Distinguishing and Integrating Aleatoric and Epistemic Variation in Uncertainty Quantification/ Sparse Gradient Image Reconstruction from Fourier and Edge Measurements

by

Kamaljit Chowdhary

B.A., New York University; New York, NY 2004

M.Sc., Brown University; Providence, RI 2007

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2012

This dissertation by Kamaljit Chowdhary is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____ _____

Jan S. Hesthaven, Ph.D., Advisor

Recommended to the Graduate Council

Date_____ _____

Paul Dupuis, Ph.D., Advisor

Date_____ _____

Edward Walsh, Ph.D., Reader

Approved by the Graduate Council

Date_____ _____

Peter M. Weber, Dean of the Graduate School

Vitae

Kamaljit Chowdhary was born on Septmember 5, 1982 in Kenosha, Wisconsin. He graduated from New York University in December 2003 *magna cum laude* with a degree in mathematics.

Acknowledgements

I would like to sincerely thank my advisors Professor Jan Hesthaven and Professor Paul Dupuis for all their help, guidance, support, and patience with me during my six years here at Brown. They opened up my eyes to the world of research in applied mathematics and allowed me to take the time to learn. I am forever grateful for the skills my advisors have taught me, which will in no doubt prepare me for an exciting career in applied mathematics. I would also like to thank my collaborator Professor Edward Walsh for his extreme patience and help with my research.

Of course, I would like to thank my family and friends, including my mom, dad, brother, cousin, girlfriend, former and current graduate students for their support, patience and discussions, especially through the stresses of graduate school. They added balance to my graduate student life, without which I would probably go crazy.

Abstract of "Distinguishing and Integrating Aleatoric and Epistemic Variation in Uncertainty Quantification/ Sparse Gradient Image Reconstruction from Fourier and Edge Measurements" by Kamaljit Chowdhary, Ph.D., Brown University, May 2012

This dissertation is divided into two separate parts. In **Part I**, we develop methods to obtain information on a system when the distributions of some variables are known exactly, others are known only approximately, and perhaps others are not modeled as random variables at all. The main tool used is the duality between risk-sensitive integrals and relative entropy, and we obtain explicit bounds on standard performance measures (variances, exceedance probabilities) over families of distributions whose distance from a nominal distribution is measured by relative entropy. The evaluation of the risk-sensitive expectations is based on polynomial chaos expansions, which help keep the computational aspects tractable. In **Part II**, we propose a new algorithm to reconstruct sparse gradient images from Fourier and edge measurements. In many imaging applications, such as functional Magnetic Resonance Imaging (fMRI), full, uniformly-sampled Cartesian Fourier measurements are acquired to reconstruct an image. In order to reduce scan time and increase temporal resolution for fMRI studies, one would like to accurately reconstruct these images from the smallest possible set of Fourier measurements. Compressed Sensing (CS) has given rise to techniques that can provide exact and stable recovery of sparse images from a relatively small set of Fourier measurements. In particular, if the images are sparse with respect to their gradient, e.g., piece-wise constant, total-variation minimization techniques can be used to recover those images from a highly incomplete set of Fourier measurements. Our algorithm aims to further reduce the number of Fourier measurements required for exact or stable recovery by utilizing prior edge information from a high resolution reference image. This reference image, or more precisely, the fully sampled Fourier measurements of this reference image, is obtained prior to an fMRI study in order to provide edge information for the region of interest. By combining this edge information with CS techniques for sparse gradient images, numerical experiments show that we can further reduce the number of Fourier measurements required for exact or stable recovery by an additional factor of 1.6 to 3 , compared with CS techniques alone, without edge information.

Contents

List of Tables

List of Figures

Part I

Distinguishing and Integrating Aleatoric and Epistemic Variation in Uncertainty Quantification

Chapter One

Introduction

Uncertainty quantification refers to a broad set of techniques for understanding the impact of uncertainties in complicated mechanical, physical, and computational systems. In this context "uncertainty" can take on many meanings, but we follow the convention of dividing uncertainty into aleatoric and epistemic categories. In short, aleatoric uncertainty refers to inherent uncertainty due to stochastic or probabilistic variability. This type of uncertainty is *irreducible* in that there will always be positive variance since the underlying variables are truly random. Epistemic uncertainty refers to limited knowledge we may have about the model or system. This type of uncertainty is *reducible* in that if we have more information, e.g., take more measurements, then this type of uncertainