based on the following lemma): if we define $S_0 \triangleq \{i : \varepsilon_i = 0\}$ and $S_1 \triangleq \{i : \varepsilon_i = 1\}$, every matrix in $\mathcal{Q}_{\boldsymbol{\varepsilon}}$ makes $\boldsymbol{W} = (\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ include $|S_1|$ major PCA projections while being orthogonal to the subspace spanned by the $|S_0|$ minor principal components.

Lemma 2.11. Let $\boldsymbol{Q}$ be an $m \times n$ tight frame, the column vectors of which are $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$. Define $S_0 \triangleq \{i : \|\boldsymbol{q}_i\| = 0\}$ and $S_1 \triangleq \{i : \|\boldsymbol{q}_i\| = 1\}$. We consider another $m \times n$ tight frame $\boldsymbol{W} \triangleq (\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Sigma}^{-\frac{1}{2}}$ where $\boldsymbol{\Sigma}$ is any $n \times n$ positive definite diagonal matrix. If we denote, by $\boldsymbol{e}_i$, the unit vector containing 1 in the ith element and 0 elsewhere, any vector $\boldsymbol{e}_i$ lies in the rowspace of $\boldsymbol{W}$ if $i \in S_1$ and in the nullspace of $\boldsymbol{W}$ if $i \in S_0$.

Proof. See Section 2.6.1.12. □

The degree of freedom lies only in the remaining subspace, i.e., $S_{(0,1)} \triangleq \{i : 0 < \varepsilon_i < 1\}$, but the multiplexing in $S_{(0,1)}$ is also restrictive: even in $S_{(0,1)}$, all the matrices in $\mathcal{Q}_{\boldsymbol{\varepsilon}}$ follow a common energy distribution (ℓ_2-norm) computed by the water-filling. We suggest

a randomization for the other free factors. Further optimization – rigorously maximizing (2.6) for $\boldsymbol{Q} \in \mathcal{Q}_{\boldsymbol{\varepsilon}}$ – may make a little improvement while it is very difficult to do analytically. The randomization has performed best in a special case with ε_i's being constant for all $i \in S_{(0,1)}$ (see Example 2.4) and also in the most uncertain circumstance where the full exploitation of signal statistics is infeasible (e.g., i.i.d. modeling given an order-statistics).

Specifically, we begin with a random Gaussian matrix for $\boldsymbol{W}$ and convert it to $\boldsymbol{Q}$. We then project $\boldsymbol{Q}$ onto $\mathcal{Q}_{\boldsymbol{\varepsilon}}$ by iterating the following two steps:

1. Reweight the columns of $\boldsymbol{Q}$ so that $\|\boldsymbol{q}_i\|^2 = \varepsilon_i$ for all i.

2. Make $\boldsymbol{Q}$ be a valid tight frame by normalization.

After a small number of iterations, the matrix $\boldsymbol{Q}$ becomes a tight frame as well as satisfies the target column weights. Given $\boldsymbol{Q}$, the measurement matrix $\boldsymbol{W}$ can be reconstructed by $\boldsymbol{W} = (\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-\frac{1}{2}}\boldsymbol{Q}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. We already know that the resulting matrix $\boldsymbol{W}$ will include $|S_1|$ major principal components and will completely exclude the $|S_0|$ minor principal components, so the above algorithm may be run only for the remaining subspace to find an $(m - |S_1|) \times (n - |S_0| - |S_1|)$ matrix.

This latter type of near-optimal scheme endures the broken ϵ-approximability and thus is more general than the former type. As an upper-bound, h^* may still remain tight even if not achievable and even if the signal is not well approximable to $\mathcal{P}_m$. This is often the case in practice, although we have no general proof.

2.4 Numerical Experiments

In Figure 2.1, we considered a couple of signal classes which have the same order statistics. One was a nonwhite Gaussian signal and the other was a set of i.i.d. samples from a non-Gaussian density. The relative performances of the PCA and random projections were completely different for the two. In fact, the results could be predicted by our InfoMax-based theory: The PCA for Gaussian, and random for i.i.d. are actually InfoMax optimal measurements (Lemma 2.3, Observations 2.4 and 2.9).

We want to redraw readers' attention to the fact that ℓ_1-regularization with random projections has only worked when m is sufficiently large. Otherwise, it has performed poorly, even worse than the canonical linear reconstruction. This is related to the claim, in the classical theory of CS, that compressible signals can be *stably* reconstructed from random measurements using ℓ_1-regularization if $m > O(k \log(n/k))$ where k denotes sparsity of the signal. If m is small, such a stability may not be attainable and goodness of random projections is no more ensured. In contrast, our results claim the InfoMax optimality of random projections for any m as long as n is large. To our expectation, we see in Figure 2.1 (bottom left) that random projections consistently enable a better reconstruction than the PCA (or ICA) projection, if decoded by the MMSE estimate rather than by ℓ_1-regularization. We explain, in Section 2.6.3, how we implemented the MMSE estimation.

We consider another i.i.d. example in which x_i has a Laplacian density $p(x_i) \propto e^{-\sqrt{2}|x_i|}$. As shown in Figure 2.5 (left), the order statistics of the i.i.d. Laplacian decay so slowly that

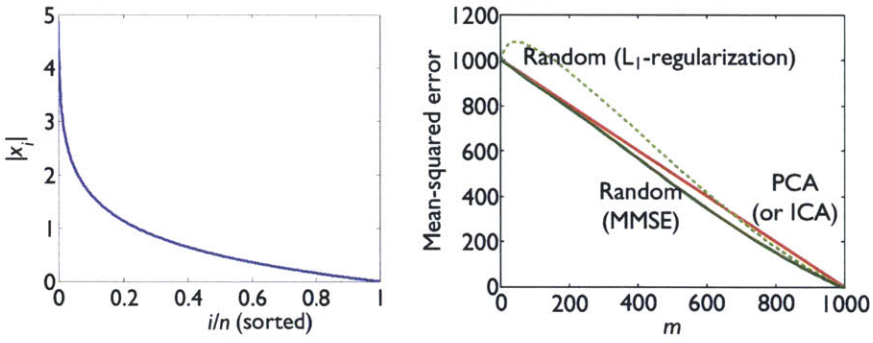


Figure 2.5: Experiment with an i.i.d. Laplacian signal, $p(x_i) \propto e^{-\sqrt{2}|x_i|}$. Left: order statistics. Right: recovery error versus number of measurements for PCA and random projections. For recovery, two alternative methods (MMSE and ℓ_1-regularization) have been employed. When the PCA projections are used, MSEs from the two recovery methods happen to coincide.

it is not classified as a compressible signal (e.g., see [35]), and therefore the classical theory of CS may not apply. Indeed, ℓ_1-regularization with random projections fails in most range of m. On the other hand, the InfoMax optimality does not require such a strict sense of compressibility and makes us expect that random projections are still the best choice for the i.i.d. Laplacian signal. Our expectation proves correct in Figure 2.5. If we employ the

MMSE estimate, random projections are still better than the PCA (or ICA) projection, in the entire range of m, although the margin is not so large.

Next, we suppose two cases of groupwise i.i.d. signals. Both cases are commonly composed of four octave bands, with the size $n_j \propto 2^j$ and the variance $\widetilde{\sigma}_j^2 \propto (1/2)^j$, for $j = 1, \ldots, 4$ (refer to Example 2.3 for the notational definition), but they are discriminated by the negentropy J_x (or by the probability density $p(\overline{x}_i)$). Supposing $p(\overline{x}_i) \propto e^{-C_s|\overline{x}_i|^s}$, where $C_s = (\Gamma(3/s)/\Gamma(1/s))^{s/2}$ to satisfy the unit variance (see Appendix A), we consider $s = 0.5$ for one case and $s = 0.7$ for the other case.

As shown in Figure 2.6, for $s = 0.5$, random projections perform better than the PCA projection in a wider range of m, while for $s = 0.7$, the reverse is true. For the groupwise i.i.d signals, the InfoMax projections are the nonuniform-density groupwise random projections. The number of projections per group is specifically determined by the water-filling. The InfoMax projections are adaptive to s because the underlying water-filling depends on the negentropy J_x (see Equation 2.8). In our example, J_x is related to s, by $J_x = \frac{1}{2}[\log(\pi s^2 \Gamma(3/s)) - \log(2\Gamma^3(1/s)) + 1 - 2/s]$ (cf. Equation A.2). In the figure, the InfoMax projection consistently achieves a lower-bound of the PCA and random projections in terms of MSE. As we mentioned in Section 2.2.1, there is no theoretical guarantee that InfoMax always leads to the smallest MSE. At the bottom of Figure 2.6, we have plotted $\widehat{h}(\boldsymbol{W}\boldsymbol{x})$ (Equation 2.6) for each measurement scheme. The relative performance, predicted from the entropy,[3] among the measurements remarkably coincides with the MMSE recovery performance.

Finally, we also consider more general non-i.i.d., non-Gaussian signals whose marginal density is given by $p(x_i) \propto e^{-\sqrt[4]{120}|x_i/\sigma_i|^{0.5}}$. We suppose that the variances $\{\sigma_i^2\}$ follow a power law, i.e., $\sigma_i \propto (1/i)^\gamma$ for various values of γ. Our theoretical results say that random projections would be asymptotically InfoMax optimal if $\gamma = 0$, while the PCA projection would be InfoMax optimal if γ is sufficiently large. In-between, we have argued that a randomized choice in $\mathcal{Q}_{\boldsymbol{\varepsilon}}$ would be nearly InfoMax optimal, with the weights $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ determined by a water-filling scheme depending on the variance σ_i^2 and the

[3]The differential entropy only matters in a *relative* sense between two equal-number projections. There is no meaningful implication whether it increases or decreases along with the number of projections.

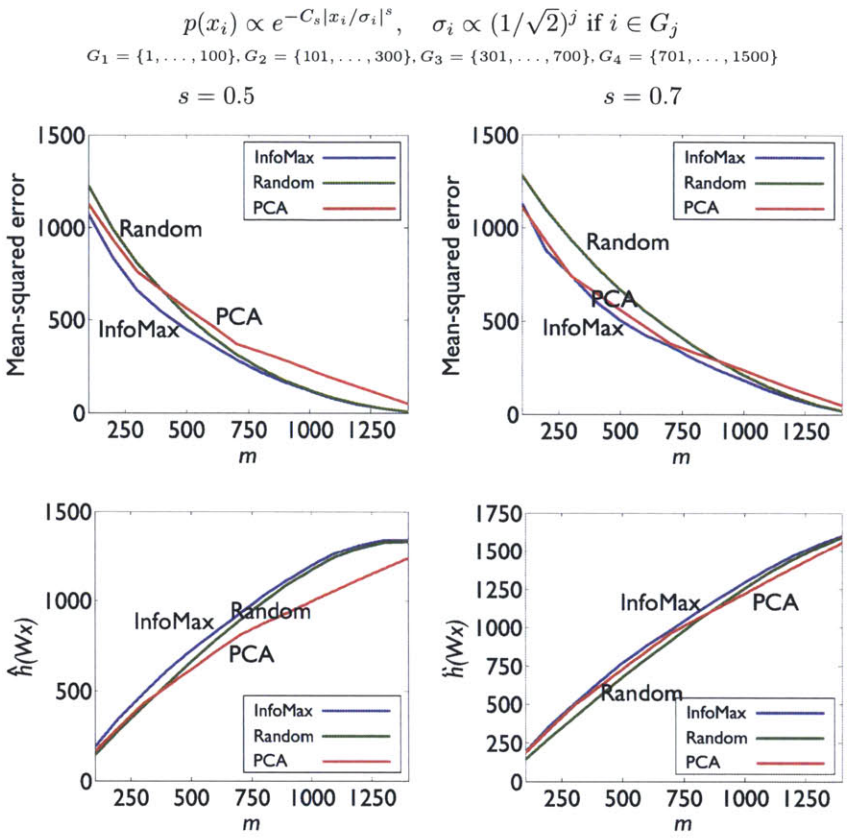


Figure 2.6: Experiments with groupwise i.i.d. signals. We suppose four groups $G_1 = \{1,\ldots,100\}$, $G_2 = \{101,\ldots,300\}$, $G_3 = \{301,\ldots,700\}$, $G_4 = \{701,\ldots,1500\}$ (the total dimension of the signals is 1,500). The marginal density has the following form: $p(x_i) \propto e^{-C_s|x_i/\sigma_i|^s}$, where $\sigma_i \propto (1/\sqrt{2})^j$ if $i \in G_j$. For s, which controls the non-Gaussianity J_x of the input distribution, we consider two values: $s = 0.5$ (left) and $s = 0.7$ (right). Top: MMSE recovery performance of PCA, random, and InfoMax-based projections. Bottom: plot of $\widehat{h}(\boldsymbol{W}\boldsymbol{x})$, based on Equation (2.6), for PCA, random, and InfoMax-based projections. Here, the InfoMax-based projections are the nonuniform-density groupwise random projections with the density regulated by the water-filling (see Example 2.3 in the previous section).

negentropy J_x of the signal. We employ the MMSE estimate, for recovery, to see the best possible reconstruction performance of each measurement scheme.

As shown in Figure 2.7, for $\gamma = 0.25$ (when non-Gaussianity is a dominant factor of sparsity), random projections perform better than the PCA projection in a wide range of m, while with small m, the PCA projection outperforms random projections. In this example, we have computed the InfoMax-based projections using the randomized selection in $\mathcal{Q}_\varepsilon$

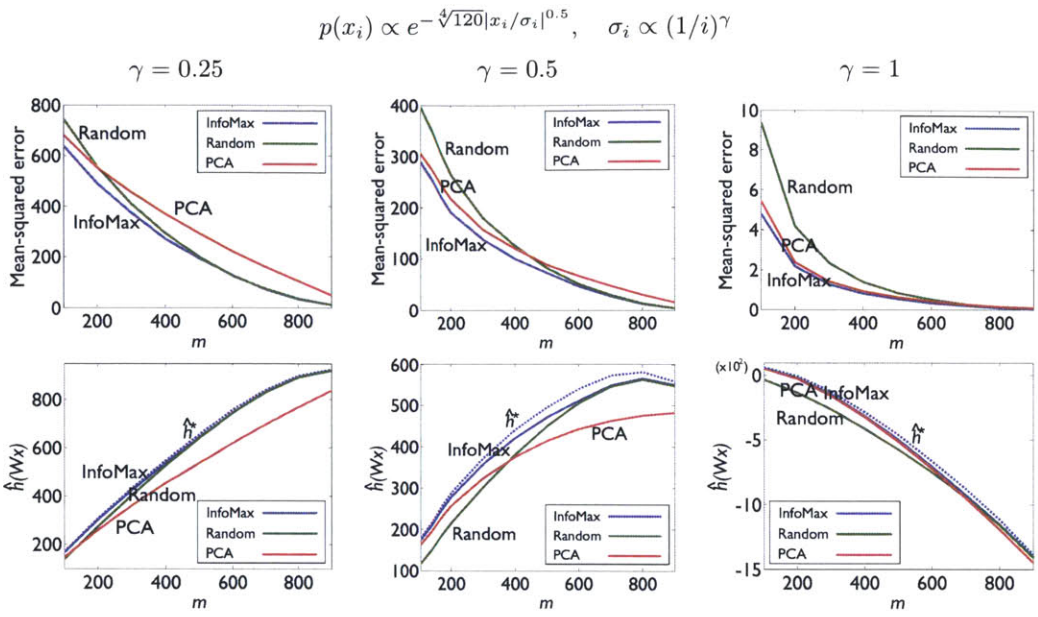


Figure 2.7: Experiments with non-i.i.d., non-Gaussian signals. The supposed marginal density is $p(x_i) \propto e^{-\sqrt[4]{120}|x_i/\sigma_i|^{0.5}}$, $i = 1, \ldots, 1000$, where $\sigma_i \propto (1/i)^{\gamma}$ for $\gamma = 0.25$ (left), $\gamma = 0.5$ (middle), and $\gamma = 1$ (right). Top: MMSE recovery performance of PCA, random, and InfoMax-based projections. Bottom: plot of $\widehat{h}(\boldsymbol{W}\boldsymbol{x})$, based on Equation (2.6), for the PCA, random, and InfoMax-based projections. Here, the InfoMax-based projections have been computed based on the randomized selection in $\mathcal{Q}_{\boldsymbol{\varepsilon}}$ (according to the algorithm given at the end of Section 2.3.2). In the figures on the bottom, the dotted blue curve denotes $\widehat{h}^*$ in the respective setting (not exactly achievable in these examples).

(see the algorithm given at the end of Section 2.3.2). The gap between $\widehat{h}(\boldsymbol{W}\boldsymbol{x})$ and $\widehat{h}^*$ is kept very small in most range of m (in Figure 2.6, bottom left), confirming the *near* InfoMax optimality of the computed set of projections. The InfoMax-based projection achieves a lower-bound of the PCA and random projections in terms of MSE.

With a larger value of γ, variance asymmetry becomes dominant over non-Gaussianity, and the PCA projection outperforms random projections in a wider range of m, but consistently the InfoMax-based projection achieves a lower-bound of the two projections in terms of MSE. Particularly when $\gamma = 0.5$, $\widehat{h}(\boldsymbol{W}\boldsymbol{x})$ may not be considered to be so close to $\widehat{h}^*$. This suggests a possibility that there may exist a slightly better set of projections, in terms of InfoMax, than the one we have used here, although none can exactly achieve $\widehat{h}^*$. It might be possible to find such a set of projections with more elaborate selection of $\boldsymbol{Q}$ in

$\mathcal{Q}_\varepsilon$.

The relative performance (among InfoMax, PCA, and random projections) shown in Figure 2.7 is typical for any other non-i.i.d., non-Gaussian signals, although the performance margin varies case by case.

2.5 Discussion

Suppose we take a small number of linear projections of signals in a dataset, and then use the projections plus our knowledge of the dataset to reconstruct the signals. What are the best projections to use? Given a sparsity assumption of the signals, the theory of CS tells us that random projections can provide a remarkably good reconstruction, even at the sub-Nyquist rate, as long as the number of projections are sufficiently large in dependence of the sparsity. The theory gives a novel sufficient condition for the reconstruction of sparse signals. It is often much tighter than the Nyquist sampling theorem, which is based on a bandlimitedness assumption, of the signals, in Fourier domain. A key to the success is evidently the better knowledge (or model) of the dataset.

If our knowledge of the dataset is even more accurate (e.g., we split the source of the sparsity into two factors, second-order or higher-order statistics of the signals), what can we say beyond the existing theory? In this chapter, we suppose that the knowledge of the dataset is given in the form of a probability density; then we apply the InfoMax principle to find an undercomplete linear measurement that is maximally informative about the unmeasured dimensions of the signals. We have focused our attention on the signals that have a sparse representation in an orthonormal basis, as is common in the literature of CS, and managed to analytically solve the InfoMax problem by exploiting a central limit behavior approximation of the mixture $\boldsymbol{W}\boldsymbol{x}$, on the basis of Jones and Sibson's seminal work [84].

Our findings are summarized as follow: In the settings where the coefficients of the sparsifying basis are i.i.d., the InfoMax principle finds that random projections are asymptotically optimal. On the other hand, in non-i.i.d. settings, InfoMax produces a novel set of projections, which consistently outperform PCA or random projections in signal re-

construction based on the MMSE estimate. In general, the InfoMax optimal projection approximately consists of a certain number of PCA projections plus the remaining number of projections restricted to multiplexing over a particular linear subspace, with every parameter governed by a type of water-filling. During our theoretical development, we also make connections to the existing CS approaches, some of which have remained heuristic; we provide them a common theoretical ground, InfoMax, and a common principled way of optimization, water-filling.

2.6 Appendix to Chapter 2

2.6.1 Proofs

2.6.1.1 Proof of Lemma 2.1

For any orthonormal square matrix $\boldsymbol{R}$, $h(\boldsymbol{R}\boldsymbol{W}\boldsymbol{x}) \overset{(a)}{=} h(\boldsymbol{W}\boldsymbol{x}) + \log|\det(\boldsymbol{R})| \overset{(b)}{=} h(\boldsymbol{W}\boldsymbol{x})$, where the equalities are due to (a) an entropy formula for the invertible transformation (Lemma 2.12) and (b) that the determinant of any orthonormal square matrix is equal to one. If $\boldsymbol{W}_o$ achieves the maximum entropy, so does $\boldsymbol{R}\boldsymbol{W}_o$ with the same power budget $\mathrm{Tr}(\boldsymbol{R}\boldsymbol{W}_o\boldsymbol{W}_o^T\boldsymbol{R}^T) = \mathrm{Tr}(\boldsymbol{R}^T\boldsymbol{R}\boldsymbol{W}_o\boldsymbol{W}_o^T) = \mathrm{Tr}(\boldsymbol{W}_o\boldsymbol{W}_o^T)$.

2.6.1.2 Proof of Lemma 2.2

Suppose any non-tight frame matrix $\boldsymbol{W}_o$ such that $\mathrm{Tr}(\boldsymbol{W}_o\boldsymbol{W}_o^T) = m$. Let $\lambda_i(\cdot)$ denote the eigenvalues of a matrix. By Jensen's inequality,

$$\begin{aligned}\frac{1}{m}\log\det(\boldsymbol{W}_o\boldsymbol{W}_o^T) &= \frac{1}{m}\sum_{i=1}^{m}\log\lambda_i(\boldsymbol{W}_o\boldsymbol{W}_o^T) \\ &\leq \log\frac{1}{m}\sum_{i=1}^{m}\lambda_i(\boldsymbol{W}_o\boldsymbol{W}_o^T) = \log\frac{\mathrm{Tr}(\boldsymbol{W}_o\boldsymbol{W}_o^T)}{m} = 0,\end{aligned}$$

where the equality holds only if $\lambda_i(\boldsymbol{W}_o\boldsymbol{W}_o^T)$'s are all equal – that is to say, $\boldsymbol{W}_o\boldsymbol{W}_o^T = \boldsymbol{I}$, which is not true in our assumption. Strictly, therefore, $\log\det(\boldsymbol{W}_o\boldsymbol{W}_o^T) < 0$. If we let $\boldsymbol{W} = (\boldsymbol{W}_o\boldsymbol{W}_o^T)^{-1/2}\boldsymbol{W}_o$, the matrix $\boldsymbol{W}$ obviously forms a tight frame as well as satisfies

the power budget constraint. Based on the formula in Lemma 2.12, we can write

$$h(\boldsymbol{W}\boldsymbol{x}) = h((\boldsymbol{W}_o\boldsymbol{W}_o^T)^{-1/2}\boldsymbol{W}_o\boldsymbol{x}) = h(\boldsymbol{W}_o\boldsymbol{x}) - \frac{1}{2}\log\det(\boldsymbol{W}_o\boldsymbol{W}_o^T) > h(\boldsymbol{W}_o\boldsymbol{x}),$$

which shows that $\boldsymbol{W}_o$ cannot be optimal.

2.6.1.3 Proof of Lemma 2.6

According to random matrix theory [139], all the eigenvalues of $\boldsymbol{H}\boldsymbol{H}^T$ become 1 for the given condition, i.e., $m < O(\sqrt{n})$ and $n \to \infty$, so $\boldsymbol{W} = \boldsymbol{H}$ and $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$.

On the other hand, Dasgupta et al. [45] have shown that, for $m < O(\sqrt{n})$, $\boldsymbol{H}\boldsymbol{x}$ asymptotically converges to a scale-mixture of zero-mean Gaussians with spherical covariances that have the same profile as the distribution of $\|\boldsymbol{x}\|/\sqrt{n}$. Because $\|\boldsymbol{x}\|/\sqrt{n} \to \sigma_x$ as $n \to \infty$, the mixture density of $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x}$ will collapse to a single Gaussian with zero-mean and covariance $\sigma_x^2\boldsymbol{I}$.

2.6.1.4 Proof of Proposition 2.7

Let us prove the observation specifically assuming that $\sigma_x = 1$ (we will generalize this result later). If we let $\boldsymbol{z} \triangleq \boldsymbol{W}\boldsymbol{x}$, the random vector $\boldsymbol{z}$ has zero for its mean and the identity for its covariance due to the tight frame property of $\boldsymbol{W}$. We can also compute a few of higher-order moments. Define $\kappa_x \triangleq \mathbb{E}[x_i^4]$ and let δ denote the Kronecker delta that gives one if all its subscripts have the same value and zero otherwise. Then, $\mathbb{E}[x_i x_j x_k] = 0$ (by the assumed symmetry of the density function) and $\mathbb{E}[x_i x_j x_k x_l] = \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + (\kappa_x - 3)\delta_{ijkl}$, $\forall i, j, k, l$. Since $z_i = \sum_{i'} W_{ii'}x_{i'}$, we can write

$$\mathbb{E}[z_i z_j z_k] = \sum_{i',j',k'} W_{ii'}W_{jj'}W_{kk'}\underbrace{\mathbb{E}[x_{i'}x_{j'}x_{k'}]}_{=0} = 0, \tag{2.11}$$

and

$$\begin{aligned}\mathbb{E}\left[z_i z_j z_k z_l\right] &= \sum_{i',j',k',l'} W_{ii'} W_{jj'} W_{kk'} W_{ll'} \mathbb{E}\left[x_{i'} x_{j'} x_{k'} x_{l'}\right] \\ &= \sum_{i',j',k',l'} W_{ii'} W_{jj'} W_{kk'} W_{ll'} \Big(\delta_{i'j'}\delta_{k'l'} + \delta_{i'k'}\delta_{j'l'} + \delta_{i'l'}\delta_{j'k'} + (\kappa_x - 3)\delta_{i'j'k'l'}\Big) \\ &= \delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk} + (\kappa_x - 3)\sum_{i'} W_{ii'} W_{ji'} W_{ki'} W_{li'}, \end{aligned} \tag{2.12}$$

where we also utilized the tight frame property of $\boldsymbol{W}$ (i.e., the orthonormality of the row vectors of $\boldsymbol{W}$).

For a scalar random variable z which has zero-mean and unit-variance, Jones and Sibson [84] approximated its negentropy, based on the Gram-Charlier Type A series expansion [13] of the density in the vicinity of Gaussian, by

$$J_z \approx \frac{1}{12}(\mathbb{E}\left[z^3\right] - 0)^2 + \frac{1}{48}(\mathbb{E}\left[z^4\right] - 3)^2 \tag{2.13}$$

where "0" and"3" correspond to the third- and fourth-order moment of the standard Gaussian density. For m-dimensional case, this approximation is extended to [84, 107]

$$J(\boldsymbol{z}) \approx \frac{1}{12}\sum_{i,j,k}(\mathbb{E}\left[z_i z_j z_k\right])^2 + \frac{1}{48}\sum_{i,j,k,l}(\mathbb{E}\left[z_i z_j z_k z_l\right] - \delta_{ij}\delta_{kl} - \delta_{ik}\delta_{jl} - \delta_{il}\delta_{jk})^2. \tag{2.14}$$

If we plug (2.11) and (2.12) into (2.14),

$$\begin{aligned} J(\boldsymbol{z}) &\approx \frac{1}{48}\sum_{i,j,k,l}\left((\kappa_x - 3)\sum_{i'} W_{ii'} W_{ji'} W_{ki'} W_{li'}\right)^2 \\ &= \frac{(\kappa_x - 3)^2}{48}\sum_{i',j'}\left(\sum_i W_{ii'} W_{ij'}\right)\left(\sum_j W_{ji'} W_{jj'}\right)\left(\sum_k W_{ki'} W_{kj'}\right)\left(\sum_l W_{li'} W_{lj'}\right) \\ &= \frac{(\kappa_x - 3)^2}{48}\sum_{i',j'}(\boldsymbol{w}_{i'}^T \boldsymbol{w}_{j'})^4, \end{aligned} \tag{2.15}$$

where $\frac{(\kappa_x - 3)^2}{48}$ approximates J_x according to (2.13). By substituting J_x back for $\frac{(\kappa_x - 3)^2}{48}$, we

obtain

$$J(\boldsymbol{z}) \approx J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4. \tag{2.16}$$

By definition of negentropy, $h(\boldsymbol{z}) = \frac{m}{2}\log(2\pi e) - J(\boldsymbol{z})$, where the constant $\frac{m}{2}\log(2\pi e)$ corresponds to the entropy of m-dimensional standard Gaussian density, and thus we immediately have

$$h(\boldsymbol{z}) \approx \frac{m}{2}\log(2\pi e) - J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4. \tag{2.17}$$

In general case with $\sigma_x \neq 1$, if we let $\overline{\boldsymbol{x}} \triangleq \boldsymbol{x}/\sigma_x$, the above result still gives us that $h(\boldsymbol{W}\overline{\boldsymbol{x}}) \approx \frac{m}{2}\log(2\pi e) - J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4$. Using the formula in Lemma 2.12, $h(\boldsymbol{W}\boldsymbol{x}) = h(\boldsymbol{W}\sigma_x\overline{\boldsymbol{x}}) = m\log\sigma_x + h(\boldsymbol{W}\overline{\boldsymbol{x}}) \approx \frac{m}{2}\log(2\pi e\sigma_x^2) - J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4$. Now, the constant $\frac{m}{2}\log(2\pi e\sigma_x^2)$ is the Gaussian entropy with which $h(\boldsymbol{W}\boldsymbol{x})$ should be compared, because the covariance of $\boldsymbol{W}\boldsymbol{x}$ is $\sigma_x^2\boldsymbol{I}$; therefore, $J(\boldsymbol{W}\boldsymbol{x}) = \frac{m}{2}\log(2\pi e\sigma_x^2) - h(\boldsymbol{W}\boldsymbol{x}) \approx J_x \sum_{i,j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4$, invariant to σ_x.

2.6.1.5 Proof of Lemma 2.8

We claim that, for a symmetric, idempotent matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, of rank m,

$$\sum_{i,j} A_{ij}^4 \geq \frac{m^4}{n^3} + \frac{m^2(n-m)^2}{n^3(n-1)} \tag{2.18}$$

where the equality is achieved if and only if

$$|A_{ij}| = \begin{cases} m/n, & \text{if } i = j \\ \sqrt{\frac{m(n-m)}{n^2(n-1)}}, & \text{otherwise.} \end{cases} \tag{2.19}$$

To prove this, we first use a simple inequality that

$$\sum_{i=1}^{n} x_i^2 \geq \frac{1}{n}\left(\sum_i x_i\right)^2, \tag{2.20}$$

which is merely a rearrangement of the following: $n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 = \sum_{i<j}(x_i - x_j)^2 \geq 0$, where it is obvious that the equality holds if and only if $x_1 = \cdots = x_n$. Considering $\{A_{ii}\}$, which satisfies $\sum_i A_{ii} = m$ (due to Lemma 2.14.iii in Section 2.6.2), we can lower-bound $\sum_i A_{ii}^2$ by

$$\sum_i A_{ii}^2 \geq \frac{1}{n}\left(\sum_i A_{ii}\right)^2 = \frac{m^2}{n} \tag{2.21}$$

where the equality holds if and only if $A_{ii} = m/n$ for all i.

Then, $\sum_{i,j} A_{ij}^4$ can be lower-bounded through the following several steps:

$$\sum_{i,j} A_{ij}^4 = \sum_i A_{ii}^4 + \sum_{i\neq j} A_{ij}^4 \tag{2.22}$$

$$\overset{(a)}{\geq} \frac{1}{n}\left(\sum_i A_{ii}^2\right)^2 + \frac{1}{n(n-1)}\left(\sum_{i\neq j} A_{ij}^2\right)^2 \tag{2.23}$$

$$\overset{(b)}{=} \frac{1}{n}\left(\sum_i A_{ii}^2\right)^2 + \frac{1}{n(n-1)}\left(m - \sum_i A_{ii}^2\right)^2 \tag{2.24}$$

$$\overset{(c)}{\geq} \frac{m^4}{n^3} + \frac{m^2(n-m)^2}{n^3(n-1)} \tag{2.25}$$

where the inequality (a) is due to (2.20); the equality (b) due to Lemma 2.14.iv; and the inequality (c) due to the facts that $\sum_i A_{ii}^2 \geq m^2/n$, by (2.21), and that $f(x) = \frac{1}{n}x^2 + \frac{1}{n(n-1)}(m-x)^2$ is strictly increasing for $x \in [m^2/n, \infty)$ with $f(m^2/n) = \frac{m^4}{n^3} + \frac{m^2(n-m)^2}{n^3(n-1)}$. Tracing back the equality conditions of (a) and (c), we can easily find that the overall equality is achieved if and only if (2.19) is satisfied.

Finally, note that $\boldsymbol{w}_i^T\boldsymbol{w}_j$ is the (i,j)th entry of $\boldsymbol{W}^T\boldsymbol{W}$ which is a symmetric, idempotent matrix. If we plug $\boldsymbol{w}_i^T\boldsymbol{w}_j$ in place of A_{ij}, Lemma 2.8 immediately follows.

2.6.1.6 Proof of Observation 2.9

In Lemma 2.8, a lower-bound of ν is given as $m\beta^3 + m\frac{\beta(1-\beta)^2}{(n-1)} \to m\beta^3$ as n goes to infinity. Here, we will show that $\nu(\boldsymbol{W}) = m\beta^3$, which is sufficient to prove this observation. A direct proof on the asymptotic equiangular tight frame property of $\boldsymbol{W}$ needs more

complicated tricks.

Let us evaluate $\nu(\boldsymbol{W}) = \sum_{i,j}(\boldsymbol{w}_i^T\boldsymbol{w}_j)^4$ as below:

1) For $i = j$: Let $\boldsymbol{\Psi}_i \triangleq \sum_{j\neq i}\boldsymbol{h}_j\boldsymbol{h}_j^T$. Then, $\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{\Psi}_i + \boldsymbol{h}_i\boldsymbol{h}_i^T$, where $\boldsymbol{\Psi}_i$ and $\boldsymbol{h}_i$ are mutually independent. By applying the matrix inversion lemma (or Sherman-Morrison formula [15]), we obtain

$$(\boldsymbol{H}\boldsymbol{H}^T)^{-1} = (\boldsymbol{\Psi}_i + \boldsymbol{h}_i\boldsymbol{h}_i^T)^{-1} = \boldsymbol{\Psi}_i^{-1} - \frac{\boldsymbol{\Psi}_i^{-1}\boldsymbol{h}_i\boldsymbol{h}_i^T\boldsymbol{\Psi}_i^{-1}}{1+\boldsymbol{h}_i^T\boldsymbol{\Psi}_i^{-1}\boldsymbol{h}_i}$$

and subsequently

$$\boldsymbol{w}_i^T\boldsymbol{w}_i = \boldsymbol{h}_i^T(\boldsymbol{H}\boldsymbol{H}^T)^{-1}\boldsymbol{h}_i = \frac{\boldsymbol{h}_i^T\boldsymbol{\Psi}_i^{-1}\boldsymbol{h}_i}{1+\boldsymbol{h}_i^T\boldsymbol{\Psi}_i^{-1}\boldsymbol{h}_i}.$$

A central result of random matrix theory [139] applies here: The empirical spectral distribution of $\boldsymbol{\Psi}_i$ almost surely converges to the so-called Marčenko-Pastur law whose density is

$$p_{\text{MP}}(z;\beta) = \frac{\sqrt{(z-a)(b-z)}}{2\pi\beta z}, \quad a < z < b,$$

where $a \triangleq (1-\sqrt{\beta})^2$ and $b \triangleq (1+\sqrt{\beta})^2$. As a result, we can say that

$$\frac{1}{m}\text{Tr}\left(\boldsymbol{\Psi}_i^{-1}\right) \to \mathbb{E}\left[\lambda(\boldsymbol{\Psi}_i^{-1})\right] = \int_a^b \frac{1}{z}p_{\text{MP}}(z;\beta)dz = \frac{1}{1-\beta}.$$
$$\frac{1}{m}\text{Tr}\left(\boldsymbol{\Psi}_i^{-2}\right) \to \mathbb{E}\left[\lambda(\boldsymbol{\Psi}_i^{-2})\right] = \int_a^b \frac{1}{z^2}p_{\text{MP}}(z;\beta)dz = \frac{1}{(1-\beta)^3}$$

where $\lambda(\cdot)$ denotes the eigenvalue. According to Lemma 2.15 (in Section 2.6.2),

$$\text{Var}(\boldsymbol{h}_i^T\boldsymbol{\Psi}_i^{-1}\boldsymbol{h}_i) = \frac{2}{n^2}\text{Tr}(\boldsymbol{\Psi}_i^{-2}) \to \frac{2}{n}\cdot\frac{\beta}{(1-\beta)^3} \to 0$$

while

$$\mathbb{E}\left[\boldsymbol{h}_i^T\boldsymbol{\Psi}_i^{-1}\boldsymbol{h}_i\right] = \text{Tr}(\boldsymbol{\Psi}_i^{-1}\mathbb{E}\left[\boldsymbol{h}_i\boldsymbol{h}_i^T\right]) = \frac{1}{n}\text{Tr}(\boldsymbol{\Psi}_i^{-1}) \to \frac{\beta}{1-\beta}.$$

The above two equations imply that $\boldsymbol{h}_i^T\boldsymbol{\Psi}_i^{-1}\boldsymbol{h}_i \to \frac{\beta}{1-\beta}$ and subsequently that $\boldsymbol{w}_i^T\boldsymbol{w}_i =$

$\frac{\boldsymbol{h}_i^T \boldsymbol{\Psi}_i^{-1} \boldsymbol{h}_i}{1+\boldsymbol{h}_i^T \boldsymbol{\Psi}_i^{-1} \boldsymbol{h}_i} \to \frac{\frac{\beta}{1-\beta}}{1+\frac{\beta}{1-\beta}} = \beta$, for all i. Therefore, $\sum_{i=j} (\boldsymbol{w}_i^T \boldsymbol{w}_j)^4 = n\beta^4 = m\beta^3$.

2) For $i \neq j$: Let $\boldsymbol{\Omega}_{ij} = \sum_{k \neq i,j} \boldsymbol{h}_k \boldsymbol{h}_k^T$. Then, $\boldsymbol{H}\boldsymbol{H}^T = \boldsymbol{\Omega}_{ij} + \boldsymbol{h}_i \boldsymbol{h}_i^T + \boldsymbol{h}_j \boldsymbol{h}_j^T$. Similarly as in the previous case, we apply the matrix inversion lemma to obtain

$$\boldsymbol{w}_i^T \boldsymbol{w}_j = \frac{\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j}{(1 + \boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_i)(1 + \boldsymbol{h}_j^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j) - (\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j)^2}.$$

In the asymptotic condition, $\boldsymbol{\Omega}_{ij}$ has the same spectral distribution as $\boldsymbol{\Psi}_i$. Therefore, $\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_i = \boldsymbol{h}_j^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j \to \frac{\beta}{1-\beta}$. We can also loosely bound $(\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j)^2$ by $(\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j)^2 \leq (\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_i)(\boldsymbol{h}_j^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j) \to \frac{\beta^2}{(1-\beta)^2}$ using Cauchy-Schwartz inequality. Therefore,

$$|\boldsymbol{w}_i^T \boldsymbol{w}_j| \leq \frac{\left|\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j\right|}{1/(1-\beta)^2 - \beta^2/(1-\beta)^2} = \frac{1-\beta}{1+\beta} \left|\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j\right| \leq (1-\beta) \left|\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j\right|.$$

Let us write $\boldsymbol{\Omega}_{ij}^{-1} = \boldsymbol{U}_{ij} \boldsymbol{D}_{ij} \boldsymbol{U}_{ij}^T$ in its SVD, where $\boldsymbol{D}_{ij} = \mathrm{diag}(\lambda_1(\boldsymbol{\Omega}_{ij}^{-1}), \ldots, \lambda_m(\boldsymbol{\Omega}_{ij}^{-1}))$. For any instance of $\boldsymbol{U}_{ij}$, which is always orthonormal, two random vectors $\boldsymbol{f} \triangleq \boldsymbol{U}_{ij}^T \boldsymbol{h}_i$ and $\boldsymbol{g} \triangleq \boldsymbol{U}_{ij}^T \boldsymbol{h}_j$ are still Gaussian and mutually independent. Then,

$$\begin{aligned}
\mathbb{E}\left[|\boldsymbol{h}_i^T \boldsymbol{\Omega}_{ij}^{-1} \boldsymbol{h}_j|^4\right] &= \mathbb{E}\left[|\boldsymbol{f}^T \boldsymbol{D}_{ij} \boldsymbol{g}|^4\right] \\
&= \mathbb{E}\left[(\sum_{k=1}^{m} \lambda_k(\boldsymbol{\Omega}_{ij}^{-1}) f_k g_k)^4\right] \\
&= \sum_{k_1,k_2,k_3,k_4=1}^{m} \lambda_{k_1}(\boldsymbol{\Omega}_{ij}^{-1}) \lambda_{k_2}(\boldsymbol{\Omega}_{ij}^{-1}) \lambda_{k_3}(\boldsymbol{\Omega}_{ij}^{-1}) \lambda_{k_4}(\boldsymbol{\Omega}_{ij}^{-1}) \\
&\qquad \times \mathbb{E}\left[f_{k_1} f_{k_2} f_{k_3} f_{k_4}\right] \mathbb{E}\left[g_{k_1} g_{k_2} g_{k_3} g_{k_4}\right] \\
&= \frac{3}{n^4} (\sum_{k=1}^{m} \lambda_k^2(\boldsymbol{\Omega}_{ij}^{-1}))^2 + \frac{6}{n^4} \sum_{k=1}^{m} \lambda_k^4(\boldsymbol{\Omega}_{ij}^{-1}) \\
&\to \frac{3m^2 (\mathbb{E}\left[\lambda^2(\boldsymbol{\Omega}_{ij}^{-1})\right])^2 + 6m \mathbb{E}\left[\lambda^4(\boldsymbol{\Omega}_{ij}^{-1})\right]}{n^4}
\end{aligned}$$

where we have made use that $\mathbb{E}\left[f_{k_1} f_{k_2} f_{k_3} f_{k_4}\right] = \mathbb{E}\left[g_{k_1} g_{k_2} g_{k_3} g_{k_4}\right] = (\delta_{k_1 k_2} \delta_{k_3 k_4} + \delta_{k_1 k_3} \delta_{k_2 k_4} + \delta_{k_1 k_4} \delta_{k_2 k_3})/n^2$ for Kronecker delta δ that produces one if its subscripts have the same value

and zero otherwise. Since

$$\mathbb{E}\left[\lambda^2(\mathbf{\Omega}_{ij}^{-1})\right] = \int \frac{1}{z^2} p_{\text{MP}}(z;\beta)dz = \frac{1}{(1-\beta)^3}$$
$$\mathbb{E}\left[\lambda^4(\mathbf{\Omega}_{ij}^{-1})\right] = \int \frac{1}{z^4} p_{\text{MP}}(z;\beta)dz = \frac{1+3\beta+\beta^2}{(1-\beta)^7},$$

we obtain

$$\mathbb{E}\left[|\boldsymbol{w}_i^T\boldsymbol{w}_j|^4\right] \leq (1-\beta)^4 \cdot \frac{3m^2\frac{1}{(1-\beta)^6} + 6m\frac{1+3\beta+\beta^2}{(1-\beta)^7}}{n^4} = \frac{3\beta^2}{n^2(1-\beta)^2} + \frac{6\beta(1+3\beta+\beta^2)}{n^3(1-\beta)^3}.$$

Therefore, $\sum_{i\neq j}|\boldsymbol{w}_i^T\boldsymbol{w}_j|^4 \to n(n-1)\mathbb{E}\left[|\boldsymbol{w}_i^T\boldsymbol{w}_j|^4\right] \leq \frac{3\beta^2}{(1-\beta)^2} + \frac{6}{n}\frac{\beta(1+3\beta+\beta^2)}{(1-\beta)^3}$.

Overall,

$$\nu(\boldsymbol{W}) = \sum_{i=j}(\boldsymbol{w}_i^T\boldsymbol{w}_j)^4 + \sum_{i\neq j}(\boldsymbol{w}_i^T\boldsymbol{w}_j)^4 \leq m\beta^3 + \frac{3\beta^2}{(1-\beta)^2} + \frac{6}{n}\frac{\beta(1+3\beta+\beta^2)}{(1-\beta)^3} \to m\beta^3$$

as $m, n \to \infty$. Since this upper-bound is also a lower-bound of $\nu(\boldsymbol{W})$ as we explained at the beginning of the proof, $\nu(\boldsymbol{W}) = m\beta^3$ in the limit of $m, n \to \infty$. Inserting this value into the entropy approximation in Proposition 2.7, we have $\widehat{h}(\boldsymbol{W}\boldsymbol{x}) = \frac{m}{2}\log(2\pi e\sigma_x^2) - mJ_x\beta^3$, as claimed.

2.6.1.7 Proof of Lemma 2.10

By Lemma 2.16 (in Section 2.6.2),

$$-\log\det(\sum_i \frac{1}{\sigma_i^2}\boldsymbol{q}_i\boldsymbol{q}_i^T) \leq -\log\prod_i(\frac{1}{\sigma_i^2})^{\|\boldsymbol{q}_i\|^2} = \sum_i \|\boldsymbol{q}_i\|^2\log\sigma_i^2. \tag{2.26}$$

We can also upper-bound $-J_x\sum_{i,j}(\boldsymbol{q}_i^T\boldsymbol{q}_j)^4$ by

$$-J_x\sum_{i,j}(\boldsymbol{q}_i^T\boldsymbol{q}_j)^4 = -J_x\sum_i(\boldsymbol{q}_i^T\boldsymbol{q}_i)^4 - J_x\sum_{i\neq j}(\boldsymbol{q}_i^T\boldsymbol{q}_j)^4 \tag{2.27}$$
$$\leq -J_x\sum_i(\boldsymbol{q}_i^T\boldsymbol{q}_i)^4 \tag{2.28}$$

$$= -J_x \sum_i \|\boldsymbol{q}_i\|^8. \tag{2.29}$$

Putting both together, we obtain an interim result that the approximated entropy in Equation (2.6) is upper-bounded by

$$\widehat{h}_U \triangleq \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_i \|\boldsymbol{q}_i\|^2 \log\sigma_i^2 - J_x \sum_i \|\boldsymbol{q}_i\|^8. \tag{2.30}$$

To prove Lemma 2.10, we maximize $\widehat{h}_U$ in (2.30) with respect to $\|\boldsymbol{q}_i\|^2$, which is constrained by

$$0 \le \|\boldsymbol{q}_i\|^2 \le 1, \quad \sum_i \|\boldsymbol{q}_i\|^2 = m \tag{2.31}$$

since it is a diagonal entry of the symmetric, idempotent matrix $\boldsymbol{Q}^T\boldsymbol{Q}$ (see Lemma 2.14.i–iii in Section 2.6.2). From the first order condition $\partial\widehat{h}_U/\partial\|\boldsymbol{q}_i\|^2 = 0$ with a Lagrange multiplier ξ, we obtain

$$\varepsilon_i \triangleq \arg\max_{\|\boldsymbol{q}_i\|^2} \widehat{h}_U(\{\|\boldsymbol{q}_i\|^2\}; \{\sigma_i^2, J_x\}) \tag{2.32}$$

$$= \text{median}\left(\sqrt[3]{\frac{\log\sigma_i^2 + \xi}{8J_x}},\ 0,\ 1\right), \quad \forall i = 1, \ldots, n \tag{2.33}$$

with the maximum value being

$$\widehat{h}_U(\{\varepsilon_i\}; \{\sigma_i^2, J_x\}) = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_i \varepsilon_i \log\sigma_i^2 - J_x \sum_i \varepsilon_i^4. \tag{2.34}$$

Finally, the Lagrange multiplier ξ should be determined to satisfy (2.31), i.e., $\sum_i \varepsilon_i = m$.

2.6.1.8 Details on Example 2.1

In Equation (2.8), $\varepsilon_1 = \cdots = \varepsilon_n$ because $\sigma_1^2 = \cdots = \sigma_n^2$. Further, the value of ε_i should be equal to m/n by the condition $\sum_i \varepsilon_i = m$. Then, Equation (2.7) gives $\widehat{h}^* = \frac{m}{2}\log(2\pi e) + \frac{m}{2}\log\sigma_x^2 - mJ_x(\frac{m}{n})^3$.

In Equation (2.5), we showed that if $n \to \infty$ with $m/n \to \beta < 1$, a tight frame obtained by normalizing an $m \times n$ random Gaussian matrix makes $\widehat{h} = \frac{m}{2}\log(2\pi e\sigma_x^2) - mJ_x\beta^3$, which is the same as $\widehat{h}^*$ in the above.

2.6.1.9 Details on Example 2.2

1) If $\frac{\log\sigma_{m+1}^2+\xi}{8J_x} \leq 0$, $\varepsilon_{m+1} = \text{median}(\sqrt[3]{\frac{\log\sigma_{m+1}^2+\xi}{8J_x}}, 0, 1) = 0$. Because ε_i's monotonically decrease (by the assumption that $\sigma_1 \geq \cdots \geq \sigma_n$) and cannot be negative, $\varepsilon_{m+1} = \cdots = \varepsilon_n = 0$. Then, to make $\sum_i \varepsilon_i = m$, all the other ε_i's should be one (ε_i's cannot be greater than one).

2) If $\frac{\log\sigma_{m+1}^2+\xi}{8J_x} \geq 0$, by the condition (2.9),

$$\sqrt[3]{\frac{\log\sigma_m^2+\xi}{8J_x}} \geq \sqrt[3]{\frac{\log(\sigma_{m+1}^2 e^{8J_x})+\xi}{8J_x}} = \sqrt[3]{1+\frac{\log\sigma_{m+1}^2+\xi}{8J_x}} \geq 1$$

and therefore, $\varepsilon_m = \text{median}(\sqrt[3]{\frac{\log\sigma_m^2+\xi}{8J_x}}, 0, 1) = 1$. Since $\varepsilon_m \leq \cdots \leq \varepsilon_1 \leq 1$, it follows that $\varepsilon_1 = \cdots = \varepsilon_m = 1$. All the remaining ε_i's should be zero because $\sum_i \varepsilon_i = m$ and $\varepsilon_i \geq 0$ for all i.

In either case, we have $\varepsilon_1 = \cdots = \varepsilon_m = 1$ and $\varepsilon_{m+1} = \cdots = \varepsilon_n = 0$. Then, Equation (2.7) gives $\widehat{h}^* = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^m \log\sigma_i^2 - mJ_x$.

The entropy of the PCA projection is computed by

$$h = \sum_{i=1}^m h(x_i) = \sum_{i=1}^m \left(\frac{1}{2}\log(2\pi e\sigma_i^2) - J_x\right),$$

which is the same as $\widehat{h}^*$ in the above.

2.6.1.10 Details on Example 2.3

Given $\widetilde{\varepsilon}_j$, the upper-bound $\widehat{h}^*$ in Equation (2.7) can be simply rewritten as

$$\widehat{h}^* = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_j n_j\widetilde{\varepsilon}_j \log\widetilde{\sigma}_j^2 - J_x\sum_j n_j\widetilde{\varepsilon}_j^4.$$

In the proposed scheme, the jth group is measured by the normalization of an $m_j \times n_j$ random Gaussian matrix and in the asymptotic condition ($n_j \to \infty$), the approximated entropy becomes equal to $\frac{m_j}{2}\log(2\pi e\widetilde{\sigma}_j^2) - m_j J_x(\frac{m_j}{n_j})^3$ (see Equation 2.5).

Summing over all the (independent) groups, we obtain

$$\widehat{h} = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_j m_j \log\widetilde{\sigma}_j^2 - J_x \sum_j m_j \left(\frac{m_j}{n_j}\right)^3,$$

which is equal to $\widehat{h}^*$ in the above, if $m_j = n_j\widetilde{\varepsilon}_j$ or $m_j/n_j = \widetilde{\varepsilon}_j$. Finally, note that, for any given $\widetilde{\varepsilon}_j \geq 0$, we can take an integer m_j which makes $|\frac{m_j}{n_j} - \widetilde{\varepsilon}_j|$ be arbitrarily small since $n_j \to \infty$ (i.e. infinite resolution).

2.6.1.11 Details on Example 2.4

Let $\xi = -\log\sigma^2_{m+m_1+1}$. Then,

$$\varepsilon_i = \text{median}\left(\sqrt[3]{\frac{\log\sigma_i^2 - \log\sigma^2_{m+m_1+1}}{8J_x}}, 0, 1\right) = \begin{cases} 1, & i \leq m - m_0 \\ 0, & i \geq m + m_1 + 1 \\ \frac{m_0}{m_0+m_1}, & \text{elsewhere,} \end{cases}$$

and $\sum_i \varepsilon_i = m$, which validates our choice of $\xi = -\log\sigma^2_{m+m_1+1}$. With this set of $\{\varepsilon_i\}$, Equation (2.7) gives $\widehat{h}^* = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\sum_{i=1}^m \log\sigma_i^2 - mJ_x + m_0 J_x\left(1 - \left(\frac{m_0}{m_0+m_1}\right)^3\right)$.

On the other hand, the entropy of the major $(m - m_0)$ principal components is equal to

$$h_1 = \sum_{i=1}^{m-m_0} h(x_i) = \sum_{i=1}^{m-m_0}\left(\frac{1}{2}\log(2\pi e\sigma_i^2) - J_x\right), \tag{2.35}$$

and the entropy of the normalization of m_0 random projections over the next $(m_0 + m_1)$ principal components is computed by (refer to Equation 2.5)

$$\widehat{h}_2 = \frac{m_0}{2}\log(2\pi e\sigma_m^2) - J_x\frac{m_0^4}{(m_0+m_1)^3} \tag{2.36}$$

in the asymptotic case. The total entropy is simply the addition of (2.35) and (2.36) and becomes equal to $\widehat{h}^*$.

2.6.1.12 Proof of Lemma 2.11

Let $\boldsymbol{P} \triangleq \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Q}^T(\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-1}\boldsymbol{Q}\boldsymbol{\Sigma}^{-\frac{1}{2}}$. The matrix $\boldsymbol{P}$ is a symmetric, idempotent matrix that operates the projection onto the rowspace of $\boldsymbol{W}$ (e.g., see [131]). We also write $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ without loss of generality.

1) $i \in S_0$: This implies that $\boldsymbol{q}_i = \boldsymbol{0}$ and subsequently that

$$\boldsymbol{P}\boldsymbol{e}_i = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Q}^T(\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-1}\underbrace{\boldsymbol{Q}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}_i}_{=\frac{1}{\sigma_i}\boldsymbol{q}_i=\boldsymbol{0}} = \boldsymbol{0}. \tag{2.37}$$

The vector $\boldsymbol{e}_i$ is an eigenvector of the matrix $\boldsymbol{P}$ with the eigenvalue zero. This means that the rowspace of $\boldsymbol{W}$ completely excludes the ith coordinate.

2) $i \in S_1$: Note that $\sum_j(\boldsymbol{q}_i^T\boldsymbol{q}_j)^2 = \boldsymbol{q}_i^T\left(\sum_j \boldsymbol{q}_j\boldsymbol{q}_j^T\right)\boldsymbol{q}_i = \boldsymbol{q}_i^T\boldsymbol{q}_i = \|\boldsymbol{q}_i\|^2 = 1$ because $\boldsymbol{Q}$ is a tight frame satisfying $\boldsymbol{Q}\boldsymbol{Q}^T = \sum_j \boldsymbol{q}_j\boldsymbol{q}_j^T = \boldsymbol{I}$. It follows that

$$1 = \sum_j(\boldsymbol{q}_i^T\boldsymbol{q}_j)^2 = (\boldsymbol{q}_i^T\boldsymbol{q}_i)^2 + \sum_{j\neq i}(\boldsymbol{q}_i^T\boldsymbol{q}_j)^2 = 1 + \sum_{j\neq i}(\boldsymbol{q}_i^T\boldsymbol{q}_j)^2 \tag{2.38}$$

and that $\boldsymbol{q}_i^T\boldsymbol{q}_j = 0$ for all $j \neq i$. As a consequence, we can write

$$\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\underbrace{\boldsymbol{Q}^T\boldsymbol{q}_i}_{=\boldsymbol{e}_i} = \frac{1}{\sigma_i^2}\boldsymbol{q}_i, \tag{2.39}$$

or

$$\boldsymbol{q}_i = \frac{1}{\sigma_i^2}(\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-1}\boldsymbol{q}_i. \tag{2.40}$$

Therefore,

$$\boldsymbol{P}\boldsymbol{e}_i = \boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Q}^T(\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-1}\underbrace{\boldsymbol{Q}\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{e}_i}_{=\frac{1}{\sigma_i}\boldsymbol{q}_i} \tag{2.41}$$

$$= \sigma_i\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{Q}^T\underbrace{(\boldsymbol{Q}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}^T)^{-1}\frac{1}{\sigma_i^2}\boldsymbol{q}_i}_{=\boldsymbol{q}_i} \tag{2.42}$$

$$= \sigma_i \boldsymbol{\Sigma}^{-\frac{1}{2}} \underbrace{\boldsymbol{Q}^T \boldsymbol{q}_i}_{=\boldsymbol{e}_i} \tag{2.43}$$

$$= \sigma_i \boldsymbol{\Sigma}^{-\frac{1}{2}} \boldsymbol{e}_i \tag{2.44}$$

$$= \boldsymbol{e}_i. \tag{2.45}$$

The vector $\boldsymbol{e}_i$ is an eigenvector of the matrix $\boldsymbol{P}$ with the eigenvalue one. This means that the rowspace of $\boldsymbol{W}$ includes the ith coordinate.

2.6.2 Miscellaneous Lemmas

Lemma 2.12. For any invertible matrix $\boldsymbol{A}$ and for any random vector $\boldsymbol{x}$, the following decomposition holds:

$$h(\boldsymbol{A}\boldsymbol{x}) = h(\boldsymbol{x}) + \log|\det(\boldsymbol{A})|. \tag{2.46}$$

Proof. The proof is simple, where $|\det(\boldsymbol{A})|$ appears as a Jacobian factor in the probability density of the linear transformation $\boldsymbol{A}\boldsymbol{x}$. See [43] for the complete proof. □

Lemma 2.13. Any symmetric, idempotent matrix is positive semi-definite.

Proof. If $\boldsymbol{A}$ is symmetric and idempotent,

$$\boldsymbol{A} = \boldsymbol{A}^2 = \boldsymbol{A}^T\boldsymbol{A},$$

and, for any vector $\boldsymbol{x}$,

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = ||\boldsymbol{A}\boldsymbol{x}||^2 \geq 0,$$

which proves the positive semi-definiteness of $\boldsymbol{A}$. □

Lemma 2.14. Let $\boldsymbol{A}$ be a symmetric, idempotent matrix of rank m. Then, its entry A_{ij} should satisfy the following properties: (i) $A_{ii} \geq 0$, for all i; (ii) $|A_{ij}| \leq 1$, for all i, j; (iii) $\sum_i A_{ii} = m$; and (iv) $\sum_{i,j} A_{ij}^2 = m$.

Proof. If we let $\boldsymbol{v}$ be an eigenvector of $\boldsymbol{A}$ with the eigenvalue λ, by the idempotent property of $\boldsymbol{A}$, $\lambda\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} = \boldsymbol{A}^2\boldsymbol{v} = \lambda^2\boldsymbol{v}$, and we see that λ should be either 0 or 1. Since $\boldsymbol{A}$ is symmetric, the singular values should also be either 0 or 1 by the following:

$$\underbrace{\sigma(\boldsymbol{A})}_{\text{singular value}} = \sqrt{\underbrace{\lambda(\boldsymbol{A}^T\boldsymbol{A})}_{\text{eigenvalue}}} = \sqrt{\lambda(\boldsymbol{A})} = \begin{cases} 0, & \text{if } \lambda(\boldsymbol{A}) = 0, \\ 1, & \text{if } \lambda(\boldsymbol{A}) = 1. \end{cases}$$

The rank of a matrix is equal to the number of nonzero singular values and, specifically for the symmetric, idempotent case, to the number of nonzero eigenvalues as well. Therefore, our matrix $\boldsymbol{A}$ should have m number of 1's and the remaining number of 0's as its singular values and as its eigenvalues.

(i) For all i,

$$A_{ii} = \boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{e}_i \geq 0$$

by the positive semi-definiteness of $\boldsymbol{A}$ (see Lemma 2.13), where $\boldsymbol{e}_i$ denotes a unit vector with one at the ith entry.

(ii) For all i, j,

$$|A_{ij}| \leq \|\boldsymbol{A}\boldsymbol{e}_j\|_\infty \leq \|\boldsymbol{A}\boldsymbol{e}_j\|_2 \leq \sigma_{\max}(\boldsymbol{A})\|\boldsymbol{e}_j\|_2 = 1.$$

(iii) The trace is equal to the sum of the eigenvalues, and thus $\sum_i A_{ii} = \mathrm{Tr}(\boldsymbol{A}) = m$.

(iv) Denoting the Frobenius norm of a matrix by $\|\cdot\|_F$,

$$\sum_{i,j} A_{ij}^2 = \|\boldsymbol{A}\|_F^2 = \mathrm{Tr}(\boldsymbol{A}^T\boldsymbol{A}) = \mathrm{Tr}(\boldsymbol{A}) = m$$

where the symmetric, idempotent property of $\boldsymbol{A}$ was used in the second-last equality. □

Lemma 2.15. Let $x_1, \ldots, x_n$ be i.i.d. Gaussian random variables, with zero-mean and variance σ^2. If $\boldsymbol{A}$ is any $n \times n$ symmetric matrix, then

$$\mathrm{Var}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = 2\sigma^4\mathrm{Tr}(\boldsymbol{A}^2).$$

Proof. To evaluate $\mathrm{Var}(\cdot)$, we compute $\mathbb{E}\left[(\cdot)^2\right]$ and $\mathbb{E}\left[\cdot\right]$ in the followings:

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})^2\right] &= \sum_{i,j,k,l} A_{ij}A_{kl}\mathbb{E}\left[x_i x_j x_k x_l\right] \\
&= \sigma^4(\sum_i A_{ii})^2 + 2\sigma^4\sum_{i,j} A_{ij}^2 \\
&= \sigma^4\left(\mathrm{Tr}(\boldsymbol{A})^2 + 2\mathrm{Tr}(\boldsymbol{A}^2)\right), \\
\mathbb{E}\left[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right] &= \mathbb{E}\left[\mathrm{Tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^T)\right] \\
&= \sigma^2\mathrm{Tr}(\boldsymbol{A}).
\end{aligned}$$

Finally, $\mathrm{Var}(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}) = \mathbb{E}\left[(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})^2\right] - \left(\mathbb{E}\left[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right]\right)^2 = 2\sigma^4\mathrm{Tr}(\boldsymbol{A}^2)$. □

Lemma 2.16. Suppose a set of symmetric, positive semi-definite matrices, $\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_n \in \mathbb{R}^{m\times m}$, which add up to the identity, i.e., $\sum_{i=1}^n \boldsymbol{A}_i = \boldsymbol{I}$. Then, the following inequality holds for any $c_i > 0$:

$$\det(\sum_{i=1}^n c_i\boldsymbol{A}_i) \geq \prod_{i=1}^n c_i^{\mathrm{Tr}(\boldsymbol{A}_i)}. \tag{2.47}$$

Proof. Assume $c_1 \leq c_2 \leq \cdots \leq c_n$ without loss of generality and denote the eigenvalues of $\boldsymbol{A} \triangleq \sum_i c_i\boldsymbol{A}_i$ by $\{\lambda_k\}$, with each λ_k associated with a unit-length eigenvector $\boldsymbol{v}_k$. Since $\boldsymbol{A}$ is symmetric, $\{\boldsymbol{v}_k\}$ forms an orthonormal set. Therefore, $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_m]$ becomes an orthonormal square matrix, i.e., satisfying $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$.

For all k, λ_k can be written as

$$\begin{aligned}
\lambda_k &= \boldsymbol{v}_k^T\boldsymbol{A}\boldsymbol{v}_k \\
&= \boldsymbol{v}_k^T(\sum_i c_i\boldsymbol{A}_i)\boldsymbol{v}_k \\
&= \sum_i c_i\boldsymbol{v}_k^T\boldsymbol{A}_i\boldsymbol{v}_k \\
&= \sum_i c_i\rho_{ki}
\end{aligned}$$

where $\rho_{ki} \triangleq \boldsymbol{v}_k^T\boldsymbol{A}_i\boldsymbol{v}_k$. Note that ρ_{ki}'s have the following properties:

(i) For all k and for all i, $\rho_{ki} \geq 0$, due to the positive semi-definiteness of $\boldsymbol{A}_i$;

(ii) For all k, $\sum_i \rho_{ki} = 1$ because

$$\begin{aligned} 1 &= \|\boldsymbol{v}_k\|^2 \\ &= \boldsymbol{v}_k^T (\sum_i \boldsymbol{A}_i) \boldsymbol{v}_k \\ &= \sum_i \rho_{ki}; \end{aligned}$$

(iii) For all i, $\sum_k \rho_{ki} = \mathrm{Tr}(\boldsymbol{A}_i)$ because

$$\begin{aligned} \mathrm{Tr}(\boldsymbol{A}_i) &= \mathrm{Tr}(\boldsymbol{V}^T \boldsymbol{A}_i \boldsymbol{V}) \\ &= \sum_k \boldsymbol{v}_k^T \boldsymbol{A}_i \boldsymbol{v}_k \\ &= \sum_k \rho_{ki}. \end{aligned}$$

Due to the properties (i) and (ii),

$$\sum_i c_i \rho_{ki} \geq \prod_i c_i^{\rho_{ki}},$$

which is a well-known generalization of the arithmetic and geometric mean inequality, so $\lambda_k \geq \prod_i c_i^{\rho_{ki}}$ for all k. Finally, the inequality (2.47) can be shown by

$$\begin{aligned} \det(\boldsymbol{A}) &= \prod_k \lambda_k \\ &\geq \prod_k \prod_i c_i^{\rho_{ki}} \\ &= \prod_i c_i^{\sum_k \rho_{ki}} \\ &= \prod_i c_i^{\mathrm{Tr}(\boldsymbol{A}_i)} \end{aligned}$$

where the last equality holds because of the property (iii). □

2.6.3 MMSE Estimation Based on Posterior Sampling

As we mentioned in the introduction, InfoMax does not assume any specific recovery scheme. In this chapter, we primarily use the MMSE estimate rather than ℓ_1-regularization typically used in the RIP-based theory. We believe that the choice will minimize any artifacts irrelevant to the measurement scheme itself. The error with the MMSE estimate is the best achievable reconstruction fidelity, in ℓ_2-norm, and thus gives a fundamental limit to each measurement scheme. The computational cost of the MMSE estimate is usually higher than that of ℓ_1-regularization, but it is not a focus of this study. There were, and still are, a great deal of research efforts in making the MMSE estimation efficient inside and outside the CS community [104, 72, 14, 113]. If a random matrix is used for $\boldsymbol{W}$ in the large system limit ($m, n \to \infty$ with $\beta \to m/n$), the MMSE may also be computable using the replica method [71, 114] without explicit construction of the MMSE estimate.

In this chapter, we take rather a direct approach: generating a number of posterior samples and computing the average. First, we suppose that $p(x_i)$ is representable with a mixture of Gaussians, i.e., $p(x_i) = \sum_j a_{i,j}\mathcal{N}(x_i; 0, \sigma_{i,j}^2)$; if not, we use the expectation maximization (EM) method [56] to approximate the true density by such a mixture of Gaussians. Let c_i denote the (auxiliary) indicator variable, so $c_i = j$ when x_i comes from the jth Gaussian distribution. Then, we can use the auxiliary-variable Gibbs sampling [123] to generate the posterior samples of $\boldsymbol{x}$ given $\boldsymbol{y}$. The sampling consists of two alternating steps:

(S1) Sample auxiliary variables c_i given x_i, for all i:

$$\Pr(c_i = j|x_i) \propto \frac{a_{i,j}}{\sigma_{i,j}} \exp\left(-\frac{x_i^2}{2\sigma_{i,j}^2}\right)$$

(S2) Sample $\boldsymbol{x}$ given ($\boldsymbol{c}$, $\boldsymbol{y}$): If hidden variables $\boldsymbol{c} = (c_1, \ldots, c_n)$ are given, the density of $\boldsymbol{x}$ is simply a zero-mean Gaussian with the covariance $\boldsymbol{\Sigma_c} \triangleq \mathrm{diag}(\sigma_{1,c_1}^2, \ldots, \sigma_{n,c_n}^2)$. Additionally given $\boldsymbol{y}$, the density becomes (e.g., see [105])

$$p(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{\Sigma_c W}^T(\boldsymbol{W\Sigma_c W}^T)^{-1}\boldsymbol{y}, \boldsymbol{\Sigma_c} - \boldsymbol{\Sigma_c W}^T(\boldsymbol{W\Sigma_c W}^T)^{-1}\boldsymbol{W\Sigma_c}).$$

We can efficiently sample $\boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{y})$ by computing

$$\boldsymbol{x} = \boldsymbol{\Sigma}_{\boldsymbol{c}}^{1/2} \left(\boldsymbol{\phi} + \boldsymbol{\Sigma}_{\boldsymbol{c}}^{1/2} \boldsymbol{W}^T (\boldsymbol{W} \boldsymbol{\Sigma}_{\boldsymbol{c}} \boldsymbol{W}^T)^{-1} (\boldsymbol{y} - \boldsymbol{W} \boldsymbol{\Sigma}_{\boldsymbol{c}}^{1/2} \boldsymbol{\phi}) \right),$$

based on $\boldsymbol{\phi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$, which does not involve the Cholesky decomposition of the covariance matrix.

Every run yields a single sample of $\boldsymbol{x}$. In the steady-state of the alternation between (S1) and (S2), the samples follow the posterior density $p(\boldsymbol{x}|\boldsymbol{y})$ and all we need to do is to take the mean of a number of the steady-state samples.

Chapter 3

Informative Sensing for Natural Images

In this chapter, we investigate the application of informative sensing to natural images. Specifically, we assume that a bandwise i.i.d. signal model is a good approximation to natural images. Then, we use a nonuniform-density bandwise random projection, shown in the previous chapter to be most informative for such a model.

We present, in somewhat intuitive fashion, how to effectively apply bandwise projections to images, with reference to some well-known statistics of natural images. Experimental results demonstrate that bandwise random projections consistently outperform other kinds of projections (e.g., PCA, random) in image reconstruction. In the presence of noise, we also consider optimal power distribution among sensors within a given budget. The optimization makes the measurement be robust to the noisy setting. In this aspect, we generalize the result of Linsker, who previously optimized the power distribution for Gaussian inputs, to natural images which are not Gaussian. The improvement of noise tolerance is experimentally shown.

3.1 Introduction

Consider a linear projection $\boldsymbol{y} \in \mathbb{R}^m$ of a signal $\boldsymbol{x} \in \mathbb{R}^n$. They are related by a rectangular matrix $\boldsymbol{W}$ such that $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$. We assume $m < n$, which implies that the projection performs a dimension reduction. What choice of $\boldsymbol{W}$ is expected to enable the best reconstruction of $\boldsymbol{x}$?

A nice result is available if we keep recovery to be *linear* and use the average ℓ_2-error for the recovery performance measure. Principal component analysis (PCA) finds an optimal measurement matrix: the rowspace of $\boldsymbol{W}$ must be the linear hull of the major m principal components of $\boldsymbol{x}$ (see Proposition 3.5). However, if $\boldsymbol{x}$ is not Gaussian, *nonlinear* operators, which exploit the non-Gaussianity, can remarkably improve the reconstruction. With the relaxation of the linear recovery constraint, the optimal measurement can be substantially different from the PCA projection. Compressed sensing (CS) is a fabulous demonstration of such nonlinear recovery from highly incomplete linear samples of sparse signals [51, 32]. The performance of CS with random measurements has rigorously been analyzed when the signal is exactly or approximately sparse [51, 32, 119, 40].

While natural images are approximately sparse in a wavelet basis (e.g., see Figure 1.1), recent studies have found some evidence that random measurements are not optimal for natural images. For example, Seeger and Nickisch [127] and Haupt and Nowak [74] found that standard low-pass filtering followed by subsampling (similar to the PCA projection) often gives better reconstruction results than random projections. Lustig, Donoho, and Pauly [98] noticed that including more low-frequency samples than high-frequency samples can produce better performance for real images when using a random Fourier matrix. Romberg [115] uses 1,000 low-frequency discrete cosine transform (DCT) coefficients[1] together with the remaining amount of random projections rather than purely uses random projections for all. Romberg's approach (i.e., joint use of PCA and random projections) has been followed by a few other studies (e.g., see [121, 129]).

In Chapter 2, we have made crucial progress in optimizing linear measurements in terms of uncertainty minimization. According to our results, random projections provide a best measurement scheme if the signal is assumed to be sparse in an orthonormal basis *and* if no further statistical information of the signal is available a priori. The statistics of natural images, however, are well known and certainly show more structure than simple sparsity. In this chapter, we argue that a bandwise i.i.d. signal model is a more precise description for the natural image statistics than the simple sparsity model. Hence, we use a nonuniform-density bandwise random projection, known to be most informative for such a

[1]The DCT kernels are known to well approximate the principal components of natural images [3].

bandwise i.i.d. signal (see Example 2.3), in constructing the matrix $\boldsymbol{W}$.

This chapter is based on the analytical results of Chapter 2 but has a couple of distinct contributions. First, this chapter deals with an effective implementation of the bandwise random projections. We will give readers more intuition on how the bandwise random projections work, with reference to well-known image statistics. We also provide an efficient method to determine the nonuniform density profile, which is conceptually and algorithmically simpler than the one originally given in Chapter 2.

Second, we take the measurement noise into account. A vast majority of prior work considering measurement noises focuses on bounding the recovery performance of random projections in noisy settings [143, 75, 122, 63, 144, 5, 1, 112], while only a few on the design of the measurement matrix [146, 34, 127]. Furthermore, the power distribution among sensors, within a restricted budget, has rarely been considered. In this chapter, we derive an optimal power distribution for the bandwise random measurements within the InfoMax framework. This may be considered to generalize the twenty year-old result of Linsker [94] – the optimal power distribution for Gaussian inputs – to natural images which are not Gaussian.

3.2 Informative Sensing

In [94], Linsker proposed the so-called InfoMax principle that a linear sensory system should maximize the mutual information $I(\boldsymbol{x};\boldsymbol{y})$ between the input $\boldsymbol{x}$ and the output $\boldsymbol{y}$. Let $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta}$, where $\boldsymbol{W}$ denotes an $m \times n$ sensing matrix and $\boldsymbol{\eta}$ represents the sensor noise. Recall that $I(\boldsymbol{x};\boldsymbol{y}) = h(\boldsymbol{y}) - h(\boldsymbol{y}|\boldsymbol{x})$, where $h(\boldsymbol{y})$ is the output entropy and where $h(\boldsymbol{y}|\boldsymbol{x})$ denotes the remaining entropy of the output given the input signal and thus merely the entropy of the sensor noise. The noise entropy is constant with respect to $\boldsymbol{W}$, so maximizing $I(\boldsymbol{x};\boldsymbol{y})$ is equivalent to maxmizing $h(\boldsymbol{y})$. We usually need to limit the total power of the sensors (or the squared sum of all entries of $\boldsymbol{W}$ or the trace of $\boldsymbol{W}\boldsymbol{W}^T$) less

than a given budget γ. Therefore, the InfoMax problem can be formalized as

$$\boldsymbol{W}^* = \underset{\boldsymbol{W}_{m\times n}:\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T)=\gamma}{\arg\max} h(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta}). \tag{3.1}$$

in noisy settings. The solution depends on both probability densities of the signal $\boldsymbol{x}$ and the noise $\boldsymbol{\eta}$. We assume that the noise, if any, is Gaussian with $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma_\eta^2 \boldsymbol{I})$. Such a Gaussian noise model is often assumed in the compressed sensing literature (e.g., see [75, 122, 146, 34, 125, 63, 144, 5, 1]), while sometimes more complicated noise models are also considered (e.g., bounded noise [143], Poisson noise [112]).

In Chapter 2, we found the solutions of Problem (3.1) for some special cases. A case is when the signal $\boldsymbol{x}$ is bandwise i.i.d. and the measurement is noiseless (i.e. $\sigma_\eta = 0$). Then, the InfoMax solution can be shown to be bandwise random and the number of projections per band can be determined by solving a convex program. The solution depends on how non-Gaussian the signal is and how fast the variance falls.

3.3 Informative Sensing for Natural Images

The nonuniform-density bandwise projections can be intuitively understood, in relation to the multi-resolution property of natural images. Consider a Laplacian pyramid [24], illustrated in Figure 3.1, where each level represents a bandpass image consisting of edges at a certain scale. The multi-scale edges form a set of "independent" components for natural images [19]. In general, the loss of coarse-scale edges results in larger ℓ_2-error than that of fine-scale edges because of the difference in the power (or variance) carried by the edges. Therefore, we need to allocate sensors to coarse-scale edges with more priority. On the other hand, a coarse-scale edge image is still sparse; its information may be faithfully measured with fewer sensors than its dimension (from the classical results of CS). If we have already well captured the coarse-scale edges and still have extra sensors, then we should desirably spend some for the next-scale edges. Each scale edge image is remarkably white and approximable with an i.i.d. model. Therefore, we use random projections in each scale, overall forming a set of bandwise random projections with nonuniform density per

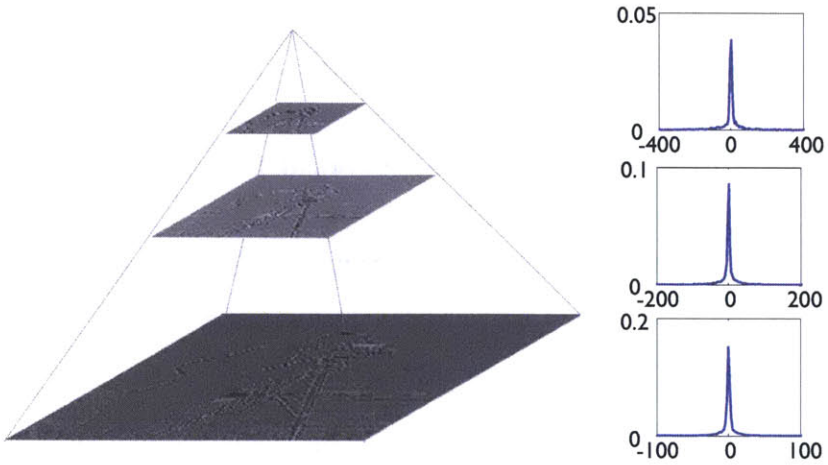


Figure 3.1: Octave Laplacian image pyramid. Shown are the bandpass images of *Camera Man* (left) and the wavelet coefficient distributions of the bandpass images (right). Each bandpass image consists of edges at a certain scale. The distribution of wavelet coefficients is peaked at zero and heavy-tailed. The shape of the distribution is remarkably similar for all different scales, while the standard deviation approximately grows to double with one-level upward in the pyramid.

band.

3.3.1 Mathematical Review

We assume that natural images consists of L independent bands. Let us focus on a particular band s, where the elements (i.e. edges) in the signal $\boldsymbol{x}_s$ are assumed to be i.i.d. We denote the common variance by σ_s^2 and the common negentropy by J_s. The negentropy denotes the non-Gaussianity of the signal elements, by measuring the Kullback-Leibler divergence between their density and a Gaussian with the same first- and second-order statistics. We also denote the dimension of the sth band by n_s, the allotted number of projections by m_s, and the allocated power by γ_s, respectively.

The following is a key result (from the previous chapter) of which we will make use: Let $\boldsymbol{W}_s$ denote an $m_s \times n_s$ random matrix, normalized to satisfy $\boldsymbol{W}_s \boldsymbol{W}_s^T = (\gamma_s/m_s)\boldsymbol{I}$.[2] Then, the entropy of the bandwise random measurement $\boldsymbol{y}_s = \boldsymbol{W}_s \boldsymbol{x}_s$ is approximately

[2]More formally, the normalization refers to $\boldsymbol{W}_s = \sqrt{\gamma_s/m_s}(\boldsymbol{H}\boldsymbol{H}^T)^{-\frac{1}{2}}\boldsymbol{H}$, where $\boldsymbol{H}$ is a random matrix of the same size, e.g., with $H_{ij} \sim \text{iid}, \mathcal{N}(0,1)$.

given, in the asymptotic condition $n_s \to \infty$, by

$$h(\boldsymbol{W}_s \boldsymbol{x}_s) \approx \frac{m_s}{2} \log(2\pi e \sigma_s^2 \gamma_s / m_s) - m_s J_s \beta_s^3 \tag{3.2}$$

where $\beta_s \triangleq m_s/n_s$. Equation (3.2) is an approximation based on Jones and Sibson's work [84]. Note that, in (3.2), we have not considered the measurement noise. We will restrict our attention to noiseless measurements until Section 3.4.

Due to the independence assumption among the bands, the total entropy simply becomes $h = \sum_{s=1}^{L} h(\boldsymbol{y}_s)$. Our objective is to maximize h with respect to m_s's and γ_s's, subject to $\sum_{s=1}^{L} m_s = m$ and $\sum_{s=1}^{L} \gamma_s = \gamma$. In fact, if there is no noise, determining γ_s's does not make a separate issue. Given any m_s's, the first-order condition that $\partial/\partial\gamma_s[h - \xi(\sum_s \gamma_s - \gamma)] = 0$, with a Lagrange multiplier ξ, requires that $\gamma_s = m_s/(2\xi)$ for all s. By the constraint $\sum_{s=1}^{L} \gamma_s = \gamma$, the Lagrange multiplier ξ must be equal to $m/(2\gamma)$, and therefore $\gamma_s = \gamma m_s/m$ for all s. We will shortly present an efficient scheme to determine m_s (and thus β_s), which is conceptually and algorithmically simpler than the one previously given in Chapter 2.

3.3.2 Optimal Profile of Measurement Density

Here, we consider how many samples should be taken in each band. If we have found the optimal numbers $m_1, \ldots, m_L$, it will be impossible to further increase the total entropy by reassigning a sensor which has been given to one band (the sth band), to another (the tth band). This local optimum condition is written mathematically as below: for any $s \neq t$,

$$h(y_{s,1}, \ldots, y_{s,m_s}) + h(y_{t,1}, \ldots, y_{t,m_t}) \;\geq\; h(y_{s,1}, \ldots, y_{s,m_s-1}) + h(y_{t,1}, \ldots, y_{t,m_t+1})$$

or, by rearrangement,

$$\Delta h_{s,m_s} \geq \Delta h_{t,m_t+1} \tag{3.3}$$

if we define

$$\Delta h_{s,j} \triangleq h(y_{s,j}|y_{s,1},\ldots,y_{s,j-1}) \tag{3.4}$$

$$= h(y_{s,1},\ldots,y_{s,j}) - h(y_{s,1},\ldots,y_{s,j-1}) \tag{3.5}$$

$$= \frac{1}{2}\log(2\pi e\sigma_s^2\gamma/m) - \frac{J_s}{n_s^3}(j^4-(j-1)^4). \tag{3.6}$$

We call $\Delta h_{s,j}$ the *net capacity* of the jth sensor, given all previous $(j-1)$ sensors, in the sth band. It is a decreasing function of j within any fixed s. From these facts, the following observation regarding how to determine the optimal number of sensors in each band comes rather straightforwardly (see Figure 3.2 for illustration):

Observation 3.1. For bandwise i.i.d. signals, the optimal number of measurements per band can be found by the following two-step algorithm: (i) evaluate $\Delta h_{s,j}$'s for all $j = 1,\ldots,n_s$, $s = 1,\ldots,L$, (ii) select (s,j)'s associated with m highest values of Δh. The number of measurements for the tth band is determined by the selected number of (s,j)'s with $s = t$.

Proof. See Section 3.7.1.1. □

In Equation (3.6), Δh's are strongly dependent on the signal statistics $\{\sigma_s^2, n_s, J_s\}_{s=1}^L$; so is the optimal profile of the measurement density.

3.3.3 Implementation and Examples

Suppose that images have dimension $\sqrt{n}\times\sqrt{n}$. To implement the bandwise random measurements, we conduct the band decomposition in DCT domain, as illustrated in Figure 3.3. Each DCT kernel in the sth band B_s represents a specific linear combination of the wavelets (e.g., see Figure 3.3, right) that lie in the frequencies between

$$\frac{2^{s-2}f_o}{\sqrt{n}} \le \sqrt{f_x^2+f_y^2} < \frac{2^{s-1}f_o}{\sqrt{n}} \tag{3.7}$$

where f_o denotes the sampling frequency in both directions. If we select a band in DCT domain, the wavelets at the specific scale are still mixed together, but the wavelets at differ-

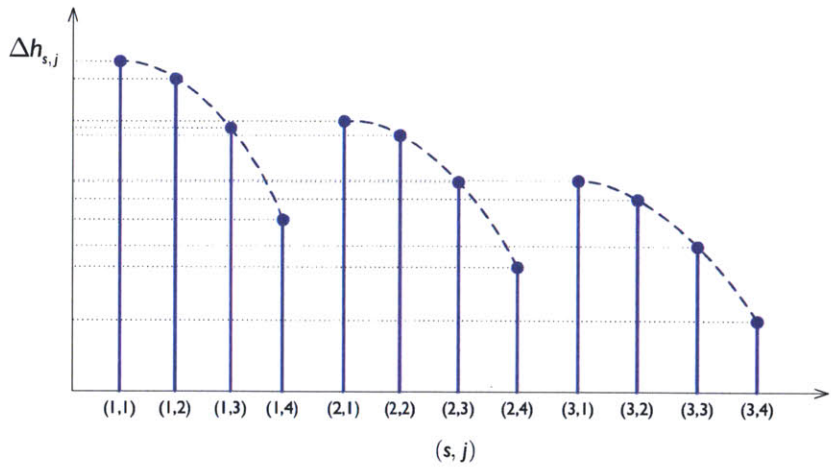


Figure 3.2: Net capacity of bandwise random projections. This is an exemplar plot, where we have drawn the net capacity of only first four sensors in each band (the number of bands is three). Note that the net capacity is a bandwise decreasing function. We should include sensors in the decreasing order of $\Delta h_{s,j}$, i.e., $(1,1)$, $(1,2)$, $(2,1)$, $(1,3)$, $(2,2)$, $(3,1)$, $(2,3)$, $(3,2)$, $(1,4)$, etc., until we use up the given budget. If the total budget m is equal to six, the optimal numbers will be $(m_1, m_2, m_3) = (3,2,1)$.

ent scales are sifted out. Then, the bandwise random multiplexing of the DCT coefficients is simply equivalent to the bandwise random multiplexing of the wavelet coefficients.

For the random multiplexing, we use a subset of noiselets [41], binary-valued pseudo-random basis, following Romberg and Candès [115, 28]. There exists a fast algorithm for the noiselet transform, which makes the computer simulation efficient.

Let us consider two images, *Camera Man* and *Einstein*, shown in Figure 3.4. The standard deviation of the wavelet coefficients notably falls to a half, with the increase of s by one (see Figure 3.1; also refer to [123, 120] for the power spectral statistics). And, in a single image, the distribution is very much alike for all different bands if the standard deviation is normalized (see Figure 3.1; also refer to [155] for the scale invariance). However, the shape of the distribution is different for different images. In Figure 3.4 (middle row), we have plotted the normalized density of the wavelet coefficients of Camera Man and Einstein. Note that Camera Man has more non-Gaussianity than Einstein ($J_s \approx 0.95$ for Camera Man and $J_s \approx 0.45$ for Einstein, in terms of negentropy), which is due to the "simple" (piecewise smooth) content of the Camera Man. On the bottom, we provide the net capacity of bandwise random projections for the two cases. Given m, we should opti-

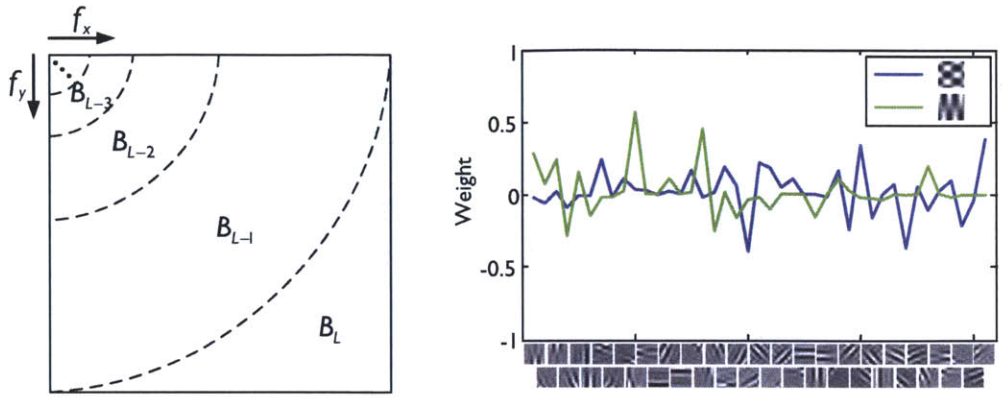


Figure 3.3: Illustration of the band decomposition in spatial frequency domain, where f_x and f_y denote the horizontal and vertical spatial frequency, respectively (left). If we denote, by f_o, the discrete sampling frequency in both directions, the sth band consists of the wavelets which lie in the spatial frequencies between $2^{s-2} f_o/\sqrt{n} \leq \sqrt{f_x^2 + f_y^2} < 2^{s-1} f_o/\sqrt{n}$. The band decomposition may be conducted in DCT domain. Each DCT kernel in B_s is a specific linear combination of the wavelets in B_s. On the right, we show, for example, how a couple of DCT kernels can be represented as linear combinations of the wavelets of the same band. Here, the wavelets have been found by fastICA [79] on a number of image samples bandpassed in DCT domain.

mally determine m_s's as described in Observation 3.1. The optimal density profile varies according to the value of J_s. As seen in Figure 3.4 (bottom), with a small value of J_s, low-frequency bands are favored far more than with a large value of J_s. The smaller value of J_s implies the higher entropy of the normalized image content. Qualitatively speaking, with reference to the image pyramid we considered at the beginning of this section, generally we need more sensors at each scale to capture the "complex" content of the image; few extra sensors will remain, for fine-scales, once after we have used up most sensors at coarser-scales.

3.4 Noisy Measurements

In practice, measurements tend to be corrupted by noises whether the noise level is high or low. If noise is involved, the biggest change in InfoMax is that the maximum entropy may not always be achievable with a tight frame matrix any more (cf. Lemma 2.2). This may be easy to understand by first reviewing Linsker's results for Gaussian signals.

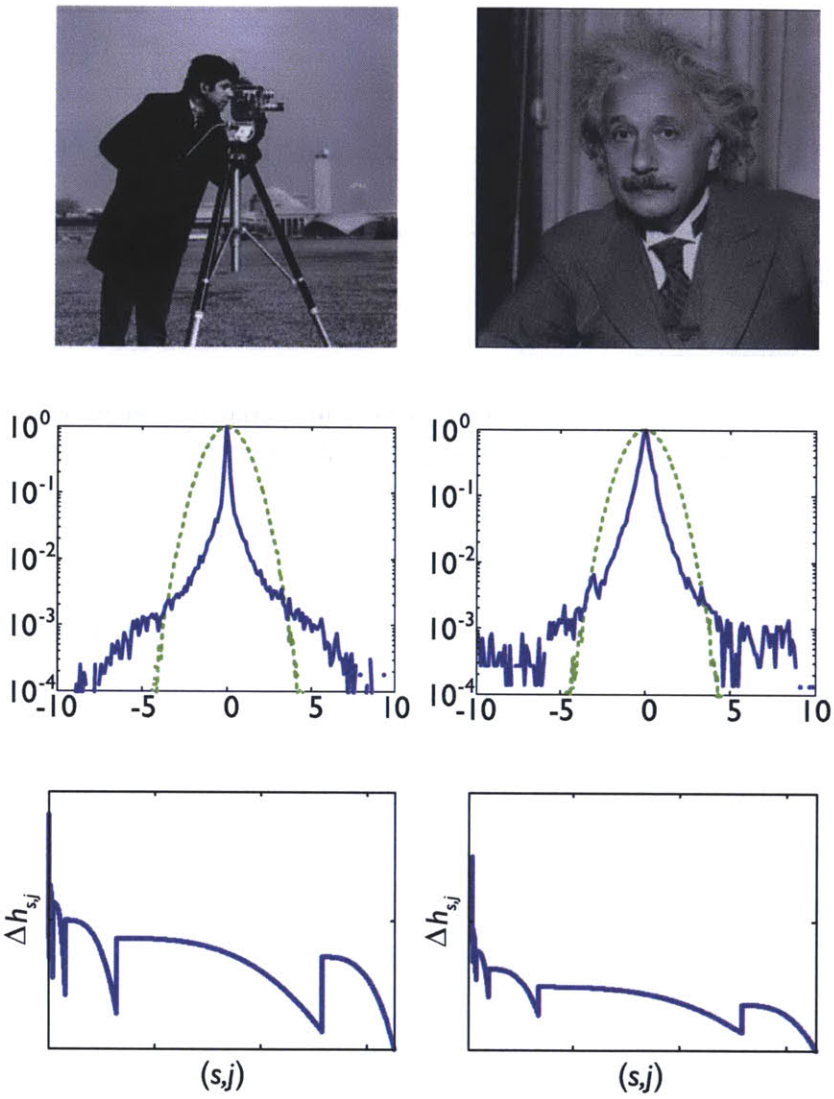


Figure 3.4: Two image examples, Camera Man (left) and Einstein (right). Top: image. Middle: normalized density of wavelet coefficients (vertical axis in log scale). For reference, a Gaussian density with the same variance is plotted together (dashed green). Note that Camera Man has more non-Gaussianity than Einstein. Bottom: Net capacity plot of the bandwise random projection for each case. Note that the profile is strongly dependent on the degree of non-Gaussianity.

Lemma 3.2 (Linsker [94]). Let $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta}$, where $\boldsymbol{W}$ is an $m \times n$ matrix satisfying $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = \gamma$. If $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$ with $\sigma_1 \geq \cdots \geq \sigma_n$, and if $\boldsymbol{\eta} \sim \mathcal{N}(0, \sigma_\eta^2 \boldsymbol{I})$, an InfoMax optimal matrix $\boldsymbol{W}$, which maximizes $h(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta})$ under the power budget constraint, consists of m major principal components of $\boldsymbol{x}$ in its rows.

Furthermore, the power allocated to the ith principal component is determined by

$$p_i^2 = \max(0, 1/\xi - \sigma_\eta^2/\sigma_i^2),\ i = 1, \ldots, m \tag{3.8}$$

where ξ is chosen to satisfy $\sum_i p_i^2 = \gamma$.

Proof. See Section 3.7.1.2 for our own proof, which is more rigorous than provided in [94]. □

Linsker explained the solution (3.8) with a so-called "water-filling" analogy: If one plots σ_η^2/σ_i^2 versus $i = 1, \ldots, m$, then p_i^2 is the depth of water at i when one pours into the vessel defined by the σ_η^2/σ_i^2 curve at a total quantity of water that corresponds to $\sum_i p_i^2 = \gamma$ and brings the water level to $1/\xi$. Lemma 3.2 tells that without noise we should use equi-power sensors in measuring m major principal components for Gaussian signals but that with noise we should allocate more power to more significant components. The power redistribution makes intuitive sense in two aspects: (i) we may want to protect more important signals from noise, (ii) minor principal components are weak signals vulnerable to noise; there is little hope to denoise them without excessive expense of sensor power, so we had better give them up.

It is obvious that Linsker's results are not directly applicable to natural images because of their non-Gaussianity. Next we consider non-Gaussian signals but assume, for the time being, that x_i's are i.i.d. We have the following lemma:

Lemma 3.3. Suppose $\boldsymbol{x} = (x_1, \ldots, x_n)$, where x_i's are i.i.d., with variance σ_x^2 and negentropy J_x. Let $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta}$, where $\boldsymbol{W}$ is an $m \times n$ matrix satisfying $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = \gamma$ and $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\eta^2 \boldsymbol{I})$. In the asymptotic condition ($n \to \infty$), random matrices, followed by the normalization so that $\boldsymbol{W}\boldsymbol{W}^T = (\gamma/m)\boldsymbol{I}$, maximize $h(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta})$ under the power budget constraint. Furthermore, the power allocated to each random projection should be equal.

Proof. See Section 3.7.1.3. □

Then, what can we say about natural images which are modeled as a bandwise i.i.d. signal? When $\sigma_\eta = 0$ (no noise), bandwise random projections were InfoMax optimal in the

asymptotic setting. Such bandwise projections had a physical meaning with regard to the multi-resolution representation of natural images, as given at the beginning of Section 3.3. That interpretation is valid even with noise involved, which provides a good reason for us to believe that bandwise random projections are still asymptotically InfoMax optimal in noisy measurement settings.

Observation 3.4. We consider the problem of finding an $m \times n$ matrix $\boldsymbol{W}$ that, for a bandwise i.i.d. signal $\boldsymbol{x}$ and for a Gaussian noise $\boldsymbol{\eta} \sim \mathcal{N}(\boldsymbol{0}, \sigma_\eta^2 \boldsymbol{I})$, maximizes $h(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta})$ while satisfying $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = \gamma$. If we restrict our attention to all *bandwise* projections and if $n_s \to \infty$ for all s, the solution will be bandwise random projections with nonuniform measurement density per band. Further, the power allocated to each sensor in the same band must be identical.

Proof. Given the bandwise restriction, we should necessarily optimize every band for the optimal solution. Lemma 3.3 applies to each single band where the signals are i.i.d. Therefore, this observation immediately follows. □

Surely, the details are subject to change. The measurement density and power distribution should be determined in dependence of the noise level.

3.4.1 Noise Effect

In this section, we will see how the noise affects the entropy of the measurement that should be maximized in the InfoMax framework. We consider band by band. Recall that we denote the number of sensors by m_s and the total power by γ_s assigned to the sth band, respectively. Let $\boldsymbol{W}_s$ be an $m_s \times n_s$ random matrix, normalized to satisfy $\boldsymbol{W}_s \boldsymbol{W}_s^T = \frac{\gamma_s}{m_s} \boldsymbol{I}$. The elements in the sth band signal $\boldsymbol{x}_s$ are i.i.d., with variance σ_s^2 and negentropy J_s.

We can rewrite the measurement $\boldsymbol{y}_s$ as $\boldsymbol{y}_s = \boldsymbol{W}_s \boldsymbol{x}_s + \boldsymbol{\eta}_s = \boldsymbol{W}_s \boldsymbol{x}'_s$, where $\boldsymbol{x}'_s = \boldsymbol{x}_s + \boldsymbol{\eta}'_s$ with $\boldsymbol{\eta}'_s \sim \mathcal{N}(\boldsymbol{0}, \frac{m_s}{\gamma_s} \sigma_\eta^2 \boldsymbol{I})$. The elements in the pre-corrupted signal $\boldsymbol{x}'_s$ are also i.i.d. The variance is easily obtained to be $\frac{m_s}{\gamma_s} \sigma_\eta^2 + \sigma_s^2$. To see how the negentropy of the corrupted signal behaves according to the noise level, we will consider that each signal element has a generalized Gaussian density with mean zero, variance σ_s^2, and shape parameter r (see

Appendix A). If $r < 2$, the distribution is heavy-tailed, with the degree determined by r (the smaller, the heavier). Without noise, the negentropy is known to be $J_s = \frac{1}{2}\log\left(\frac{\pi r^2 \Gamma(\frac{3}{r})}{2\Gamma^3(\frac{1}{r})}\right) + \frac{1}{2} - \frac{1}{r}$. The generalized Gaussian has widely been used for modeling the distribution of the wavelet coefficients [130, 21, 155].

We have computed the negentropy when a generalized Gaussian density is corrupted by a Gaussian noise for various SNRs. The corrupted density has been obtained by convolving the noise density, i.e., Gaussian, with the original signal density. Subsequently, the negentropy has been numerically computed. A couple of results, each for $r = 0.32$ and $r = 0.49$, are shown in Figure 3.5, where the negentropy is plotted as a function of SNR, i.e., $J_s(\mathsf{snr})$. The negentropy monotonically increases with SNR, from zero, i.e., perfect Gaussianity (at $\mathsf{snr} = 0$) to J_s (at $\mathsf{snr} = \infty$). The first- and second-order derivatives are also shown in Figure 3.5. We can see that the negentropy plot is strictly concave because $\frac{d^2 J_s(\mathsf{snr})}{d\mathsf{snr}^2}$ is negative for all range of SNRs.

In our case, $\mathsf{snr} = \frac{\gamma_s \sigma_s^2}{m_s \sigma_\eta^2}$. Based on the formula in Equation (3.2), we can write the entropy in the sth band as

$$h_s = \frac{m_s}{2}\log(2\pi e(m_s\sigma_\eta^2/\gamma_s + \sigma_s^2)\gamma_s/m_s) - J_s(\mathsf{snr})\frac{m_s^4}{n_s^3} \tag{3.9}$$

$$= \frac{m_s}{2}\log(2\pi e(\sigma_\eta^2 + \gamma_s\sigma_s^2/m_s)) - J_s(\gamma_s\sigma_s^2/m_s\sigma_\eta^2)\frac{m_s^4}{n_s^3}. \tag{3.10}$$

Summing (3.10) over all the bands, we finally obtain

$$h = \sum_{s=1}^{L} \frac{m_s}{2}\log(2\pi e(\sigma_\eta^2 + \gamma_s\sigma_s^2/m_s)) - J_s(\gamma_s\sigma_s^2/m_s\sigma_\eta^2)\frac{m_s^4}{n_s^3}. \tag{3.11}$$

3.4.2 Optimization

Our objective is to find $\{m_s, \gamma_s\}_{s=1}^L$ that maximizes h in Equation (3.11) subject to two constraints, $\sum_s m_s = m$ and $\sum_s \gamma_s = \gamma$. The solution seems to be very difficult to analytically compute. We rely on iterative optimizations of m_s, and γ_s/m_s, given the other.

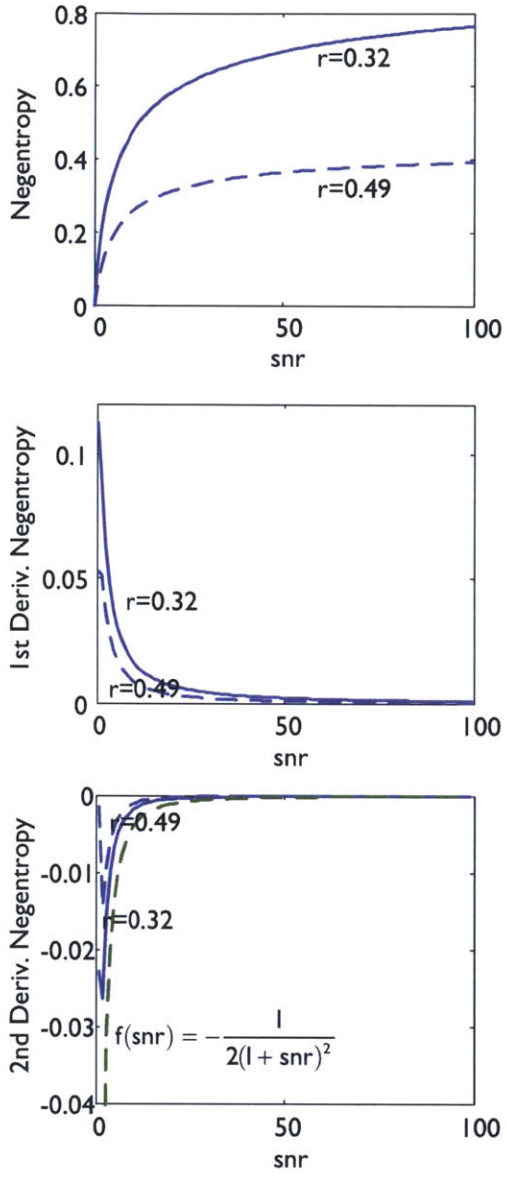


Figure 3.5: Variation of the negentropy by noise corruption. The signal is assumed to follow a generalized Gaussian density $p(x) \propto e^{-|x/C|^r}$ (see Appendix A) and the noise is assumed to be Gaussian. Shown are the functional plot of the negentropy (top), and its first-order (middle) and second-order (bottom) derivatives, versus SNR. On the bottom, the dashed green curve is $f(\mathsf{snr}) = -\frac{1}{2(1+\mathsf{snr})^2}$. Note that the second derivatives, for both values of r, do not fall below the curve.

3.4.2.1 Optimal measurement density, given power distribution

Given γ_s/m_s's, the optimization of the measurement density comes almost for free. If γ_s/m_s's are fixed, Equation (3.11) is simply in the form we had in the noiseless setting.

Thus, with the net capacity Δh, computed by

$$\Delta h_{s,j} = h(y_{s,1}, \ldots, y_{s,j}) - h(y_{s,1}, \ldots, y_{s,j-1}) \tag{3.12}$$

$$= \frac{1}{2}\log(2\pi e(\sigma_\eta^2 + \gamma_s\sigma_s^2/m_s)) - \frac{J_s(\gamma_s\sigma_s^2/m_s\sigma_\eta^2)}{n_s^3}(j^4 - (j-1)^4), \tag{3.13}$$

we can simply use the algorithm described in Observation 3.1 to obtain the optimal m_s's.

3.4.2.2 Optimal power distribution, given measurement density

Next, suppose that m_s's are given. To maximize h in Equation (3.11), with respect to γ_s/m_s under the power budget constraint $\sum_s \gamma_s = \gamma$, the first-order condition requires that, for a Lagrange multiplier ξ,

$$\underbrace{\frac{1}{2(1+\mathsf{snr}_s)} - \beta_s^3 \frac{d}{d\mathsf{snr}_s} J_s(\mathsf{snr}_s)}_{\triangleq g(\mathsf{snr}_s)} = \frac{\xi\sigma_\eta^2}{\sigma_s^2} \tag{3.14}$$

if $\gamma_s > 0$, where $\mathsf{snr}_s \triangleq \gamma_s\sigma_s^2/m_s\sigma_\eta^2$ and $\beta_s \triangleq m_s/n_s$; otherwise, $\gamma_s = 0$. The Lagrange multiplier ξ is determined to satisfy the power budget constraint, that is, $\sum_{s=1}^{L} \gamma_s = \gamma$. For Gaussian signals ($J_s \equiv 0$), Equation (3.14) gives

$$\mathsf{snr}_s = \max\left(0, \frac{\sigma_s^2}{2\xi\sigma_\eta^2} - 1\right) \quad \text{or} \quad \gamma_s/m_s = \max\left(0, \frac{1}{2\xi} - \frac{\sigma_\eta^2}{\sigma_s^2}\right), \tag{3.15}$$

which is simply the same as Equation (3.8). For non-Gaussian signals, the solution of Equation (3.14) cannot be explicitly written but can still be computed numerically. Note that the function $g(\cdot)$, defined in Equation (3.14), tends to be a strictly decreasing function given a measurement density β_s. For the generalized Gaussian either with $r = 0.32$ or with $r = 0.49$, we illustrated that

$$\frac{d^2}{d\mathsf{snr}_s^2} J_s(\mathsf{snr}_s) > -\frac{1}{2(1+\mathsf{snr}_s)^2} \geq -\frac{1}{2\beta_s^3(1+\mathsf{snr}_s)^2} \tag{3.16}$$

on the bottom of Figure 3.5. Hence,

$$\frac{d}{d\mathsf{snr}_s} g(\mathsf{snr}_s) = -\frac{1}{2(1+\mathsf{snr}_s)^2} - \beta_s^3 \frac{d^2}{d\mathsf{snr}_s^2} J_s(\mathsf{snr}_s) < 0. \tag{3.17}$$

It is numerically found that the strict decreasing property of $g(\cdot)$ holds if $r > 0.2$. This implies that the solution of (3.14) is unique and that it increases with $1/\xi$. Such a feature enables us to efficiently find the correct value of ξ, for example, with binary search. Most natural images even with very simple texture satisfy the condition that $r > 0.2$. The average value computed from a subset of Berkeley image dataset [101] is approximately 0.4 (see Section 3.5) or equivalently $J_s \approx 0.65$.

3.5 Experimental Results

In this section, we provide a set of results from image recovery experiments. We compare the performance of bandwise random projections against other kinds of projections in terms of peak-signal-to-noise ratio (PSNR), defined as

$$\mathsf{PSNR} = 10 \log_{10} \frac{255^2}{\frac{1}{n}\sum_i (x_i - \widehat{x}_i)^2} \quad \text{(dB)} \tag{3.18}$$

where $\widehat{\boldsymbol{x}} = (\widehat{x}_1, \ldots, \widehat{x}_n)$ denotes the reconstructed image (with vectorization).

Image recovery is based on regularization by total variation (TV),[3] as in many other studies (e.g., [29, 115, 20]). In the noiseless setting, we estimate $\widehat{\boldsymbol{x}}^*$ by

$$\widehat{\boldsymbol{x}}^* = \arg\min_{\widehat{\boldsymbol{x}}} \sum_{i,j} \|\nabla \mathcal{I}_{ij}(\widehat{\boldsymbol{x}})\|, \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{W}\widehat{\boldsymbol{x}} \tag{3.19}$$

where $\mathcal{I}(\widehat{\boldsymbol{x}})$ denotes the 2D matrix representation of $\widehat{\boldsymbol{x}}$. The TV regularization is known to perform better than the ℓ_1-norm regularization on the wavelet basis, avoiding high-frequency artifacts [115, 20]. This is partly because the TV sparsity model can account for the power-law spectrum.

[3] In Chapter 2, our recovery was based on the minimum mean-squared error (MMSE) estimate in all experiments. Here, we are doing TV based recovery as an approximation.

The first set of experiments was conducted on Camera Man and Einstein in the noiseless condition. In Figure 3.6, we compared the PSNR performance among four different

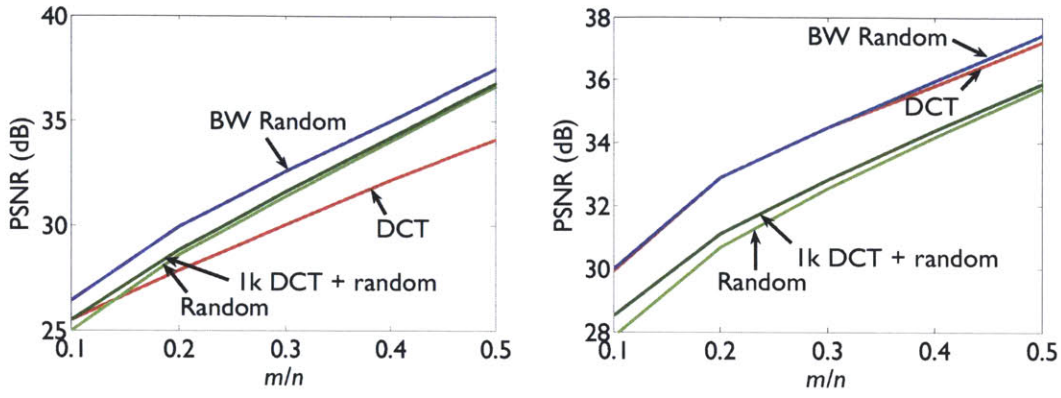


Figure 3.6: Image reconstruction results for Camera Man (left) and Einstein (right). Shown are the plots of PSNR versus the measurement rate. Compared projection schemes are low-pass DCT (red), random (light green), 1k DCT + random (dark green), and the nonuniform-density bandwise random (blue) projections. For recovery, TV regularization has been commonly used (see Equation 3.19).

projection schemes: low-pass DCT coefficients in zig-zag order (red), which approximates the PCA projection (see Footnote 1), random projections (light green), Romberg's method [115] which uses 1,000 DCT coefficients in zig-zag order plus the remaining number of random projections (dark green), and the nonuniform-density bandwise random projections (blue).

Perhaps unsurprisingly, the dark green curve (1k DCT + random projections) is above the light green (pure random projections) in every case. However, if we compare the greens (random projections or 1k DCT + random projections) against the red (DCT), their relative performance is completely different, depending on the input image. Recall that Camera Man and Einstein have very different shapes in the distribution of wavelet coefficients (see Figure 3.4, middle row). Camera Man has simple (piecewise smooth) texture. The image content throughout all spatial frequencies is well accommodated by a moderate number of random projections. Meanwhile, the DCT projection wastefully allocates available sensors to low-frequency content, which could be captured with even fewer sensors, so all high-frequency content is irrecoverably thrown away. On the other hand, Einstein has complex

texture. We should use almost all sensors for low-frequency bands. Otherwise, even low-resolution image is not faithfully recoverable.

For the two images, the net capacity diagrams are different as shown in Figure 3.4 (bottom row). In this set of experiments, we used different density profiles, each according to their respective net capacity diagram, in applying the bandwise random projections to the images. Then, the bandwise random projections (blue) outperform all other projections in most regions, while showing nearly equal performance as the DCT projection for Einstein.

Figure 3.7 shows the results of the reconstruction of Camera Man from five thousand measurements (about 7.6% of the original dimension), which clearly portray the behavioral characteristics of each measurement scheme. The image reconstructed from DCT projection almost loses the mid/high-frequency content and looks piecewise constant. In contrast, random projections, and even Romberg's method, produce pasty images. When the total number of measurements is seriously restricted, a success in recovering the high-frequency details only comes with a sacrifice of the low/mid-frequency content which is more important. Last, the bandwise random projection gives up the high-frequency content but faithfully preserves the low/mid-frequency content instead.

In the first set of experiments, we have applied different density profiles to Camera Man and Einstein, according to the complexity of the input image. If the complexity of the input image is not fixed, we may have to use the average complexity estimated from an aggregated set of the normalized wavelet coefficients. In the next experiments, ten 256×256 images, shown in Figure 3.8, are used. They are from Berkeley dataset [101] and have various complexities in terms of negentropy. Using a common set of bandwise random projections, tuned to $J_s \approx 0.65$ (mean value), for the entire set of images, we have obtained the results as shown in Table 3.1, where we use $m = 20,000$ (approximately 30% of the original dimension). As seen in the table, the bandwise random projection performs best for most images (>1dB better than the other projections on average) while worse than the DCT projection for the last two images (Im_9, Im_{10}). As aforementioned, the DCT projection is nearly optimal for fairly complex images, Im_7–Im_{10}. If we tuned the bandwise random projection to $J_s \approx 0.35$, it would give similar performance for the last two images as the DCT projection.

DCT (24.81dB)

Random (23.78dB)

1k DCT + random (24.41dB)

Bandwise random (25.73 dB)

Figure 3.7: Image reconstruction results for Camera Man. In this experiment, the number of measurements is restricted to 5k, which corresponds to 7.6% of the original dimension.

We have conducted similar sets of experiments also in noisy settings. In this case, we incorporate denoising into TV-based recovery. In Figure 3.9, we compared the performance of five projection schemes (with $m = 20,000$) on Camera Man and Einstein at various noise levels. Four of them are exactly what we have used in the first set of experiments: low-pass DCT (red), random (light green), 1,000 DCT plus 19,000 random (dark green) and nonuniform-density bandwise random (solid blue) projections. For the density profile of

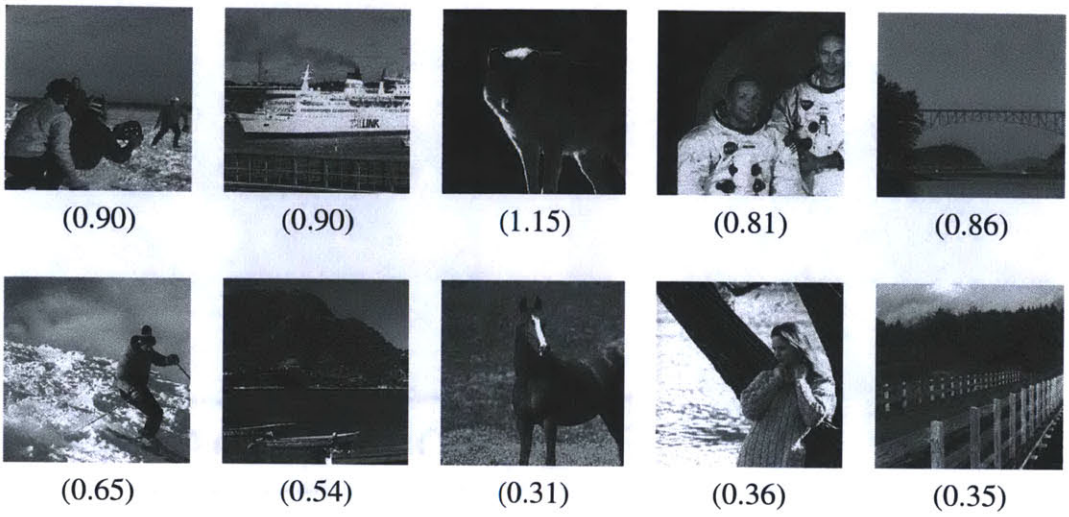


Figure 3.8: Ten images from Berkeley dataset [101], each cropped to 256×256. They are numbered Im_1–Im_{10} from left to right, top to bottom. Each parenthesized number denotes the value of J_s, the negentropy of the wavelet coefficients of the above image.

Table 3.1: The PSNR performance of image reconstruction results with $m = 20,000$. †For all images, the same set of bandwise random projections, tuned to $J_s \approx 0.65$, has been commonly used.

Method	Im_1	Im_2	Im_3	Im_4	Im_5	Im_6	Im_7	Im_8	Im_9	Im_{10}
DCT	32.74	26.81	36.94	32.45	35.69	29.77	29.61	33.23	**29.87**	**29.86**
Random	32.69	26.09	39.34	32.58	36.89	29.10	28.34	31.70	27.61	27.73
1k DCT + random	32.91	26.32	39.54	32.86	37.08	29.32	28.53	31.93	27.84	27.95
BW random†	**34.41**	**27.77**	**40.16**	**34.29**	**38.20**	**30.99**	**29.86**	**33.62**	29.53	29.43

the bandwise random projections, we have used $J_s \approx 0.95$ for Camera Man and $J_s \approx 0.45$ for Einstein, respectively. For a short while, let us consider the four only (solid curves), to see how each scheme, without noise adaptation, degrades with the level of noise. For Camera Man, even with small noise (e.g., $\sigma_\eta > 5$), the DCT projection (red) shows a better performance than random projections (light green) as well as Romberg's method (dark green), while it has performed worse in the noiseless setting (i.e., $\sigma_\eta = 0$). If $\sigma_\eta \geq 15$, the DCT projection also outperforms the bandwise random projection (solid blue). Such a good performance of the DCT projection is found also in Einstein (see Figure 3.9, right), where the DCT projection works best throughout almost all noise levels.

The remaining scheme is the noise-adapted version of the nonuniform-density bandwise

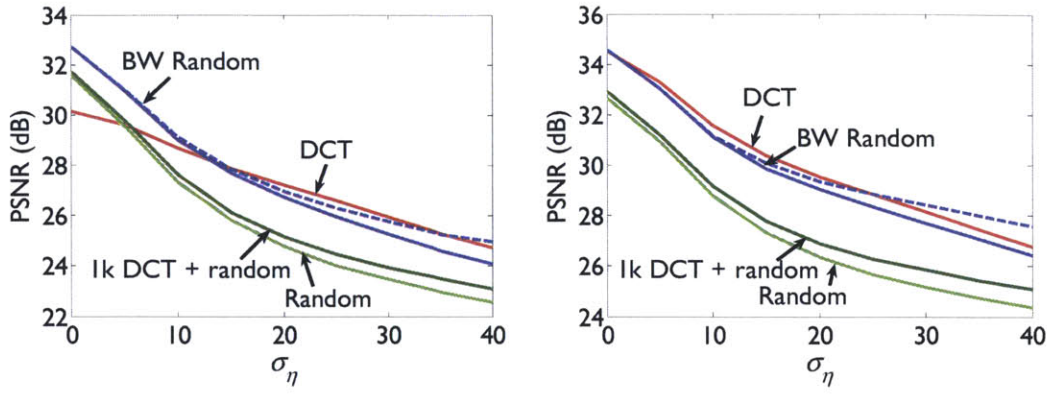


Figure 3.9: Image reconstruction results for Camera Man (left) and Einstein (right) in noisy settings. Shown are the plots of PSNR versus the noise variance. Compared projection schemes are low-pass DCT (red), random (light green), 1k DCT + random (dark green), and nonuniform-density bandwise random (solid blue), plus noise-adapted bandwise random (dashed blue) projections. For recovery, TV regularization has been commonly used. The number of measurements are set to $m = 20,000$.

random projection (dashed blue). We use the iterative optimization method, described in Section 3.4.2, for the noise adaptation (we assume that σ_η is known a priori). The total power is kept to be equal (i.e., $\gamma = m$) for all five projection schemes. The noise adaptation has a bit boosted the performance of the bandwise random projections. It prevents the reconstruction performance from degrading fast, particularly when the noise level is high.

Finally, we have run the same experiments with the images from the Berkeley dataset under two SNR regimes (one with $\sigma_\eta = 5$ and the other with $\sigma_\eta = 30$). In this case, we have used $m = 20,000$ and a fixed value $J_s \approx 0.65$, so a common set of bandwise random projections for all images. The results are summarized in Table 3.2. Even in high SNR regime, the DCT projection outperform pure random or 1k DCT plus random projections (compare with Table 3.1). This demonstrates that the DCT projection is resilient to the measurement noise while random projections are vulnerable. In high SNR regime, the bandwise random projection still remains the best, on the average performance, among the four without noise adaptation. In low SNR regime or if images are complex, the DCT projection tends to outperform all other three.

If the noise level is known at the step of designing the measurement matrix, we may use the noise adaptation techniques (i.e., power redistribution, as well as density profile

Table 3.2: The PSNR performance of image reconstruction results in noisy settings, with $m = 20,000$. †For all images, the same set of bandwise random projections, tuned to $J_s \approx 0.65$, has been commonly used. ‡BW random (NA) refers to the noise-adapted version of the bandwise random projections.

High SNR regime ($\sigma_\eta = 5$)

Method	Im_1	Im_2	Im_3	Im_4	Im_5	Im_6	Im_7	Im_8	Im_9	Im_{10}
DCT	31.83	26.46	35.06	31.56	34.17	29.26	**29.12**	**32.26**	**29.34**	**29.33**
Random	30.47	25.39	34.33	30.36	32.80	27.92	27.47	30.19	26.94	26.99
1k DCT + random	30.69	25.63	34.62	30.68	33.07	28.16	27.72	30.46	27.17	27.22
BW random†	32.35	**27.13**	**35.95**	**32.34**	**34.68**	**29.89**	29.07	32.20	28.86	28.77
BW random (NA)‡	**32.40**	27.06	35.93	32.32	34.63	29.88	29.09	32.22	29.03	28.85

Low SNR regime ($\sigma_\eta = 30$)

Method	Im_1	Im_2	Im_3	Im_4	Im_5	Im_6	Im_7	Im_8	Im_9	Im_{10}
DCT	27.32	**23.25**	30.20	26.86	29.43	25.34	**25.47**	27.37	25.03	**25.15**
Random	24.37	20.88	26.68	23.93	26.60	23.04	23.21	25.04	22.76	22.65
1k DCT + random	25.04	21.26	27.52	24.55	27.39	23.48	23.67	25.71	23.21	23.10
BW random†	26.71	22.57	29.78	26.30	28.98	24.82	24.90	26.94	24.54	24.53
BW random (NA)‡	**27.56**	22.36	**30.48**	**26.95**	**29.76**	**25.37**	24.98	**27.87**	**25.35**	24.97

optimization). The impact of the noise adaptation is not huge but still meaningful in low SNR regime. On average, the noise-adapted bandwise random projections have performed best.

3.6 Discussion

Despite the popularity in the CS framework, random measurements are not universally optimal for every class of sparse signals. Given the input distribution, we can optimize linear measurements so as to minimize the uncertainty of the signal given the measurement. With Shannon's entropy as the uncertainty criterion, this is equivalent to the InfoMax framework. In particular, if the signals are groupwise i.i.d., nonuniform-density groupwise random measurements are known to be asymptotically optimal in such a framework.

In this chapter, we have applied bandwise random measurements to natural images by properly modeling natural images with reference to their well-known statistics. This measurement scheme is more or less similar to the variable-density random Fourier sampling

suggested by Lustig, Donoho, and Pauly [98], who, however, did not find a principled method to determine the density profile. In the InfoMax framework, we introduced the so-called *net capacity* of each measurement and subsequently, based on it, presented an efficient algorithm to optimize the density profile along octave frequency bands. We also showed how the density profile should depend on the second- and higher-order moments of the wavelet coefficients. In the presence of noise, we considered the optimal distribution of power among sensors, which generalizes Linsker's results on Gaussian signals [94], to natural images which are not Gaussian. As experimentally demonstrated, the power distribution makes the measurement robust to noise.

For natural images, sparsity can be attributed to a couple of sources, i.e., variance asymmetry (power-law in spectrum) and non-Gaussianity. Subject to such "source-split" sparsity, a bandwise i.i.d. model is the distribution with the largest entropy, which should be chosen, according to the principle of maximum entropy [82], if no further information is used. The bandwise i.i.d. assumption better describes natural images than the simple sparsity assumption. However, it is not yet the best model for natural images. The wavelet coefficients are known to be actually dependent, for example, in the local neighborhood or along the tree hierarchy [130]. In recent literature of CS, the dependencies are utilized at the recovery side [10, 4]. In the future, they should desirably be exploited for the design of the measurement matrix as well, although finding the InfoMax measurement matrix for a complicated prior may be difficult.

3.7 Appendix to Chapter 3

3.7.1 Proofs

3.7.1.1 Proof of Observation 3.1

For simplicity's sake, we will assume that $\Delta h_{s,j}$'s are all distinct, although the observation does not require such an assumption. Suppose that $m_1, \ldots, m_L$ are the optimal number of sensors per band, satisfying $\sum_{s=1}^{L} m_s = m$. Let us denote, by C_m, the set of (s, j)'s selected by the algorithm. We will specifically prove that, for any s and j, $(s, j) \in C_m$ if

and only if $j \leq m_s$. Then, the claimed observation will immediately follow.

1) If $j \leq m_s$,

$$\Delta h_{s,j} \geq \Delta h_{s,m_s} > \underbrace{\Delta h_{t,m_t+1} > \cdots > h_{t,n_t}}_{\#=n_t-m_t}, \quad \text{for all } t \tag{3.20}$$

because of the inequality (3.3) as well as of the bandwise decreasing property of Δh. Therefore, at least $\sum_{t=1}^{L}(n_t - m_t) = (n - m)$ number of Δh's are smaller than $\Delta h_{s,j}$, which implies that $(s, j) \in C_m$.

2) If $j \geq m_s + 1$,

$$\underbrace{\Delta h_{t,1} > \cdots > \Delta h_{t,m_t}}_{\#=m_t} > \Delta h_{s,m_s+1} \geq h_{s,j}, \quad \text{for all } t, \tag{3.21}$$

which is also due to the inequality (3.3) as well as the bandwise decreasing property of Δh. Therefore, at least $\sum_{t=1}^{L} m_t = m$ number of Δh's are greater than $h_{s,j}$, which means that $(s, j) \notin C_m$.

3.7.1.2 Proof of Lemma 3.2

If both signal and noise are Gaussian, the measurement $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{\eta}$ will also be Gaussian for any $\boldsymbol{W}$, so the entropy of $\boldsymbol{y}$ is simply equal to $h(\boldsymbol{y}) = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\log\det(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T + \sigma_\eta^2\boldsymbol{I})$. The first-order condition, to maximize $h(\boldsymbol{y})$ with respect to $\boldsymbol{W}$ under the power budget constraint $\mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = \gamma$, is

$$(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T + \sigma_\eta^2\boldsymbol{I})^{-1}\boldsymbol{W}\boldsymbol{\Sigma} = \xi_1\boldsymbol{W}, \tag{3.22}$$

where ξ_1 is a Lagrange multiplier. Let $\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ by its singular value decomposition, where $\boldsymbol{U}$ is an orthonormal, square matrix and where $\boldsymbol{\Lambda}$ is a diagonal matrix. Equation (3.22) becomes

$$\boldsymbol{U}(\boldsymbol{\Lambda} + \sigma_\eta^2\boldsymbol{I})^{-1}\boldsymbol{U}^T\boldsymbol{W}\boldsymbol{\Sigma} = \xi_1\boldsymbol{W}. \tag{3.23}$$

Multiplying $(\boldsymbol{\Lambda} + \sigma_\eta^2 \boldsymbol{I})\boldsymbol{U}^T$ in front of both sides in Equation (3.23), we obtain

$$\boldsymbol{U}^T \boldsymbol{W} \boldsymbol{\Sigma} = \xi_1 (\boldsymbol{\Lambda} + \sigma_\eta^2 \boldsymbol{I}) \boldsymbol{U}^T \boldsymbol{W}, \tag{3.24}$$

from which we see that the row vectors of $\boldsymbol{U}^T\boldsymbol{W}$ should be m eigenvectors, or principal components, of $\boldsymbol{\Sigma}$. If we say that the ith row vector is the j_ith principal component (scaled by p_i) of $\boldsymbol{\Sigma}$, the associated eigenvalue will be $\sigma_{j_i}^2$. Then, we may write

$$\boldsymbol{\Lambda} = \boldsymbol{U}^T \boldsymbol{W} \boldsymbol{\Sigma} \boldsymbol{W}^T \boldsymbol{U} = \mathrm{diag}(p_1^2 \sigma_{j_1}^2, \ldots, p_m^2 \sigma_{j_m}^2). \tag{3.25}$$

The power budget constraint is translated, in terms of p_i, into

$$\gamma = \mathrm{Tr}(\boldsymbol{W}\boldsymbol{W}^T) = \mathrm{Tr}(\underbrace{\boldsymbol{U}^T \boldsymbol{W} \boldsymbol{W}^T \boldsymbol{U}}_{=\mathrm{diag}(p_1^2, \ldots, p_m^2)}) = \sum_{i=1}^{m} p_i^2. \tag{3.26}$$

Equation (3.25) is the condition for a local optimum. We may need to combinatorially find the global optimum. Note that $\det(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T + \sigma_\eta^2 \boldsymbol{I}) = \det(\boldsymbol{U}\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T\boldsymbol{U}^T + \sigma_\eta^2 \boldsymbol{I})$ because of the orthonormality of $\boldsymbol{U}$. Then, we can write $h(\boldsymbol{y})$, based on Equation (3.25), as

$$h(\boldsymbol{y}) = \frac{m}{2} \log(2\pi e) + \sum_{i=1}^{m} \frac{1}{2} \log(p_i^2 \sigma_{j_i}^2 + \sigma_\eta^2). \tag{3.27}$$

For any fixed p_i's (we assume that $p_1^2 \geq \cdots \geq p_m^2$ without loss of generality), the entropy h in Equation (3.27) is maximized when $\sigma_{j_1}^2, \ldots, \sigma_{j_m}^2$ are m largest eigenvalues of $\boldsymbol{\Sigma}$ in that order, so we can fix $\sigma_{j_1}^2, \ldots, \sigma_{j_m}^2$ to such eigenvalues. Finally, the first-order condition, to maximize h in Equation (3.27) with respect to p_i under the power budget constraint $\sum_i p_i^2 = \gamma$ (Equation 3.26), gives

$$\frac{p_i \sigma_{j_i}^2}{p_i^2 \sigma_{j_i}^2 + \sigma_\eta^2} = 2\xi_2 p_i \tag{3.28}$$

where ξ_2 is a Lagrange multiplier. Therefore, $p_i = 0$ or $p_i^2 = \frac{1}{2\xi_2} - \frac{\sigma_\eta^2}{\sigma_{j_i}^2}$, which is simply rewritten as $p_i^2 = \max(0, 1/\xi - \sigma_\eta^2/\sigma_{j_i}^2)$. The Lagrange multiplier $\xi = 2\xi_2$ should be determined to satisfy $\sum_i p_i^2 = \gamma$.

3.7.1.3 Proof of Lemma 3.3

The entropy $h(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta})$ can be expressed as the difference, by the negentropy $J(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta})$, from the Gaussian entropy, i.e., $h(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta}) = \frac{m}{2}\log(2\pi e) + \frac{1}{2}\log\det(\sigma_x^2\boldsymbol{W}\boldsymbol{W}^T + \sigma_\eta^2\boldsymbol{I}) - J(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta})$. By Jensen's inequality, the log determinant term is upper-bounded by

$$\frac{1}{m}\log\det(\sigma_x^2\boldsymbol{W}\boldsymbol{W}^T + \sigma_\eta^2\boldsymbol{I}) \leq \log\left(\frac{1}{m}\mathrm{Tr}(\sigma_x^2\boldsymbol{W}\boldsymbol{W}^T + \sigma_\eta^2\boldsymbol{I})\right) = \log\left(\frac{\sigma_x^2\gamma}{m} + \sigma_\eta^2\right), \tag{3.29}$$

where the equality holds if and only if all the eigenvalues of $(\sigma_x^2\boldsymbol{W}\boldsymbol{W}^T + \sigma_\eta^2\boldsymbol{I})$ are equal to $\frac{\sigma_x^2\gamma}{m} + \sigma_\eta^2$, that is, $\sigma_x^2\boldsymbol{W}\boldsymbol{W}^T + \sigma_\eta^2\boldsymbol{I} = (\frac{\sigma_x^2\gamma}{m} + \sigma_\eta^2)\boldsymbol{I}$ or $\boldsymbol{W}\boldsymbol{W}^T = \frac{\gamma}{m}\boldsymbol{I}$. On the other hand, the negentropy $J(\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta})$ is minimized when $\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta}$ is closest to a Gaussian distribution. By taking a normalized random matrix for $\boldsymbol{W}$, more precisely by taking $\boldsymbol{W} = \sqrt{\frac{\gamma}{m}}(\boldsymbol{H}\boldsymbol{H}^T)^{-\frac{1}{2}}\boldsymbol{H}$ with $H_{ij} \sim \mathcal{N}(0,1)$, we can make $\boldsymbol{W}\boldsymbol{x}$ as Gaussian as possible (according to Chapter 2). The maximal Gaussianity is also preserved for $\boldsymbol{W}\boldsymbol{x}+\boldsymbol{\eta}$ with $\boldsymbol{\eta}$ being a spherical Gaussian. Note that such a normalized random matrix $\boldsymbol{W}$ does not only minimize the negentropy but also maximizes the log determinant term since it satisfies $\boldsymbol{W}\boldsymbol{W}^T = \frac{\gamma}{m}\boldsymbol{I}$. This completes the proof of the asymptotic InfoMax optimality of (normalized) random projections. Finally, note that $\boldsymbol{W}\boldsymbol{W}^T = \frac{\gamma}{m}\boldsymbol{I}$ implies that each measurement (each row vector of $\boldsymbol{W}$) equally has $\frac{\gamma}{m}$ as its norm (or power).

3.7.2 Miscellaneous Lemmas

Proposition 3.5. Suppose that we want to reconstruct an n-dimensional signal $\boldsymbol{x}$ given the measurement $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ for some $m \times n$ matrix $\boldsymbol{W}$, with $m \leq n$. In terms of average ℓ_2-error, an optimal matrix consists of the major m principal components of $\boldsymbol{x}$ in its rows if recovery is restricted to be linear.

Proof. Let $\boldsymbol{\Sigma}$ denote the covariance of $\boldsymbol{x}$. The ℓ_2-error of the linear MMSE estimate of $\boldsymbol{x}$ based on $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ is given by $\mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-1}\boldsymbol{W}\boldsymbol{\Sigma})$ [105]. Let $\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$ by its singular value decomposition, where $\boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$

with $\sigma_1^2 \geq \cdots \geq \sigma_n^2 > 0$. Then,

$$\begin{aligned}
\boldsymbol{W}^* &= \arg\min_{\boldsymbol{W}} \mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-1}\boldsymbol{W}\boldsymbol{\Sigma}) && (3.30)\\
&= \arg\max_{\boldsymbol{W}} \mathrm{Tr}(\boldsymbol{\Sigma}\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{\Sigma}\boldsymbol{W}^T)^{-1}\boldsymbol{W}\boldsymbol{\Sigma}) && (3.31)\\
&= \arg\max_{\boldsymbol{W}} \mathrm{Tr}(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\boldsymbol{W}^T)^{-1}\boldsymbol{W}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T) && (3.32)\\
&= \left(\arg\max_{\boldsymbol{V}} \mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Lambda})\right)\boldsymbol{U}^T && (3.33)
\end{aligned}$$

where we let $\boldsymbol{V} = \boldsymbol{W}\boldsymbol{U}$. Given $\boldsymbol{V}^*$ that maximizes $\mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Lambda})$, the matrix $\boldsymbol{W}^*$ can be computed simply by $\boldsymbol{V}^*\boldsymbol{U}^T$ due to the orthonormality of $\boldsymbol{U}$. Let $\boldsymbol{M} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}^{\frac{1}{2}}$. The matrix $\boldsymbol{M}$ is a symmetric, idempotent matrix of rank m, and thus its diagonal entries should satisfy the following conditions (see Lemma 2.14): $0 \leq M_{ii} \leq 1$ and $\sum_i M_{ii} = m$. Note that

$$\begin{aligned}
\mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{V}^T(\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T)^{-1}\boldsymbol{V}\boldsymbol{\Lambda}) &= \mathrm{Tr}(\boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{M}\boldsymbol{\Lambda}^{\frac{1}{2}}) && (3.34)\\
&= \mathrm{Tr}(\boldsymbol{\Lambda}\boldsymbol{M}) && (3.35)\\
&= \sum_i \sigma_i^2 M_{ii}, && (3.36)
\end{aligned}$$

which is obviously maximized when all the weights are concentrated on m largest σ_i^2's (i.e., $M_{ii} = 1$ for $i \leq m$ and zero otherwise), for example by taking

$$\boldsymbol{V}^* = [\overbrace{\boldsymbol{I}}^{m}\overbrace{\boldsymbol{0}}^{n-m}]. \tag{3.37}$$

Finally, by (3.33), $\boldsymbol{W}^*$ is composed of the first m rows of $\boldsymbol{U}^T$ or the major m eigenvectors of $\boldsymbol{\Sigma}$. □

Chapter 4

Learning Color Filter Arrays

Most digital cameras sense one color component per pixel, in a mosaic pattern as a whole, and then use "demosaicking" to reproduce a full color image. The color component sensed at each pixel is determined by the color filter array (CFA). The vast majority of cameras use the Bayer pattern CFA, but there exist a variety of other patterns as well, each proposed to replace the Bayer pattern.

In this chapter, we regard the way such digital cameras handle color as a special case of compressed sensing. Like in the previous chapters, we use the InfoMax principle as the design criterion. We seek to minimize the expected uncertainty that we would face in demosaicking. We first model the probability density of natural scenes in color, to mathematically define the uncertainty. Then, we use an efficient greedy algorithm to optimize the CFA in terms of the expected uncertainty. Our approach is validated by experimental results, in which our learned CFAs show significant improvements in performance over existing CFAs.

4.1 Introduction

Three independent light spectra produce color perceivable by human eyes. There exist three-chip cameras, or even single-chip cameras with layered detectors, which simultaneously measure three spectra at every pixel, but they require expensive physical devices such as precision beam splitters and multiple sensor arrays. Many digital cameras, instead,

overlay a color filter array (CFA) over the sensors and let each sensor detect only a single spectrum. Then, sensor outputs look like a mosaic (e.g. Figure 4.1, right), which stores,

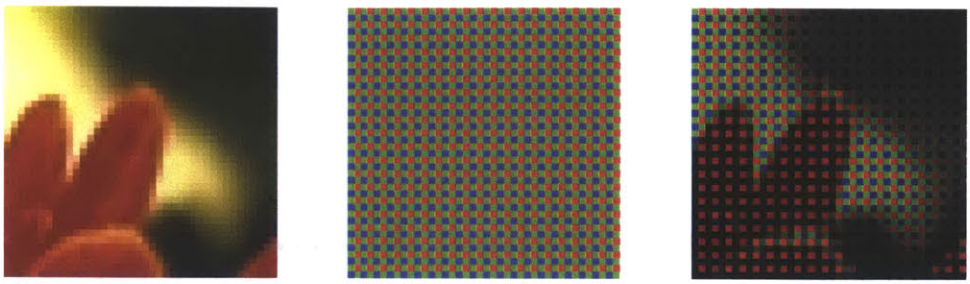


Figure 4.1: Color image sensing in a single-chip camera. Left: original scene. Middle: CFA installed in the camera. Right: actual measurement.

at each pixel, the intensity of a specific spectrum of the overlaid color filter. The sensor outputs should go through so-called *demosaicking* to recover full color. The demosaicking process usually performs interpolation between sensor outputs from the same type of color filters (see [68, 152]).

The most common CFA is the Bayer pattern [16] which arranges red, green, and blue filters in a chessboard layout (**B** in Figure 4.2), with green twice more than red or blue. The vast majority of cameras adopt the Bayer pattern, and a tremendous amount of work has been devoted to demosaicking the Bayer pattern (refer to [68, 89] and the papers cited therein). In relatively few studies, the optimality of the Bayer pattern has been doubted and alternative CFAs that enable better reconstruction have been sought.

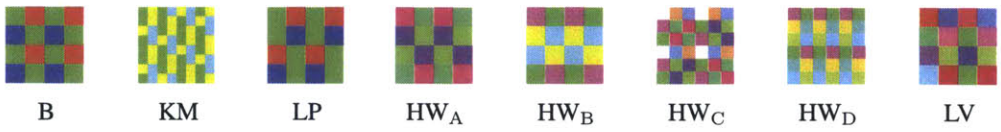


Figure 4.2: Examples of existing CFA patterns. **B:** Bayer, **KM:** 8×8 Knop & Morf [85], **LP:** 4×4 Lukac & Plataniotis [96], **HW$_A$–HW$_D$:** four patterns by Hirakawa & Wolfe [77], **LV:** Lu & Vetterli [95].

In 1985, Knop and Morf considered pseudo-random patterns [85], particularly restricting their attention to *shift patterns* where the filter arrangement is identical, up to shift, for every row. They showed some examples of shift patterns (e.g., **KM** in Figure 4.2) but did not give a criterion nor efficient way to optimize among infinitely many choices

of such patterns. In [96], Lukac and Plataniotis conducted extensive experimental comparisons among ten CFAs, all consisting of red, green, and blue filters. They found that several (ad-hoc) variations of the Bayer are comparable to or even slightly better than the original pattern (see **LP** in Figure 4.2 for the best variation pattern). In [77], Hirakawa and Wolfe developed a more formal framework, based on spectral analysis, to evaluate and design CFAs. The key behind their method is to minimize aliasing in luminance and chrominance channels at the same time. They found several good CFAs ($\mathbf{HW}_{\mathrm{A}}$–$\mathbf{HW}_{\mathrm{D}}$ in Figure 4.2), which, even when demosaicked linearly, often outperform the Bayer pattern that is demosaicked with a more elaborate scheme [67]. The minimization of aliasing is surely a reasonable thing, but perhaps not the best thing, to do. Surprisingly, the recent theory of compressed sensing [51, 32, 57] claims that sparse signals can be better reconstructed from subtly aliased measurements than from maximally anti-aliased ones. Color images are sparse (e.g., see [100]), so the claim may hold for color image sensing. More recently, Lu and Vetterli [95] minimized the expected reconstruction error of the optimal *linear* estimate, algorithmically, with respect to the CFA and found the pattern **LV** shown in Figure 4.2. This approach is also reasonable but still remains suboptimal in that it only depends on up to the second-order statistics of color images. The linear type of recovery is not sufficiently good in general. Most state-of-the-art demosaicking techniques involve nonlinear operators [97, 76, 39, 103, 91]. More explicitly, Mairal, Elad, and Sapiro [100] have experimentally shown that the high-order sparsity structure of natural images can play an important role in demosaicking. Such a nonlinear feature should desirably be taken into account also in designing CFAs.

In this chapter, we would like to learn a CFA in a way that it exploits the high-order statistics of natural images. We use a mixture of Gaussians (MoG) to tractably model the prior density of color image patches. Then, we attempt to minimize the uncertainty of the missing color components given the measured color components. This is a Bayesian experimental design approach (e.g., see [124]) for the CFA, which no one has attempted before, as far as we know. We consider a couple of criteria, Shannon's entropy and minimum mean-squared error (MMSE), for the uncertainty measure. A CFA that minimizes the MMSE criterion is obviously optimal in the sense of mean-squared error (MSE) or peak-

signal-to-noise-ratio (PSNR).[1] Note that we do not limit ourselves to linear recovery, so we can obtain better performance than with Lu and Vetterli's optimized pattern. To efficiently search for the optimal CFAs, we present a greedy algorithm. We also show how to universally demosaic from arbitrary CFAs based on the MoG prior of color images. Finally, we validate, by experiments, the improved performance of the newly found CFAs over any existing CFAs that we have briefly reviewed in this section.

4.2 Color Image Sensing: Mathematical Review

Let $\boldsymbol{x}$ be the rasterized vector of an m-pixel color image. If we write the color sensing process in the form of $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$, the matrix $\boldsymbol{W}$ is a rectangular matrix that has three times as many columns as rows. Mathematically, it can be represented as

$$\boldsymbol{W} = \sum_i (\boldsymbol{e}_i \boldsymbol{e}_i^T \otimes \boldsymbol{c}_i^T), \tag{4.1}$$

where $\boldsymbol{c}_i$ is a three-dimensional vector denoting the color filter at the ith pixel; $\boldsymbol{e}_i$ denotes an m-dimensional unit vector with 1 at the ith entry; and $\otimes$ denotes Kronecker (or tensor) product. The structure of $\boldsymbol{W}$ in Equation (4.1) reflects the fact that the measurements are pixelwise, not allowed to multiplex color spectra from different locations. We also have $\boldsymbol{c}_i \succeq \boldsymbol{0}$, for all i, because negative weights are not feasible in color filters.

4.3 Image Prior

As manifest in Section 4.2, recovering $\boldsymbol{x}$ from $\boldsymbol{y}$ is an ill-posed problem in itself, which has infinitely many solutions. If we know the prior density $p(\boldsymbol{x})$, we can use it in addition to the measurement to uniquely determine the most appropriate solution for a given loss function.

Image priors have actively been investigated in the recent past (e.g., see [154, 53, 2, 59, 102, 118]) and have proved effective in image restoration tasks such as denoising [117, 58,

[1]The PSNR is a strictly decreasing function of the MSE. Refer to Equation (4.9) for the mathematical definition of PSNR.

73, 147], deblurring [86, 38], and also color demosaicking [100]. In particular, Zoran and Weiss [156] have shown that a mixture of Gaussians (MoG) provides a good prior, while being very tractable, for a rather small size of image patches. The goodness of the MoG model has been only validated with gray-scale images but may be generalizable to color images as well. In this chapter, we represent the prior density of color images in a 5×5 local neighborhood by an MoG with N clusters, i.e.,

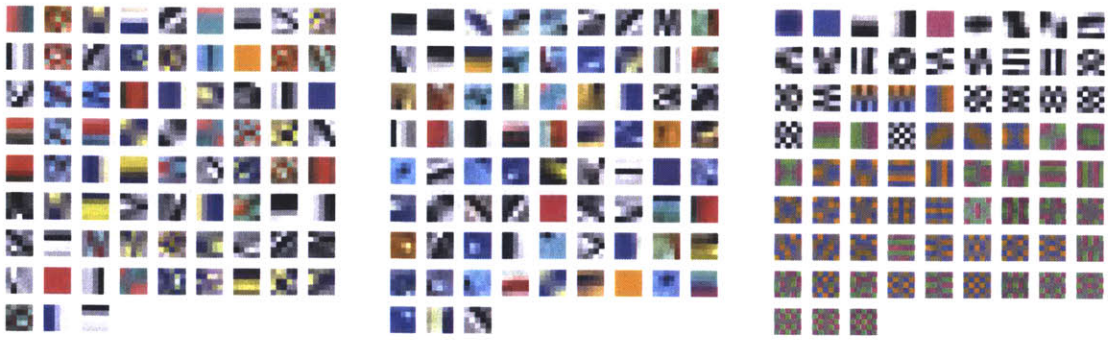
$$p(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{4.2}$$

with $\alpha_i \geq 0$ and $\sum_{i=1}^{N} \alpha_i = 1$. We set $N = 75$. From [156], we expect each Gaussian cluster to be ellipsoidal, in the signal space, with a certain orientation. In this sense, we accommodate as many orientations as the dimensions. We find the remaining set of parameters $\{\alpha_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}_{i=1}^{N}$ by applying the expectation maximization (EM) method [56] on a million samples from Berkeley training dataset [101]. The Berkeley training dataset is a subset of the Corel PhotoCD library, and photographs were taken using analog, film-based cameras and hence do not contain any demosaicking artifacts.

To see how well the learned MoG preserves the high-order statistics of natural scenes, we conducted independent component analysis (ICA) [79] on a million samples, now randomly generated from the learned MoG prior. As shown in Figure 4.3 (top left), the synthesis basis functions are edge filters, very similar to those obtained directly from real image samples (in Figure 4.3, top middle; cf. [19]), and the coefficient distributions are kurtotic. Meanwhile, if we used a single Gaussian model, only based on up to the second-order statistics, the ICA basis would be simply a set of Fourier filters (Figure 4.3, top right), and the sparsity structure of a natural scene along the edges would be lost.

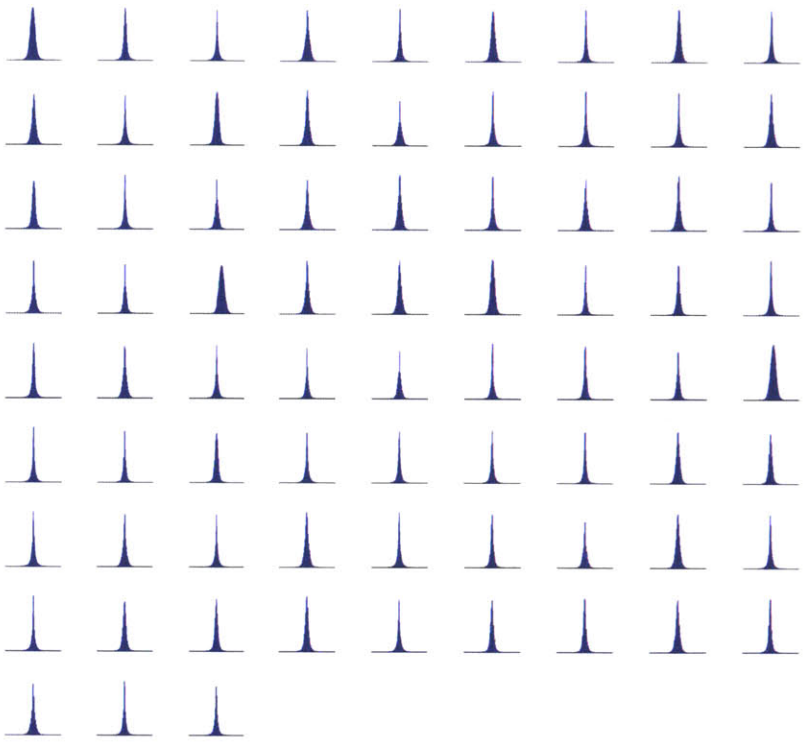
4.4 CFA Design

Given the MoG prior, how should we design a CFA so that it enables the best reconstruction? We measure error in (R,G,B) space given a linear measurement which is assumed to be a linear combination of R, G, and B filters. There are some filters that are not linear



MoG Nonparametric Gaussian

ICA synthesis basis



Distributions of ICA synthesis coefficients (MoG)

Figure 4.3: High-order statistics of the learned MoG prior. ICA synthesis basis functions and corresponding coefficient distributions are shown on the top left and on the bottom, respectively. To generate the ICA basis functions, we have used the fast ICA algorithm by Hyvärinen [79]. Despite simplicity, the MoG prior well preserves the sparsity structure of natural scenes along edges. Compare with the ICA basis functions derived from a non-parametric model (top middle) and from a single Gaussian model (top right). Note that the single Gaussian has completely lost the edge structure.

combinations of R, G, and B filters, which we cannot evaluate because we do not have their statistics. But if we restrict ourselves to filters that are linear combinations of R, G, B filters, then we have all the statistics we need. We consider several ways in this section.

Randomization. The color sensing is notionally a compressed sensing: We only acquire incomplete color samples of a natural scene which clearly possesses a sparsity structure (Figure 4.3). We know that the theory of compressed sensing is not directly applicable, due to the restriction on the measurement matrices (Equation 4.1). Nevertheless, we are still tempted to try a random pattern and eager to see whether the randomness also works in the CFA-type setting. This is somewhat similar to Knop and Morf's pseudo-random CFAs, but we do not restrict our attention only to the shift patterns. The colors in a CFA are likely to be all distinct.

Bayesian experimental design. At the decoder's side, we are expected to use the measurement as well as the image prior to perform inference on the missing color components. On each 5×5 local neighborhood, we are specifically given twenty-five measurements $\boldsymbol{y} = (y_1, \ldots, y_{25})$, one per pixel. In the vectorized notation, y_{13} particularly denotes the measurement at the center pixel. If we define u_{13} and v_{13} as the pair of the complementary elements, in defining color, at the center pixel, we can form the posterior density $p(u_{13}, v_{13}|y_1, \ldots, y_{25})$ for Bayesian inference. A reasonable objective at the encoder's side is then to design a CFA so as to minimize the expected uncertainty of u_{13}, v_{13} given the measurement $y_1, \ldots, y_{25}$. The uncertainty minimization framework is commonly called Bayesian experimental design (refer to [124]; see also [146, 34, 127] in the context of compressed sensing). We consider a couple of criteria for the uncertainty measure: the conditional entropy [93] and the conditional variance. When the conditional entropy is the underlying uncertainty measure, Bayesian experimental design implements the InfoMax principle (see Section 2.2.1).

4.4.1 Learning Bayes Optimal CFA

Like many others, we search for a 4×4 CFA which, we believe, is small enough to be tractable yet large enough to yield interesting sampling patterns. The CFA is replicated

along both horizontal and vertical directions until they match the full image size. Then, a specific CFA generates sixteen (cyclic) patterns in the 5×5 local neighborhood (see Figure 4.4). Each pattern, denoted by $k = 1, \ldots, 16$, is equally likely. Let $f_k(\widetilde{\boldsymbol{c}}_1, \ldots, \widetilde{\boldsymbol{c}}_{16})$

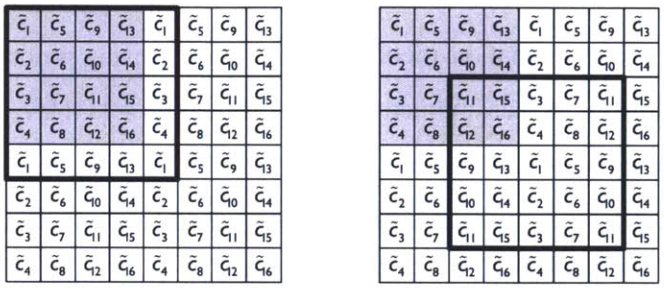


Figure 4.4: Replication of a 4×4 CFA in both directions. The CFA generates sixteen patterns in the 5×5 local neighborhood. Shown are a couple of the patterns (inside 5×5 square boxes).

denote the uncertainty of the missing color components at the center pixel given twenty five local measurements when $\widetilde{\boldsymbol{c}}_1, \ldots, \widetilde{\boldsymbol{c}}_j$'s are arranged according to the kth pattern. Then, the objective is to minimize $f(\widetilde{\boldsymbol{c}}_1, \ldots, \widetilde{\boldsymbol{c}}_{16}) = \frac{1}{16} \sum_{k=1}^{16} f_k(\widetilde{\boldsymbol{c}}_1, \ldots, \widetilde{\boldsymbol{c}}_{16})$ with respect to the CFA, $\{\widetilde{\boldsymbol{c}}_j\}_{j=1}^{16}$.

For efficiency, we discretize the search space by choosing $\widetilde{\boldsymbol{c}}_j$'s from a finite set $\mathcal{C}$ that consists of thirteen color filters shown in Figure 4.5. The problem becomes then a combi-

Figure 4.5: Restricted set of color filters: red, green, blue, white (or gray), yellow, magenta, cyan, red+magenta, red+yellow, green+yellow, blue+magenta, blue+cyan, green+cyan. They have been frequently used in the CFA design (see Figure 4.2 for example).

natorial optimization among as many as about 13^{16} CFAs. To further alleviate the search complexity, we consider a greedy algorithm, as illustrated in Figure 4.6. Suppose, for the time being, that we have a way to evaluate f_k's (and thus f as well) given a CFA $\{\widetilde{\boldsymbol{c}}_j\}_{j=1}^{16}$. We initially choose an arbitrary pattern (e.g., all white) and evaluate f for that pattern CFA. At each step, we consider only a "local" move (i.e. update of a single color filter) from the CFA currently at our hands. For all possible local moves,[2] we evaluate f's and find the CFA

[2]The number is computed by $13 \cdot 16 - 1 = 207$.

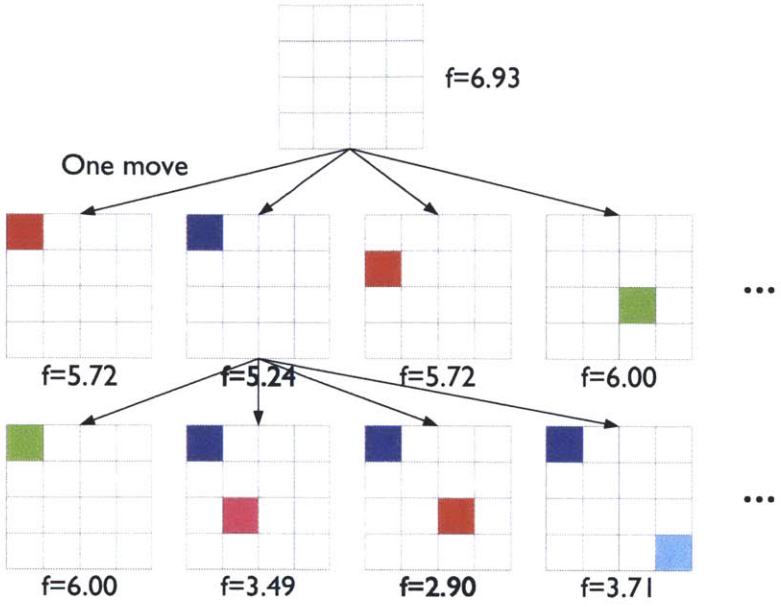


Figure 4.6: Illustration of our greedy search algorithm, where we want to minimize the value of f with respect to the CFA. Starting with an initial CFA, we update at most one filter at a time so as to reduce f as much as possible. The boldfaced number denotes the minimum value of f at each level of search and thus the CFA above it is selected at that time. The greedy search algorithm is guaranteed to converge to a locally optimal solution.

that produces the minimum value, say f^*. If f^* is smaller than the current f, it suggests that the new CFA is better, in terms of uncertainty, than the current one, so we select it and continue to search for the next local move. Otherwise, the current CFA is locally optimal and becomes the final output of the algorithm.

If we denote the depth of search (or the depth of hierarchy in Figure 4.6) by d, the overall search complexity will be proportional to $13 \cdot 16 \cdot d$, much smaller than 13^{16}. The depth d can be affected by the initialization but is typically 1–2 times as many as the number of pixels in the CFA.

Now let us discuss how to evaluate f given a CFA. Recall that, given a CFA, the color filter $\boldsymbol{c}_i$ at the ith pixel actually depends on the neighborhood pattern (Figure 4.4), and thus we will precisely denote it by $\boldsymbol{c}_{i,k}$ wherever clarity is needed. Given a particular neighborhood pattern k, we will compute f_k as below. The overall uncertainty measure f is simply the average of f_k's.

Conditional entropy. If we use Shannon's entropy h for the uncertainty measure, f_k is

defined as $f_k = h(u_{13}, v_{13}|y_1, \ldots, y_{25})$ where $y_1, \ldots, y_{25}$ and u_{13}, v_{13} are all based on the specific neighborhood pattern k. We expand the conditional entropy $h(u_{13}, v_{13}|y_1, \ldots, y_{25})$ as $h(u_{13}, v_{13}|y_1, \ldots, y_{25}) = h(y_1, \ldots, y_{25}, u_{13}, v_{13}) - h(y_1, \ldots, y_{25})$ (e.g., see [43]). Then, we can write

$$f_k = h(\boldsymbol{W}'_k \boldsymbol{x}) - h(\boldsymbol{W}_k \boldsymbol{x}) \tag{4.3}$$

with

$$\boldsymbol{W}_k = \begin{bmatrix} \boldsymbol{e}_1^T \otimes \boldsymbol{c}_{1,k}^T \\ \vdots \\ \boldsymbol{e}_{25}^T \otimes \boldsymbol{c}_{25,k}^T \end{bmatrix}, \quad \boldsymbol{W}'_k = \begin{bmatrix} \boldsymbol{W}_k \\ \boldsymbol{e}_{13}^T \otimes \boldsymbol{c}_{13,k}'^T \\ \boldsymbol{e}_{13}^T \otimes \boldsymbol{c}_{13,k}''^T \end{bmatrix}, \tag{4.4}$$

where $\boldsymbol{c}'_{13,k}$ and $\boldsymbol{c}''_{13,k}$ denote orthonormal vectors both in the nullspace of $\boldsymbol{c}_{13,k}^T$ (complementary color components).

If $\boldsymbol{x}$ has an MoG prior (Equation 4.2), a random vector generated by $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}$ also has an MoG prior with $p(\boldsymbol{z}) = \sum_i \alpha_i \mathcal{N}(\boldsymbol{z}; \boldsymbol{\Phi}\boldsymbol{\mu}_i, \boldsymbol{\Phi}\boldsymbol{\Sigma}_i\boldsymbol{\Phi}^T)$. We define ω as the hidden indicator variable on the Gaussian cluster from which the vector $\boldsymbol{x}$ (and $\boldsymbol{z}$) actually comes. Then, $h(\boldsymbol{z}) = h(\boldsymbol{z}|\omega) + H(\omega) - H(\omega|\boldsymbol{z})$, where $H(\cdot)$ denotes the entropy for discrete random variables.[3] The conditional entropy $h(\boldsymbol{z}|\omega)$ is easy to compute, while the entropy $h(\boldsymbol{z})$ is not. Explicitly, $h(\boldsymbol{z}|\omega) = \frac{r}{2}\log(2\pi e) + \frac{1}{2}\sum_i \alpha_i \log\det(\boldsymbol{\Phi}\boldsymbol{\Sigma}_i\boldsymbol{\Phi}^T)$, where r is the number of rows in $\boldsymbol{\Phi}$. Therefore, f_k in Equation (4.3) becomes $f_k = \log(2\pi e) + \frac{1}{2}\sum_i \alpha_i \left(\log\det(\boldsymbol{W}'_k\boldsymbol{\Sigma}_i\boldsymbol{W}'^T_k) - \log\det(\boldsymbol{W}_k\boldsymbol{\Sigma}_i\boldsymbol{W}_k^T)\right) - H(\omega|\boldsymbol{W}'_k\boldsymbol{x}) + H(\omega|\boldsymbol{W}_k\boldsymbol{x})$. The quantity $H(\omega|\boldsymbol{W}'_k\boldsymbol{x}) - H(\omega|\boldsymbol{W}_k\boldsymbol{x})$ represents the change in the uncertainty of the cluster indicator ω when two color components u_{13}, v_{13} are added to the existing set of the observations $y_1, \ldots, y_{25}$. Here, we will assume that it is negligible[4] and will simply use

[3] This is because the mutual information $I(\boldsymbol{z}; \omega)$ between $\boldsymbol{z}$ and ω can be expanded in two ways: $I(\boldsymbol{z}; \omega) = h(\boldsymbol{z}) - h(\boldsymbol{z}|\omega) = H(\omega) - H(\omega|\boldsymbol{z})$.

[4] We argue that we can tell the Gaussian cluster given the measurement $(y_1, \ldots, y_{25})$ with somewhat high certainty. This implies that $H(\omega|\boldsymbol{W}_k\boldsymbol{x}) \approx 0$. Because $0 \leq H(\omega|\boldsymbol{W}'_k\boldsymbol{x}) \leq H(\omega|\boldsymbol{W}_k\boldsymbol{x})$, we have $0 \approx -H(\omega|\boldsymbol{W}_k\boldsymbol{x}) \leq H(\omega|\boldsymbol{W}'_k\boldsymbol{x}) - H(\omega|\boldsymbol{W}_k\boldsymbol{x}) \leq 0$ and $H(\omega|\boldsymbol{W}'_k\boldsymbol{x}) - H(\omega|\boldsymbol{W}_k\boldsymbol{x}) \approx 0$. Our argument is somewhat analogous to the state-of-the-art demosaicking techniques which first determine the edge direction "with certainty" based only on the local measurement and then use edge-preserving interpolation (see also Section 4.5).

$f_k \approx \log(2\pi e) + \frac{1}{2}\sum_i \alpha_i \left(\log\det(\boldsymbol{W}'_k\boldsymbol{\Sigma}_i\boldsymbol{W}'^T_k) - \log\det(\boldsymbol{W}_k\boldsymbol{\Sigma}_i\boldsymbol{W}_k^T)\right)$.

Conditional variance (or MMSE). Another reasonable uncertainty measure is the conditional variance, i.e., $f_k = \mathbb{E}\left[|u_{13} - \overline{u}_{13}(\boldsymbol{y})|^2 + |v_{13} - \overline{v}_{13}(\boldsymbol{y})|^2\right]$, where $\overline{u}_{13}(\boldsymbol{y}) \triangleq \mathbb{E}\left[u_{13} \mid y_1, \ldots, y_{25}\right]$ and $\overline{v}_{13}(\boldsymbol{y}) \triangleq \mathbb{E}\left[v_{13} \mid y_1, \ldots, y_{25}\right]$, This is also known as the minimum mean-squared error (MMSE) criterion. Given a particular instance $\boldsymbol{y} = (y_1, \ldots, y_{25})$, we can compute $\overline{u}_{13}$ and $\overline{v}_{13}$, in closed form, by

$$\overline{u}_{13}(\boldsymbol{y}) = (\boldsymbol{e}_{13}^T \otimes \boldsymbol{c}'^T_{13,k}) \sum_i \alpha'_{i,k} \left(\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i\boldsymbol{W}_k^T(\boldsymbol{W}_k\boldsymbol{\Sigma}_i\boldsymbol{W}_k^T)^{-1}(\boldsymbol{y} - \boldsymbol{W}_k\boldsymbol{\mu}_i)\right) \tag{4.5}$$

$$\overline{v}_{13}(\boldsymbol{y}) = (\boldsymbol{e}_{13}^T \otimes \boldsymbol{c}''^T_{13,k}) \sum_i \alpha'_{i,k} \left(\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i\boldsymbol{W}_k^T(\boldsymbol{W}_k\boldsymbol{\Sigma}_i\boldsymbol{W}_k^T)^{-1}(\boldsymbol{y} - \boldsymbol{W}_k\boldsymbol{\mu}_i)\right) \tag{4.6}$$

where $\alpha'_{i,k}$ is the posterior cluster probability, i.e.,

$$\alpha'_{i,k} = \frac{\alpha_i \mathcal{N}(\boldsymbol{y}; \boldsymbol{W}_k\boldsymbol{\mu}_i, \boldsymbol{W}_k\boldsymbol{\Sigma}_i\boldsymbol{W}_k^T)}{\sum_i \alpha_i \mathcal{N}(\boldsymbol{y}; \boldsymbol{W}_k\boldsymbol{\mu}_i, \boldsymbol{W}_k\boldsymbol{\Sigma}_i\boldsymbol{W}_k^T)}. \tag{4.7}$$

Then, we are able to evaluate f_k based on Monte Carlo method. We draw a number of samples from the learned MoG prior; apply the CFA to each sample; compute the MMSE estimates (Equations 4.5, 4.6); and then take the sample average of $|u_{13} - \bar{u}_{13}|^2 + |v_{13} - \bar{v}_{13}|^2$.

Instead of drawing samples from the MoG prior, we may use real image patches as well. More simply, we can apply the augmented CFA to the entire images, rather than 5×5 image patches, in a training dataset and then use the MMSE estimates (Equations 4.5, 4.6) to recover the full images. The MSE between the original images and their reconstruction is simply f which we seek to compute. In this chapter, we take the last trick using ten images from Berkeley training dataset (Figure 4.7). Ten images are not many (good in terms of learning speed) but 200,000 pixels in them are sufficiently many to learn sixteen color filters on.

4.4.2 Results

1) Randomization: We chose a 4×4 random CFA in a way that each element in $\widetilde{\boldsymbol{c}}_j$ is i.i.d. uniformly distributed in [0,1]. A particular instance is shown in Figure 4.8**R**.

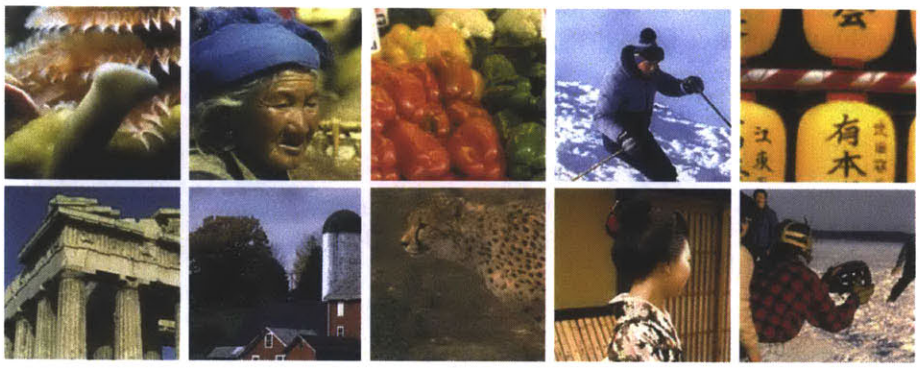


Figure 4.7: Ten images from Berkeley training dataset [101].

2) Bayesian experimental design: Starting with a 4×4 random CFA, we ran our greedy search algorithm (Figure 4.6), which gave us two patterns (each according to conditional entropy and conditional variance) shown in Figure 4.8$\mathbf{P}_{\mathrm{A}}$ and 4.8$\mathbf{P}_{\mathrm{B}}$.

Figure 4.8: A random CFA $\mathbf{R}$ and two learned CFAs $\mathbf{P}_{\mathrm{A}}$ and $\mathbf{P}_{\mathrm{B}}$ (all 4×4). The pattern $\mathbf{P}_{\mathrm{A}}$ minimizes the conditional entropy of the missing color components given the measurement and the pattern $\mathbf{P}_{\mathrm{B}}$ minimizes the conditional variance. Compare with other patterns in Figure 4.2.

As a sanity check, we evaluated the conditional entropy and conditional variance (i.e. our criteria for the CFA design) for various CFAs. The results are shown in Tables 4.1 and 4.2.

4.5 Color Demosaicking

Perhaps a naive way of demosaicking is to linearly interpolate sensor outputs from the same type of color filters. For the Bayer pattern CFA, we can define convolution kernels as

$$F_{\mathrm{Red}} = F_{\mathrm{Blue}} = \frac{1}{4}\begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \quad F_{\mathrm{Green}} = \frac{1}{4}\begin{bmatrix} 0 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \tag{4.8}$$

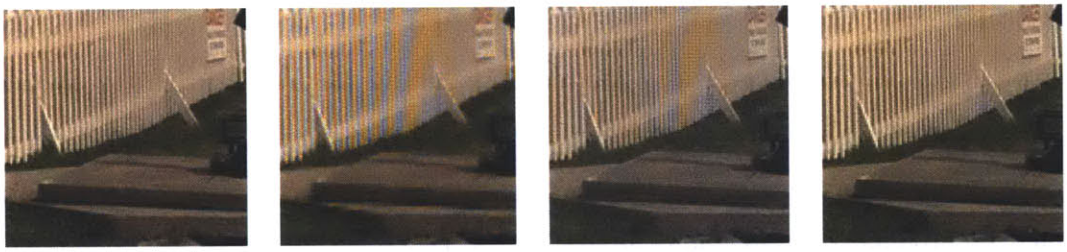
Table 4.1: Conditional entropy (estimate) for various CFAs. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe ($\mathbf{HW}_{\mathrm{A}}$–$\mathbf{HW}_{\mathrm{D}}$), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns ($\mathbf{P}_{\mathrm{A}}$, $\mathbf{P}_{\mathrm{B}}$).

B	KM	LP	HW_{A}	HW_{B}	HW_{C}	HW_{D}	LV	R	P_{A}	P_{B}
4.453	5.371	4.525	4.539	4.880	4.873	5.426	6.470	5.330	**4.236**	4.327

Table 4.2: Conditional variance for various CFAs on training images. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe ($\mathbf{HW}_{\mathrm{A}}$–$\mathbf{HW}_{\mathrm{D}}$), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns ($\mathbf{P}_{\mathrm{A}}$, $\mathbf{P}_{\mathrm{B}}$).

B	KM	LP	HW_{A}	HW_{B}	HW_{C}	HW_{D}	LV	R	P_{A}	P_{B}
40.48	78.38	39.46	36.58	39.65	47.56	57.44	40.76	71.15	32.38	**31.06**

which apply to red, green, and blue channels, separately, with missing pixels filled by zeros [6]. In Figure 4.9, we provide an example of the bilinear reconstruction from Bayer mosaic. The reconstruction suffers from two major artifacts: blurring and color fringing.



Original | Bilinear (22.45dB) | Linear MMSE (31.41 dB) | DLMMSE [152] (39.45 dB)

Figure 4.9: Original *Fence* image and three reconstructions from its Bayer mosaic. The recovery scheme used for each reconstruction, as well as the PSNR performance, are indicated. See Equation (4.9) for the mathematical definition of PSNR. The linear MMSE reconstruction has been obtained with Knop and Morf's regression method, but any other implementations would produce similar results.

A number of modern demosaicking techniques (e.g., [97, 76, 152, 39, 103, 91]) employ nonlinear operators particularly such as edge classification, whether implicit or explicit, to preserve edges during the interpolation. They greatly reduce the blurring and color fringing artifacts. For example, see Figure 4.9 (rightmost), where we have used the directional linear MMSE (DLMMSE), proposed by Zhang and Wu [152], which is one of such edge-preserving interpolations. Retrospecting from the viewpoint of the image prior, these

techniques exploit the high-order statistics of natural scenes; however, the prior is tightly coupled with the Bayer pattern, which makes them inapplicable to other CFAs.

To reconstruct from non-Bayer CFAs, Knop and Morf [85] used linear regression. Given a pseudo-random pattern, they learned regression matrices which produce the color of the center pixel from the measurements on the local neighborhood. Because a CFA generates as many local neighborhood patterns as its size (refer to Figure 4.4), the number of such regression matrices should also be equal to the CFA size. The regression was conducted so as to minimize the MSE. In this sense, their reconstruction is a particular way of implementing the linear MMSE (LMMSE) estimate. Essentially the same idea has been formulated in various aspects and implemented in various ways (e.g., see [23, 133, 48, 138]). The recovery schemes used by Lukac & Plataniotis [96], by Hirakawa & Wolfe [77], and by Lu & Vetterli [95] are also similar to (or a special case of) these approaches. The performance of LMMSE typically lies in between those of the bilinear interpolation and DLMMSE, as an example is shown in Figure 4.9. In its favor, LMMSE applies to any CFA, however its performance is not the best. Nayar and Narasimhan [108, 106] proposed an extension to the LMMSE, by allowing polynomial-order kernels in the regression, in a more generic framework of "assorted pixels," where more information (e.g., brightness, polarization) are involved besides color. But the improvement seems to grow very slowly with the polynomial order and with the size of training set. Although omitted in Figure 4.9, the reconstruction with the second-order polynomial kernels remains very similar to the LMMSE result.

Note that we were able to compute the MMSE estimate, in closed form, of the missing color components given the learned MoG prior. Provided that the learned prior is close to the true density, this should be optimal in terms of MSE. Therefore, we use the MoG-based MMSE to universally demosaic from an arbitrary CFA.

Our estimate may be conceptually connected to LMMSE and DLMMSE. First, if we used a single Gaussian model for the prior, our estimate would eventually be the same as the LMMSE. Certainly, the MoG prior is closer to the true density than the single Gaussian model (see Figure 4.3), and thus the MoG-based MMSE will be guaranteed to be better than the LMMSE. Second, if we used a hard decision in evaluating the posterior cluster

probability (Equation 4.7), e.g., by setting the maximum to one and the others to zero, our estimate would be similar to the DLMMSE. The hard decision on the Gaussian cluster is analogous to the determination of the edge direction in DLMMSE. The remaining process of the DLMMSE is essentially to apply the LMMSE with specific second-order statistics to preserve the directional edge during the interpolation.

We will present shortly, in Section 4.6, how the results of the MoG MMSE method compare with those of the LMMSE and DLMMSE.

4.6 Experimental Results

First, we evaluated the performance of recovery schemes in terms of peak-signal-to-noise-ratio (PSNR), defined as

$$\mathsf{PSNR} = 10 \log_{10} \frac{3 \cdot 255^2}{\frac{1}{m} \|\boldsymbol{x} - \widehat{\boldsymbol{x}}\|^2} \tag{4.9}$$

where $\widehat{\boldsymbol{x}}$ denotes the reconstructed image. We compared the LMMSE, DLMMSE [152], and nonlinear MMSE based on the learned MoG prior (Equations 4.5, 4.6) using twenty images shown in Figure 4.10. The test images were originally scanned from film-based photos, containing ground-truth R, G, B values at each position. In this set of experiments, we commonly used the Bayer CFA because the DLMMSE, like most other conventional demosaicking schemes, works only for the Bayer pattern.

The results are provided in Figure 4.11. For all twenty images, both nonlinear schemes perform far better than the linear MMSE, and for most images (except for the eighth and tenth) and on average, the MoG-based MMSE works slightly better than the DLMMSE. More desirably, the MoG-based MMSE is applicable to any type of CFA. Therefore, we commonly use the MoG-based MMSE in the next set of experiments where we evaluate various CFAs.

The evaluation of CFAs was also conducted on the twenty images in Figure 4.10. We included eight existing CFAs, shown in Figure 4.2, and a 4×4 random CFA ($\mathbf{R}$) plus the learned CFAs ($\mathbf{P}_{\mathrm{A}}$, $\mathbf{P}_{\mathrm{B}}$) in Figure 4.8. We uniformly quantized each CFA output in eight

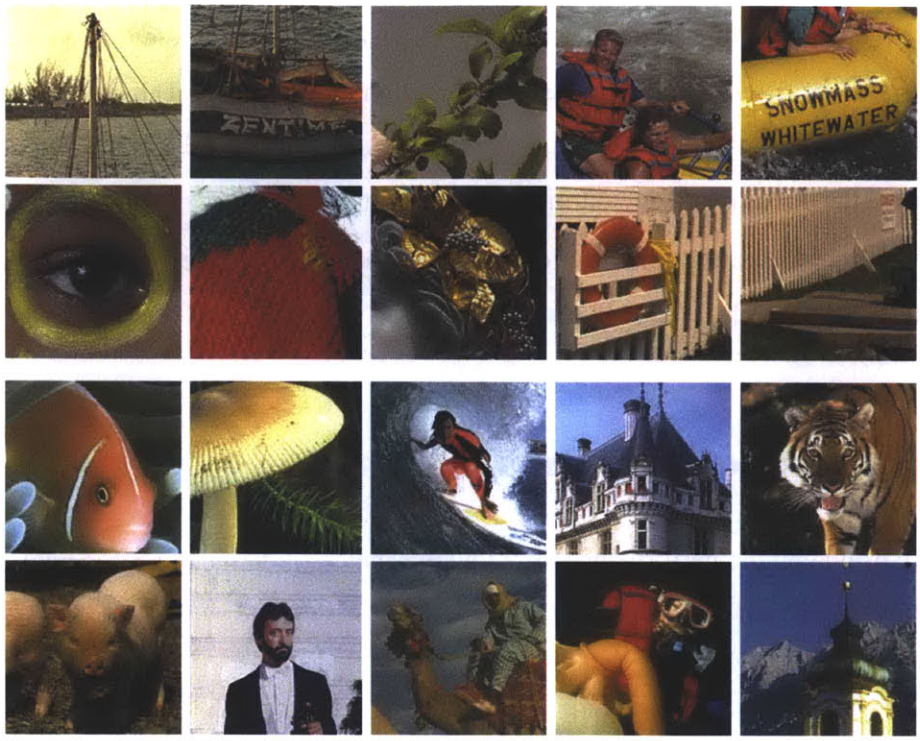


Figure 4.10: Test images. First half are from Kodak PhotoCD image set, while the others are from Berkeley test dataset. They will be numbered 1–20, from left to right, top to bottom.

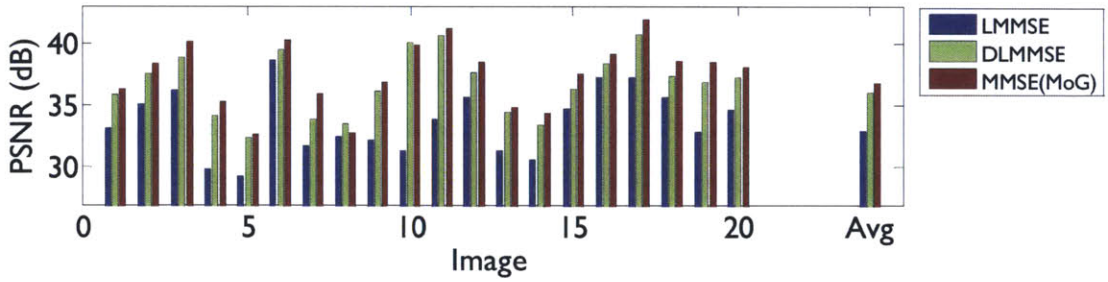


Figure 4.11: Comparison, in terms of PSNR, of three demosaicking schemes on Bayer pattern. The compared schemes are LMMSE, DLMMSE [152], and the MoG-based MMSE. The MoG-based MMSE works best for a majority of images.

bits for fair comparisons.

Detailed comparisons on the performance of CFAs (per test image) are provided in Table 4.3. As aforementioned, the MoG-based MMSE has been used for recovery. The learned CFAs ($\mathbf{P}_{\mathrm{A}}$, $\mathbf{P}_{\mathrm{B}}$) quite consistently outperform Bayer ($\mathbf{B}$) and give the best performance among all CFAs on average and also in terms of the number of top PSNR values,

although a CFA designed by Hirakawa & Wolfe ($\mathbf{HW}_{\mathrm{A}}$) is somewhat comparable. The pattern $\mathbf{P}_{\mathrm{B}}$, designed to minimize the conditional variance, shows slightly better PSNR performance than the pattern $\mathbf{P}_{\mathrm{A}}$.[5] The performance of the random CFA (**R**) may not look great, but we still consider it to be meaningful. At least, it provides a certain level of faithful recovery for natural color images which are sparse. We would like to remind readers that the random CFA does not make the measurement matrix $\boldsymbol{W}$ fully random (due to the pixelwise constraint; see Section 4.2) and thus that this result is not firmly based on the theory of compressed sensing.

Table 4.3: PSNR performance of various CFAs on test images. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe ($\mathbf{HW}_{\mathrm{A}}$–$\mathbf{HW}_{\mathrm{D}}$), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns ($\mathbf{P}_{\mathrm{A}}$, $\mathbf{P}_{\mathrm{B}}$).

CFA	B	KM	LP	HW_{A}	HW_{B}	HW_{C}	HW_{D}	LV	R	P_{A}	P_{B}
1	36.41	34.05	35.09	36.41	35.77	35.98	34.01	35.38	34.51	**36.90**	36.81
2	38.41	35.61	38.92	39.43	38.87	37.94	37.52	39.20	36.25	39.69	**39.91**
3	40.25	35.50	39.41	39.52	40.07	38.03	38.48	39.13	36.48	39.99	**40.41**
4	35.14	33.03	35.82	34.50	34.45	34.07	33.70	34.93	32.99	36.03	**36.11**
5	32.81	30.15	32.59	33.14	33.03	32.09	31.52	32.59	30.22	33.13	**33.83**
6	**40.30**	34.46	39.18	38.97	39.08	37.71	37.53	38.82	35.41	39.93	39.71
7	35.93	33.62	36.30	36.07	35.44	35.15	34.66	36.28	33.93	36.49	**36.87**
8	32.83	30.88	33.00	33.71	33.69	32.34	31.85	32.78	30.68	33.58	**34.08**
9	36.99	34.07	37.56	38.34	37.29	36.55	35.80	37.18	34.50	38.34	**38.78**
10	39.84	37.50	41.16	42.07	41.21	41.14	38.92	41.47	38.82	**42.62**	42.43
11	41.27	38.10	41.11	41.11	39.80	39.81	37.88	40.58	38.18	**42.02**	42.01
12	**38.60**	33.07	37.72	38.00	37.85	37.09	35.56	37.71	34.68	38.26	38.44
13	35.02	32.12	34.70	34.75	34.28	33.39	32.80	34.80	31.89	**36.32**	35.59
14	34.32	31.97	36.03	37.15	36.48	35.63	34.17	35.92	33.25	37.37	**37.93**
15	37.69	35.13	37.74	38.47	38.02	37.39	36.64	38.15	35.68	38.94	**39.07**
16	39.07	37.23	39.78	40.65	40.12	38.94	38.60	39.96	36.63	40.85	**41.27**
17	41.91	37.03	42.31	42.19	42.38	41.33	39.53	41.55	38.52	**43.22**	42.75
18	38.68	36.09	38.83	39.21	38.59	37.97	36.26	38.65	36.08	**39.98**	39.74
19	38.66	35.77	38.56	38.38	37.85	37.47	36.83	38.65	36.10	**39.75**	39.34
20	38.18	34.73	38.57	38.77	38.75	37.28	36.66	37.47	35.79	39.43	**39.54**
Avg.	36.83	33.96	36.94	37.27	36.92	36.13	35.31	36.80	34.38	37.80	**37.98**

We also compared the CFAs using the structural similarity (SSIM) index [145] as a

[5]The conditional variance is directly related to the PSNR. If the MMSE estimate is used, the denominator in Equation (4.9) corresponds to the conditional variance.

quality metric. The average performance is shown in Table 4.4. The relative order remains nearly the same as in the PSNR performance.

Table 4.4: Average SSIM performance of various CFAs on test images. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe ($\mathbf{HW_A}$–$\mathbf{HW_D}$), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns ($\mathbf{P_A}$, $\mathbf{P_B}$).

B	KM	LP	HW_A	HW_B	HW_C	HW_D	LV	R	P_A	P_B
.9853	.9777	.9850	.9866	.9862	.9848	.9823	.9859	.9796	.9868	**.9872**

Finally, we present several reconstruction results demosaicked from each CFA in Figures 4.13–4.16, for the subjective evaluation. Per image, only an important portion has been shown together with the error residual. Our learned CFAs provide better visual quality in comparison with others. For the Bayer pattern, we find that the MoG-based MMSE performs relatively well, but the errors could be further reduced with some other alternative CFAs. For example, a small level of color fringing artifacts in Figure 4.16**B** could successfully be suppressed in Figure 4.16$\mathbf{P_A}$ and 4.16$\mathbf{P_B}$.

4.7 Discussion

Many cameras use the Bayer pattern CFA and there has been much work on how to reconstruct from the Bayer mosaic. However, is Bayer the right thing to do? In this chapter, we learned the color image prior for a natural scene and attempted to optimize the CFA given the prior. We argued that the conditional entropy and conditional variance are good criteria for the CFA design. Then, we proposed a greedy algorithm to find a CFA that minimizes each criterion. We also provided a good universal demosaicking scheme based on the learned prior. Finally, we validated by experiments that our learned CFAs enable better reconstruction than Bayer and other existing CFAs (see the summarized results in Figure 4.12).

Historically, there has been prior work on optimizing the CFA in terms of linear MMSE. The approach is reasonable but not free from the criticism that it only optimizes a restrictive (possibly not the right) criterion. The present work goes beyond that. It sought a truly

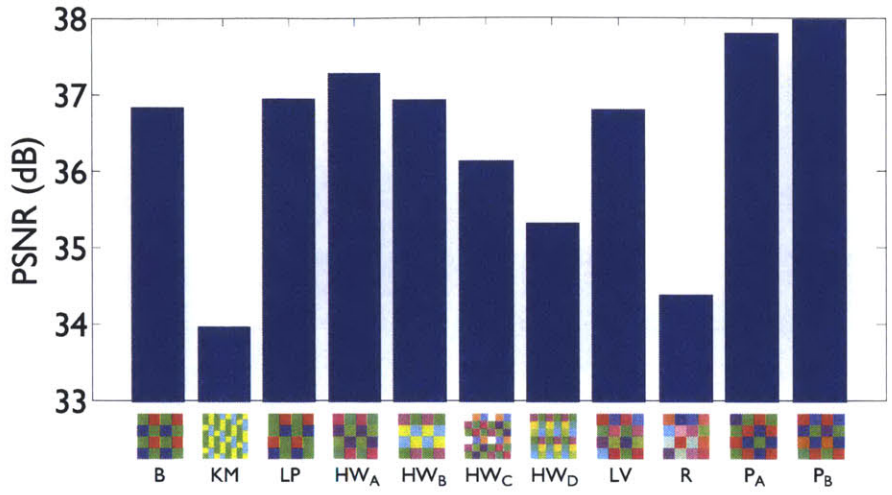


Figure 4.12: Summary of the average PSNR performance of various CFAs on test images.

optimal CFA, with some restrictions on search space and algorithm, because of technical feasibility, but not on the criterion itself.

In [148], Willet et al. invented a single-chip hyperspectral camera, where more than three spectra per pixel are estimated from the sensor outputs. They used a hand-designed Bayer-like pattern to build a spectral filter array. This is not within the scope of this study, but our results suggest that a better spectral filter array may be learnable using the statistics of target data and our proposed algorithm may be useful for such a generalized application as well.

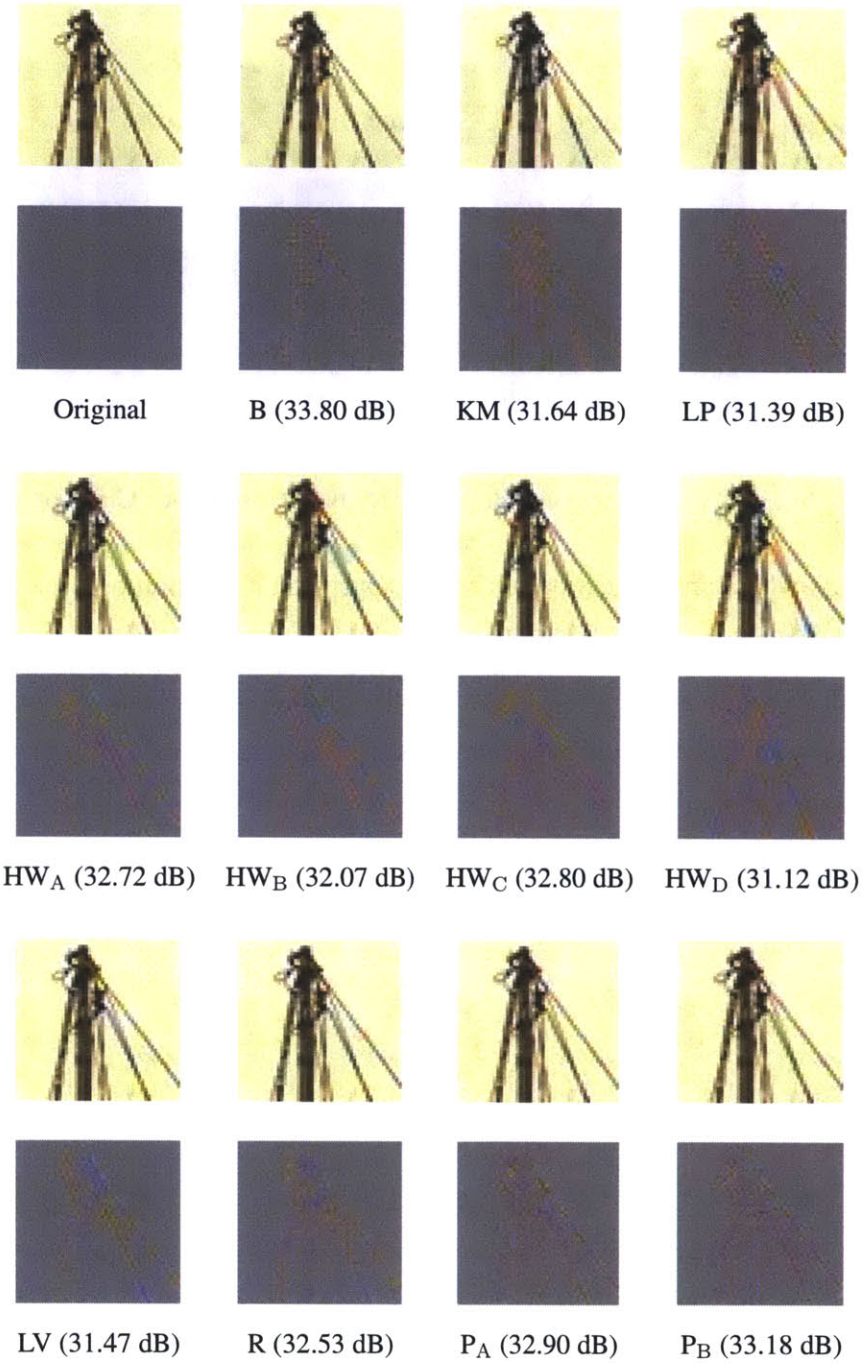


Figure 4.13: Reconstruction and error residual of a selected part from Image 1, for various CFAs. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe (**HW$_A$**–**HW$_D$**), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns (**P$_A$**, **P$_B$**). For recovery, the MoG-based MMSE has been commonly used. Error residual has been measured by reconstruction minus original, with zero displayed in mid-gray. The PSNR score, assessed only on the selected part, is given in parenthesis.

Figure 4.14: Reconstruction and error residual of a selected part from Image 4, for various CFAs. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe (**HW$_A$**–**HW$_D$**), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns (**P$_A$**, **P$_B$**). For recovery, the MoG-based MMSE has been commonly used. Error residual has been measured by reconstruction minus original, with zero displayed in mid-gray. The PSNR score, assessed only on the selected part, is given in parenthesis.

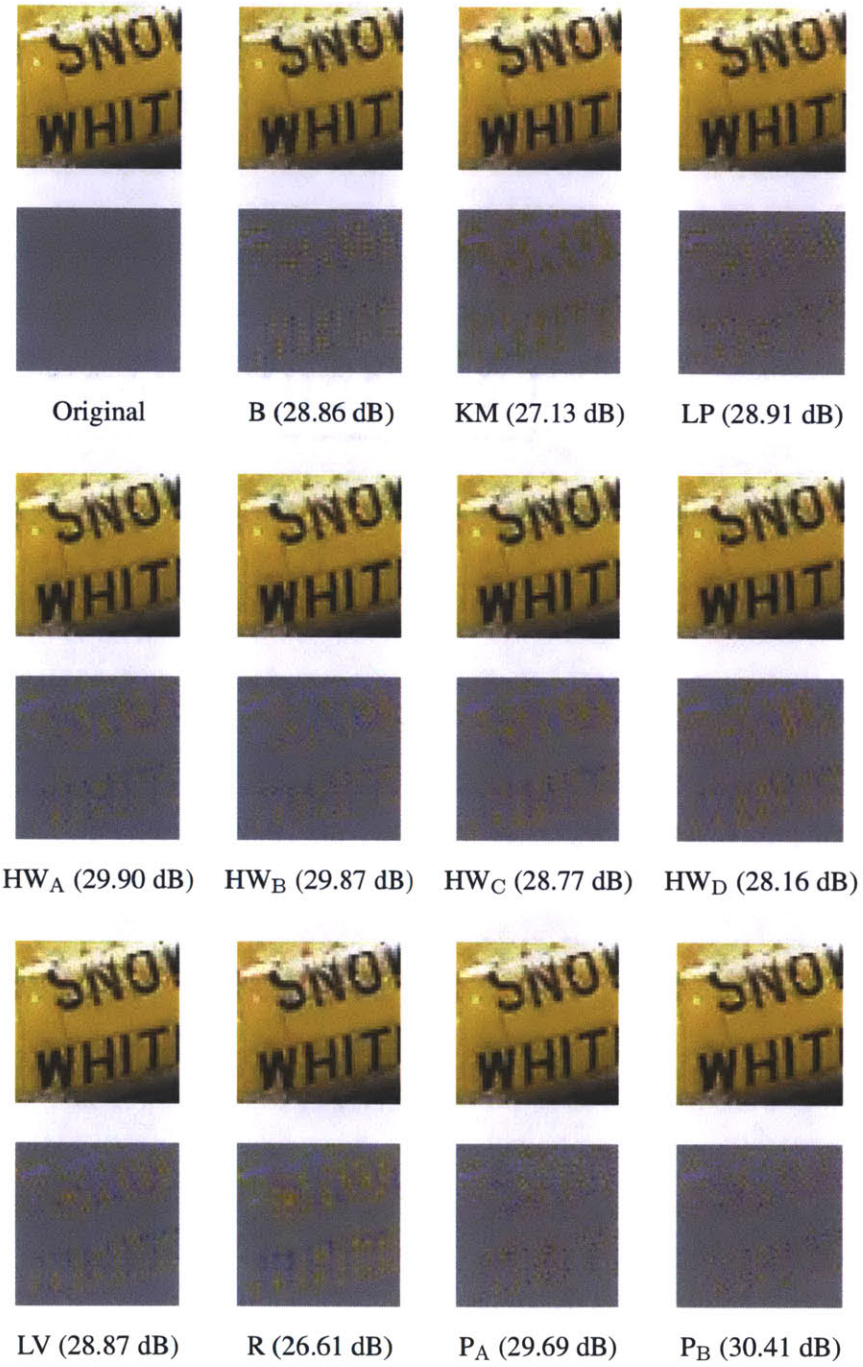


Figure 4.15: Reconstruction and error residual of a selected part from Image 5, for various CFAs. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe (**HW$_A$**–**HW$_D$**), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns (**P$_A$**, **P$_B$**). For recovery, the MoG-based MMSE has been commonly used. Error residual has been measured by reconstruction minus original, with zero displayed in mid-gray. The PSNR score, assessed only on the selected part, is given in parenthesis.

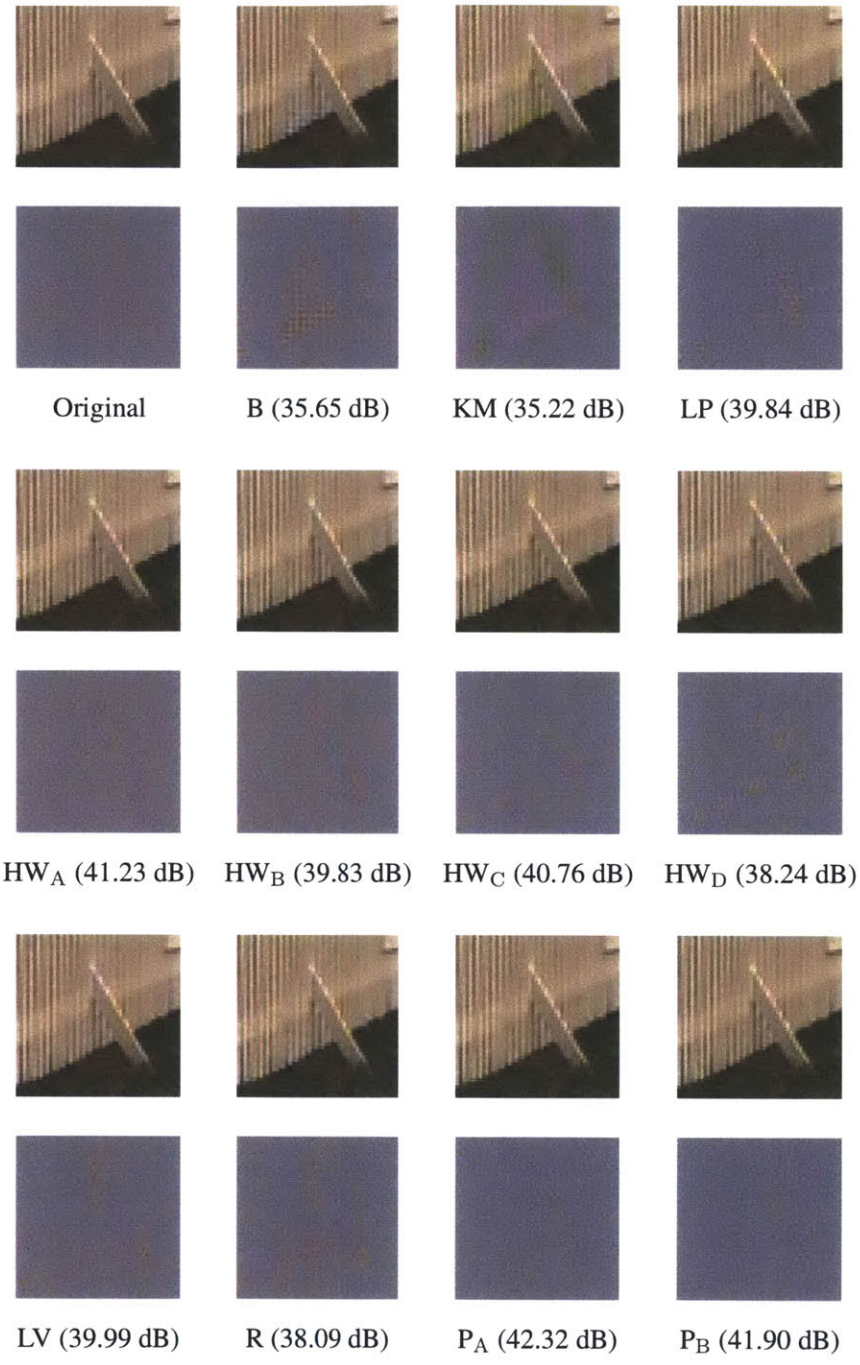


Figure 4.16: Reconstruction and error residual of a selected part from Image 10, for various CFAs. The compared CFAs are Bayer (**B**), Knop & Morf (**KM**), Lukac & Plataniotis (**LP**), Hirakawa & Wolfe (**HW**$_A$–**HW**$_D$), Lu & Vetterli (**LV**), random (**R**), and the proposed patterns (**P**$_A$, **P**$_B$). For recovery, the MoG-based MMSE has been commonly used. Error residual has been measured by reconstruction minus original, with zero displayed in mid-gray. The PSNR score, assessed only on the selected part, is given in parenthesis.

Chapter 5

Conclusions

The rationale for the classical theory of compressed sensing is that it is based on a good model (i.e. sparsity) of the input signals. When the sparsity prior better describes the input signals than limited bandwidth in Fourier domain, the signals can be recovered from sub-Nyquist rate random projections. The same rationale applies to informative sensing. What are the best projections if we have further accurate information (i.e., probability density) which can distinguish input signals from the others of the same sparsity level? The goal of this thesis was to answer this question, by providing a set of principles, analytical results, and computational algorithms.

As the central principle, we proposed that the uncertainty of the hidden signal should be minimized given the undercomplete projection. This is the view of Bayesian experimental design and also of the InfoMax principle if Shannon's entropy is used to measure the uncertainty. This formalism is generally applicable to any signals, not only to sparse ones, if the prior density is well-defined.

In the analytical part, we focused on signals which have a sparse representation in an orthonormal basis and managed to solve the InfoMax up to an approximation. The sparsity model in an orthonormal basis is common in the compressed sensing literature. Thus, we were really to see what we could tell, beyond the classical theory, about the optimal measurement matrix. Our findings are summarized as follows:

1. If the coefficients of the sparsifying basis are i.i.d., random projections are asymptot-

ically InfoMax optimal.

2. If the coefficients of the sparsifying basis are not i.i.d., InfoMax may produce instead a novel set of projections. In general, the set can be approximately represented as a combination of a certain number of PCA projections plus the remaining number of projections restricted to multiplexing over a particular linear subspace. The optimal parameters (the number of PCA projections, the linear subspace over which multiplexing is taken, etc.) are determined by a sort of water-filling.

Particularly if the coefficients are groupwise i.i.d., groupwise random projections with nonuniform sampling rate per group are asymptotically InfoMax optimal. Such a groupwise i.i.d. pattern roughly appears in natural images if the wavelet basis is partitioned into groups according the scale. Consequently, we applied the groupwise random projections to the sensing of natural images. In the presence of noise, we presented an algorithm to optimally distribute power among the sensors within a given budget, which generalized Linsker's result (on Gaussian signals) to non-Gaussian natural images.

In the last part of the thesis, we designed color filter arrays (CFAs) for the use in single-chip digital cameras. The CFA-type sensing is notionally a special case of compressed sensing, but the classical theory of compressed sensing is hard to apply because the feasible measurement matrices are constrained by physical nature. Informative sensing was still applicable. We showed how to learn a CFA that minimizes the uncertainty of the missing color components, given the measured ones.

Throughout all the parts, we provided experimental results that the "informative" projections consistently outperform others in signal reconstruction. We also found a few theoretical connections to the existing approaches in the literature of compressed sensing, some of which had remained heuristic.

In summary, we gave some analytical results and algorithms on specific signal models and on specific applications, but a number of other issues still remain unexplored. For example, the groupwise i.i.d. prior may not be the best model for natural images, while being better than the simple sparsity or limited bandwidth prior. Perhaps the dependencies among wavelet coefficients in the local neighborhood or along the tree hierarchy may be

further exploited.

Needless to say, recovery is very important in the whole system. Informative sensing itself did not assume any specific recovery scheme, but the maximal effects were attainable with a good recovery scheme based on accurate priors. The efficient computation of the MMSE estimate, one of the emerging research trends in compressed sensing, is where we also need future research for informative sensing.

Appendix A

Generalized Gaussian Distribution

The generalized Gaussian density is a family of probability density functions (pdfs) parameterized by

$$p(x) = \frac{1}{Z} e^{-|x/C|^r} \tag{A.1}$$

if the random variable x is normalized to have zero-mean and unit-variance. In (A.1), C and Z are functions of r, each given by $C = \sqrt{\frac{\Gamma(1/r)}{\Gamma(3/r)}}$ and $Z = \frac{2\Gamma(1/r)\sqrt{\Gamma(1/r)}}{r\sqrt{\Gamma(3/r)}}$, respectively. The generalized Gaussian includes several well-known pdfs in its family: Gaussian with

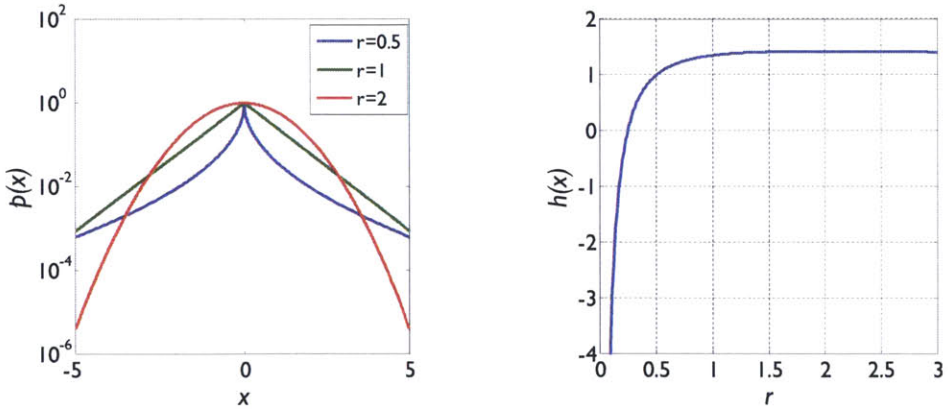


Figure A.1: Generalized Gaussian distribution. Left: probability density function (in log scale), Right: entropy as a function of the shape parameter r.

$r = 2$, Laplacian with $r = 1$, uniform with $r \to \infty$, degenerated delta function with $r \to 0$.

Generally, as r decreases, the distribution becomes more heavy-tailed (see Figure A.1, left). The entropy is known to be

$$h(x) = \mathbb{E}\left[-\log p(x)\right] = \frac{1}{r} - \frac{1}{2}\log\left(\frac{r^2\Gamma(\frac{3}{r})}{4\Gamma^3(\frac{1}{r})}\right), \tag{A.2}$$

which rapidly increases with r if $r < 2$; peaks at $r = 2$; and then slowly decreases (see Figure A.1, right).

Bibliography

[1] Shuchin Aeron, Venkatesh Saligrama, and Manqi Zhao. Information theoretic bounds for compressed sensing. *IEEE Trans. Inf. Theory*, 56(10):5111–5130, Oct. 2010.

[2] Michal Aharon, Michael Elad, and Alfred Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Trans. Signal Process.*, 54(11):4311–4322, Nov. 2006.

[3] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Trans. Comput*, 23(1):90–93, Jan. 1974.

[4] Mehmet Akçakaya, Seunghoon Nam, Peng Hu, Mehdi H. Moghari, Long H. Ngo, Vahid Tarokh, Warren J. Manning, and Reza Nezafat. Compressed sensing with wavelet domain dependencies for coronary MRI: A retrospective study. *IEEE Trans. Med. Imaging*, 30(5):1090–1099, May 2011.

[5] Mehmet Akçakaya and Vahid Tarokh. Shannon-theoretic limits on noisy compressive sampling. *IEEE Trans. Inf. Theory*, 56(1):492–504, Jan. 2010.

[6] David Alleysson, Sabine Süsstrunk, and Jeanny Hérault. Linear demosaicing inspired by the human visual system. *IEEE Trans. Image Process.*, 14(4):439–449, 2005.

[7] Joseph J. Atick. Could information theory provide an ecological theory of sensory processing? *Network: Comput. Neural Syst.*, 3:213–251, 1992.

[8] Fred Attneave. Informational aspects of visual perception. *Psych. Rev.*, 61:183–193, 1954.

[9] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a Mahalanobis metric from equivalence constraints. *J. Mach. Learn. Res.*, 6:937–965, June 2005.

[10] Richard G. Baraniuk, Volkan Cevher, Marco F. Duarte, and Chinmay Hegde. Model-based compressive sensing. *IEEE Trans. Inf. Theory*, 56(4):1982 – 2001, Apr. 2010.

[11] Richard G. Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx.*, 28(3):253–263, Dec. 2008.

[12] Horace B. Barlow. Possible principles underlying the transformation of sensory messages. In W. A. Rosenblith, editor, *Sensory Communications*, pages 217–234. MIT Press, Cambridge, MA, 1961.

[13] Ole E. Barndorff-Nielsen and David R. Cox. *Asymptotic Techniques for Use in Statistics*. Monographs on Statistics and Applied Probability. Chapman and Hall, 1989.

[14] Dror Baron, Shriram Sarvotham, and Richard G. Baraniuk. Bayesian compressive sensing via belief propagation. *IEEE Trans. Signal Process.*, 58(1):269–280, Jan. 2010.

[15] Maurice Stevenson Bartlett. An inverse matrix adjustment arising in discriminant analysis. *Ann. Math. Statist.*, 22(1):107–111, 1951.

[16] Bryce E. Bayer. Color imaging array. US Patent 3 971 065, 1976.

[17] Suzanna Becker and Geoffrey E. Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, pages 161–163, 1992.

[18] Anthony J. Bell and Terrence J. Sejnowski. A non-linear information maximisation algorithm that performs blind separation. In *Advances in Neural Information Processing Systems*, volume 7, pages 467–474, 1995.

[19] Anthony J. Bell and Terrence J. Sejnowski. Edges are the independent components of natural scenes. In *Advances in Neural Information Processing Systems*, volume 9, pages 831–837, 1997.

[20] Radu Berinde and Piotr Indyk. Sparse recovery using sparse random matrices. Technical report, MIT, 2008.

[21] Matthias Bethge. Factorial coding of natural images: How effective are linear models in removing higher-order dependencies? *J. Opt. Soc. Am. A*, 23(6):1253–1268, June 2006.

[22] Petros Boufounos and Richard G. Baraniuk. 1-bit compressive sensing. In *Conf. on Information Sciences and Systems*, Princeton, NJ, Mar. 2008.

[23] David H. Brainard. Bayesian method for reconstructing color images from trichromatic samples. In *Proc. IS&T 47th Annual Conf.*, pages 375–380, Rochester, NY, 1994.

[24] Peter J. Burt and Edward H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Trans. Comm.*, 31(4):532–540, Apr. 1983.

[25] Robert Calderbank, Stephen Howard, and Sina Jafarpour. Construction of a large class of deterministic sensing matrices that satisfy a statistical isometry property. *IEEE J. Sel. Topics Signal Process.*, 4(2):358–374, Apr. 2010.

[26] Emmanuel J. Candès. Compressive sampling. In *Proc. International Congress of Mathematicians*, pages 1433–1452, Madrid, Spain, 2006.

[27] Emmanuel J. Candès and Yaniv Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory*. to appear.

[28] Emmanuel J. Candès and Justin Romberg. Sparsity and incoherence in compressive sampling. *Inverse Prob.*, 23(3):969–986, June 2007.

[29] Emmanuel J. Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, Feb. 2006.

[30] Emmanuel J. Candès, Justin Romberg, and Terrence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

[31] Emmanuel J. Candès and Terence Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, Dec. 2005.

[32] Emmanuel J. Candès and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, Dec. 2006.

[33] Jean-François Cardoso. Infomax and maximum likelihood for blind source separation. *IEEE Signal Process. Lett.*, 4(4):112–114, Apr. 1997.

[34] Rui M. Castro, Jarvis Haupt, Robert Nowak, and Gil M. Raz. Finding needles in noisy haystacks. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 5133–5136, Mar. 2008.

[35] Volkan Cevher. Learning with compressible priors. In *Advances in Neural Information Processing Systems*, 2009.

[36] Hyun Sung Chang, Yair Weiss, and William T. Freeman. Informative sensing of natural images. In *Proc. IEEE Int. Conf. on Image Processing*, pages 3025–3028, Nov. 2006.

[37] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, 20:33–61, 1998.

[38] Taeg Sang Cho, Neel Joshi, C. Lawrence Zitnick, Sing Bing Kang, Rick Szeliski, and William T. Freeman. A content-aware image prior. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010.

[39] King-Hong Chung and Yuk-Hee Chan. Color demosaicing using variance of color differences. *IEEE Trans. Image Process.*, 15(10):2944–2955, Oct. 2006.

[40] Albert Cohen, Wolfgang Dahmen, and Ronald DeVore. Compressed sensing and best k-term approximation. *J. Amer. Math. Soc.*, 22(1):211–231, Jan. 2009.

[41] Ronald Coifman, Frank Geschwind, and Yves Meyer. Noiselets. *Appl. Comput. Harmon. Anal.*, 10:27–44, 2001.

[42] Thomas H. Cormen, Charles E, Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to Algorithms*. MIT Press, Cambridge, MA, 2009.

[43] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, NY, 1991.

[44] Imre Csiszár and Paul C. Shields. *Information Theory and Statistics: a Tutorial*, volume 1 of *Foundations and Trends in Communications and Information Theory*. Now Pub. Inc., 2004.

[45] Sanjoy Dasgupta, Daniel Hsu, and Nakul Verma. A concentration theorem for projections. In *Proc. of 22nd Conf. on Uncertainty in Artificial Intelligence*, July 2006.

[46] Mark A. Davenport. *Random Observations on Random Observations: Sparse Signal Acquisition and Processing*. PhD thesis, Rice University, 2010.

[47] Mark A. Davenport, Marco F. Duarte, Yonina C. Eldar, and Gitta Kutyniok. *Introduction to Compressed Sensing*, chapter 1. Compressed Sensing: Theory and Applications. Cambridge Univ. Press, 2012.

[48] Brice Chaix de Lavarène, David Alleysson, and Jeanny Hérault. Practical implementation of LMMSE demosaicing using luminance and chrominance spaces. *Comput. Vis. Image Underst.*, 107:3–13, July 2007.

[49] Ronald A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51–150, 1998.

[50] Ronald A. DeVore. Deterministic constructions of compressed sensing matrices. *J. Complex.*, 23(4):918–925, 2007.

[51] David L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, Apr. 2006.

[52] David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proc. Natl. Acad. Sci.*, 100(5):2197–2202, Mar. 2003.

[53] David L. Donoho, Michael Elad, and Vladimir Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Trans. Inf. Theory*, 52(1):6–18, Jan. 2006.

[54] Marco Duarte, Mark Davenport, Dharmpal Takhar, Jason Laska, Ting Sun, Kevin Kelly, and Richard Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):83–91, Mar. 2008.

[55] Julio Martin Duarte-Carvajalino and Guillermo Sapiro. Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization. *IEEE Trans. Image Process.*, 18(7):1395–14087, July 2009.

[56] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2001.

[57] Michael Elad. Optimized projections for compressed sensing. *IEEE Trans. Signal Process.*, 55(12):5695–5702, Dec. 2007.

[58] Michael Elad and Michal Aharon. Image denoising via learned dictionaries and sparse representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, New York, NY, June 2006.

[59] Michael Elad, Peyman Milanfar, and Ron Rubinstein. Analysis versus synthesis in signal priors. In *Proc. European Signal Processing Conf.*, Sept. 2006.

[60] Yonina C. Eldar. Compressed sensing of analog signals in shift-invariant spaces. *IEEE Trans. Signal Process.*, 57(8):2986–2997, Aug. 2009.

[61] Deniz Erdogmus, Jose C. Principe, and Kenneth E. Hild II. Do Hebbian synapses estimate entropy? In *Proc. IEEE Workshop on Neural Networks for Signal Processing*, pages 199–208, 2002.

[62] Matthew Fickus, Dustin G. Mixon, and Janet C. Tremain. Steiner equiangular tight frames. preprint (submitted to *Linear Algebra Appl.*), 2010.

[63] Alyson K. Fletcher, Sundeep Rangan, and Vivek K. Goyal. Necessary and sufficient conditions for sparsity pattern recovery. *IEEE Trans. Inf. Theory*, 55(12):5758–5772, Dec. 2009.

[64] Wilson S. Geisler, Jiri Najemnik, and Almon D. Ing. Optimal stimulus encoders for natural tasks. *J. Vis.*, 9(13):1–16, 2009.

[65] Semyon Aranovich Geršgorin. Über die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Nauk. Otd. Fiz.-Mat. Nauk*, 7:749–754, 1931.

[66] Rémi Gribonval and Morten Nielson. Sparse representations in unions of bases. *IEEE Trans. Inf. Theory*, 49(12):3320–3325, Dec. 2003.

[67] Bahadir K. Gunturk, Yucel Altunbasak, and Russel M. Mersereau. Color plane interpolation using alternating projections. *IEEE Trans. Image Process.*, 11(9):997–1013, Sep. 2002.

[68] Bahadir K. Gunturk, John Glotzbach, Yucel Altunbasak, Ronald W. Schafer, and Russel M. Mersereau. Demosaicking: Color filter array interpolation. *IEEE Signal Process. Mag.*, 22(1):44–54, Jan. 2005.

[69] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Mutual information and minimum mean-square error in Gaussian channels. *IEEE Trans. Inf. Theory*, 51(4):1261–1282, Apr. 2005.

[70] Dongning Guo, Shlomo Shamai, and Sergio Verdú. Proof of entropy power inequalities via MMSE. In *Proc. IEEE International Symposium on Information Theory*, pages 1011–1015, July 2006.

[71] Dongning Guo and Sergio Verdú. Randomly spread CDMA: Asymptotics via statistical physics. *IEEE Trans. Inf. Theory*, 51(6):1983–2010, June 2005.

[72] Dongning Guo and Chih-Chun Wang. Asymptotic mean-square optimality of belief propagation for sparse linear systems. In *Proc. IEEE Information Theory Workshop*, pages 194–198, Oct. 2006.

[73] David K. Hammond and Eero P. Simoncelli. Image denoising with an orientation-adaptive Gaussian scale mixture model. In *Proc. IEEE Conf. on Image Processing*, Oct. 2006.

[74] Jarvis Haupt and Robert Nowak. Compressive sampling vs. conventional imaging. In *Proc. IEEE Int. Conf. on Image Processing*, pages 1269–1272, Atlanta, GA, Oct. 2006.

[75] Jarvis Haupt and Robert Nowak. Signal reconstruction from noisy random projections. *IEEE Trans. Inf. Theory*, 52(9):4036–4048, Sept. 2006.

[76] Keigo Hirakawa and Thomas W. Parks. Adaptive homogeneity-directed demosaicing algorithm. *IEEE Trans. Image Process.*, 14(3):360–369, Mar. 2005.

[77] Keigo Hirakawa and Patrick J. Wolfe. Spatio-spectral color filter array design for optimal image recovery. *IEEE Trans. Image Process.*, 17(10):1876–1890, Oct. 2008.

[78] Junzhou Huang, Xiaolei Huang, and Dimitris Metaxas. Learning with dynamic group sparsity. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 64–71, Kyoto, Japan, Sept. 2009.

[79] Aapo Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Trans. Neural Networks*, 10(3):626–634, May 1999.

[80] Piotr Indyk. Explicit constructions for compressed sensing of sparse signals. In *Proc. ACM-SIAM Symp. on Discrete Algorithms*, San Francisco, CA, Jan. 2008.

[81] Sina Jafarpour, Weiyu Xu, Babak Hassibi, and Robert Calderbank. Efficient and robust compressed sensing using optimized expander graphs. *IEEE Trans. Inf. Theory*, 55(9):4299–4308, Sept. 2009.

[82] Edwin Thompson Jaynes. Information theory and statistical mechanics. *Phys. Rev. Ser. II*, 106(4):May, June 1957.

[83] Shihao Ji, Ya Xue, and Lawrence Carin. Bayesian compressive sensing. *Proc. Nat. Acad. Sci.*, 56(6):2346–2356, June 2008.

[84] M. Chris Jones and Robin Sibson. What is projection pursuit? *J. R. Statist. Soc. A*, 150(1):1–37, 1987.

[85] Karl Knop and Rudolf Morf. A new class of mosaic color encoding patterns for single-chip cameras. *IEEE Trans. Electron. Dev.*, 32(8):1390–1395, Aug. 1985.

[86] Dilip Krishnan and Rob Fergus. Fast image deconvolution using hyper-Laplacian priors. In *Advances in Neural Information Processing Systems*, 2009.

[87] Arun Kumar, Allen R. Tannenbaum, and Gary J. Balas. Optimal flow: A curve evolution approach. *IEEE Trans. Image Process.*, 5(4):598–610, 1996.

[88] Jason N. Laska, Sami Kirolos, Marco F. Duarte, Tamer S. Ragheb, Richard G. Baraniuk, and Yehia Massoud. Theory and implementation of an analog-to-information converter using random demodulation. In *Proc. IEEE Int. Symp. on Circuits and Systems*, pages 1959–1962, New Orleans, LA, May 2007.

[89] Xin Li, Bahadir Gunturk, and Lei Zhang. Image demosaicing: A systematic survey. In *Proc. SPIE Visual Communications and Image Processing*, volume 6822, Jan. 2008.

[90] Zhaoping Li. Optimal sensory encoding. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks: The Second Edition*, pages 815–819. MIT Press, 2002.

[91] Nai-Xiang Lian, Lanlan Chang, Yap-Peng Tan, and Vitali Zagorodnov. Adaptive filtering for color filter array demosaicking. *IEEE Trans. Image Process.*, 16(10):2515–2525, Oct. 2007.

[92] Tim Lin and Felix J. Herrmann. Compressed wavefield extrapolation. *Geophysics*, 72(5):SM77–SM93, Sept./Oct. 2007.

[93] Dennis Victor Lindley. On a measure of information provided by an experiment. *Ann. Math. Statist.*, 27(4):986–1005, 1956.

[94] Ralph Linsker. An application of the principle of maximum information preservation to linear systems. In *Advances in Neural Information Processing Systems*, volume 1, pages 186–194, 1989.

[95] Yue M. Lu and Martin Vetterli. Optimal color filter array design: Quantitative conditions and an efficient search procedure. In *Proc. IS&T/SPIE Conf. on Digital Photography V*, volume 7250, San Jose, CA, Jan. 2009.

[96] Rastislav Lukac and Konstantinos N. Plataniotis. Color filter arrays: Design and performance analysis. *IEEE Trans. Consum. Electron.*, 51(4):1260–1267, Nov. 2005.

[97] Rastislav Lukac, Konstantinos N. Plataniotis, Dimitrios Hatzinakos, and Marko Aleksic. A novel cost effective demosaicing approach. *IEEE Trans. Consum. Electron.*, 50(1):256–261, Feb. 2004.

[98] Michael Lustig, David L. Donoho, and John M. Pauly. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med.*, 58(6):1182–1195, Dec. 2007.

[99] Michael Lustig, Juan M. Santos, David L. Donoho, and John M. Pauly. k-t sparse: High frame rate dynamic mri exploiting spatio-temporal sparsity. In *Proc. Annual Meeting of ISMRM*, 2006.

[100] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE Trans. Image Process.*, 17(1):53–69, Jan. 2008.

[101] David R. Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proc. IEEE Int. Conf. on Computer Vision*, pages 416–423, July 2001.

[102] Julian J. McAuley, Tibério S. Caetano, and Matthias O. Franz. Learning high-order MRF priors of color images. In *Proc. Int'l Conf. Machine Learning*, pages 617–624, June 2006.

[103] Daniele Menon, Stefano Andriani, and Giancarlo Calvagno. Demosaicing with directional filtering and a posteriori decision. *IEEE Trans. Image Process.*, 16(1):132–141, Jan. 2007.

[104] Thomas P. Minka. Expectation propagation for approximate Bayesian inference. In *Proc. of 17th Conf. on Uncertainty in Artificial Intelligence*, Aug. 2001.

[105] Donald F. Morrison. *Multivariate Statistical Methods.* McGraw-Hill, Inc., New York, NY, 1967.

[106] Srinivasa G. Narasimhan and Shree K. Nayar. Enhancing resolution along multiple imaging dimensions using assorted pixels. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(4):518–530, Apr. 2005.

[107] Guy Philip Nason. *Design and Choice of Projection Indices.* PhD thesis, University of Bath, 1992.

[108] Shree K. Nayar and Srinivasa G. Narasimhan. Assorted pixels: Multi-sampled imaging with structural models. In *Proc. European Conference on Computer Vision*, pages 636–652, May 2002.

[109] Radford M. Neal and Geoffrey E. Hinton. *A View of the EM Algorithm that Justifies Incremental, Sparse, and Other Variants*, pages 355–368. Learning in Graphical Models. MIT Press, 1999.

[110] Ramesh Neelamani, Christine E. Krohn, Jerry R. Krebs, Justin K. Romberg, Max Deffenbaugh, and John E. Anderson. Efficient seismic forward modeling using simultaneous random sources and sparsity. *Geophysics*, 75, Dec. 2010.

[111] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing.* Prentice-Hall, Inc., Englewood Cliffs, NJ, 1989.

[112] Maxim Raginsky, Rebecca M. Willett, Zachary T. Harmany, and Roummel F. Marcia. Compressed sensing performance bounds under Poisson noise. *IEEE Trans. Signal Process.*, 58(8):3990–4002, Aug. 2010.

[113] Sundeep Rangan. Estimation with random linear mixing, belief propagation and compressed sensing. arXiv:1001.2228v2 [cs.IT], May 2010.

[114] Sundeep Rangan, Alyson K. Fletcher, and Vivek K. Goyal. Asymptotic analysis of MAP estimation via the replica method and applications to compressed sensing. *IEEE Trans. Inf. Theory*, 58(3):1902–1923, Mar. 2012.

[115] Justin Romberg. Imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):14–20, Mar. 2008.

[116] Justin Romberg and Michael Wakin. Compressed sensing: A tutorial. In *IEEE Statistical Signal Processing Workshop*, Madison, WI, Aug. 2007.

[117] Stefan Roth and Michael J. Black. Fields of experts: A framework for learning image priors. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 860–867, San Diego, CA, June 2005.

[118] Stefan Roth and Michael J. Black. Fields of experts. *Int. J. Comput. Vis.*, 82(2):205–229, Apr. 2009.

[119] Mark Rudelson and Roman Vershynin. On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.*, 61(8):1025–1045, Aug. 2008.

[120] Daniel L. Ruderman. Origins of scaling in natural images. *Vis. Res.*, 37(23):3385–3398, 1997.

[121] Aswin C. Sankaranarayanan. *Robust and Efficient Inference of Scene and Object Motion in Multi-Camera Systems*. PhD thesis, University of Maryland, College Park, 2009.

[122] Shriram Sarvotham, Dror Baron, and Richard G. Baraniuk. Measurements vs. bits: Compressed sensing meets information theory. In *Proc. Allerton Conf. on Communication, Control, and Computing*, Monticello, IL, Sept. 2006.

[123] Uwe Schmidt, Qi Gao, and Stefan Roth. A generative perspective on MRFs in low-level vision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, San Francisco, CA, June 2010.

[124] Paola Sebastiani and Henry P. Wynn. Maximum entropy sampling and optimal Bayesian experimental design. *J. R. Statist. Soc. B*, 62:145–157, 2000.

[125] Matthias W. Seeger. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, 2008.

[126] Matthias W. Seeger and Hannes Nickisch. Compressed sensing and Bayesian experimental design. In *Proc. Int. Conf. on Machine Learning*, pages 912–919, June 2008.

[127] Matthias W. Seeger and Hannes Nickisch. Compressed sensing and Bayesian experimental design. In *Proc. Int. Conf. on Machine Learning*, pages 912–919, June 2008.

[128] Matthias W. Seeger, Hannes Nickisch, Rolf Pohmann, and Bernhard Schölkopf. Bayesian experimental design of magnetic resonance imaging sequences. In *Advances in Neural Information Processing Systems*, 2008.

[129] Xianbiao Shu and Narendra Ahuja. Hybrid compressive sampling via a new total variation TVL1. In *Proc. European Conf. on Computer Vision*, pages 393–404, Sept. 2010.

[130] Eero P. Simoncelli. Statistical models for images: Compression, restoration and synthesis. In *Proc. Asilomar Conf. on Signals, Systems and Computers*, pages 673–678, Nov. 1997.

[131] Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA, 2009.

[132] Mátyás A. Sustik, Joel A. Tropp, Inderjit S. Dhillon, and Robert W. Heath Jr. On the existence of equiangular tight frames. *Linear Algebra Appl.*, 426:619–635, 2007.

[133] David Taubman. Generalized Wiener reconstruction of images from colour sensor data using a scale invariant prior. In *Proc. IEEE Int. Conf. on Image Processing*, pages 801–804, Vancouver, BC, Sep. 2000.

[134] Robert Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B*, 58(1):267–288, 1996.

[135] Joel A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory*, 50(10):2231–2242, Oct. 2004.

[136] Joel A. Tropp and Anna Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory*, 53(12):4655–4666, Dec. 2007.

[137] Joel A. Tropp, Michael B. Wakin, Marco F. Duarte, Dror Baron, and Richard G. Baraniuk. Random filters for compressive sampling and reconstruction. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pages 872–875, Toulouse, France, May 2006.

[138] H. J. Trussell and Robert E. Hartwig. Mathematics for demosaicking. *IEEE Trans. Image Process.*, 11(4):485–492, Apr. 2002.

[139] Antonia M. Tulino and Sergio Verdú. *Random Matrix Theory and Wireless Communications*, volume 1 of *Foundations and Trends in Communications and Information Theory*. Now Pub. Inc., 2004.

[140] A. van der Schaaf and J.H. van Hateren. Modelling the power spectra of natural images: statistics and information. *Vision Research*, 36(17):2759–2770, 1996.

[141] Roman Vershynin. *Introduction to the Non-Asymptotic Analysis of Random Matrices*, chapter 5. Compressed Sensing: Theory and Applications. Cambridge Univ. Press, 2012.

[142] Martin Vetterli, Pina Marziliano, and Thierry Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.*, 50(6):1417–1428, June 2002.

[143] Martin J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. In *In Proc. Allerton Conference on Communication, Control and Computing*, 2006.

[144] Martin J. Wainwright. Information-theoretic limitations on sparsity recovery in high-dimensional and noisy setting. *IEEE Trans. Inf. Theory*, 55(12):5728–5741, Dec. 2009.

[145] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, Apr. 2004.

[146] Yair Weiss, Hyun Sung Chang, and William T. Freeman. Learning compressed sensing. In *Proc. Allerton Conf. on Communication, Control, and Computing*, Sept. 2007.

[147] Yair Weiss and William T. Freeman. What makes a good model of natural images? In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, Minneapolis, MN, June 2007.

[148] Rebecca M. Willett, Michael E. Gehm, and David J. Brady. Multiscale reconstruction for computational spectral imaging. In *Proc. SPIE Computational Imaging V*, Feb. 2007.

[149] Yihong Wu and Sergio Verdú. Rényi information dimension: Fundamental limits of almost lossless analog compression. *IEEE Trans. Inf. Theory*, 56(8):3721–3748, Aug. 2010.

[150] Jianping Xu, Yiming Pi, and Zongjie Cao. Optimized projection matrix for compressive sensing. *EURASIP J. Adv. Signal Process.*, 2010. Article ID 560349.

[151] Lei Yu, Jean Pierre Barbot, Gang Zheng, and Hong Sun. Compressive sensing with chaotic sequence. *IEEE Signal Process. Lett.*, 17(8):731–734, Aug. 2010.

[152] Lei Zhang and Xiaolin Wu. Color demosaicking via directional linear minimum mean square-error estimation. *IEEE Trans. Image Process.*, 14(12):2167–2178, Dec. 2005.

[153] Tong Zhang. Sparse recovery with orthogonal matching pursuit under RIP. *IEEE Trans. Inf. Theory*, 57(9):6215–6221, Sept. 2011.

[154] Song Chun Zhu and David Mumford. Prior learning and Gibbs reaction-diffusion. *IEEE Trans. Pattern Anal. Machine Intell.*, 19(11):1236–1250, Nov. 1997.

[155] Daniel Zoran and Yair Weiss. Scale invariance and noise in natural images. In *Proc. IEEE Int. Conf. on Computer Vision*, Kyoto, Japan, Sept. 2009.

[156] Daniel Zoran and Yair Weiss. From learning models of natural image patches to whole image restoration. In *Proc. IEEE Int. Conf. on Computer Vision*, Barcelona, Spain, Nov. 2011.

[157] Argyrios Zymnis, Stephen Boyd, and Emmanuel Candès. Compressed sensing with quantized measurements. *IEEE Signal Process. Lett.*, 17(2):149–152, Feb. 2010.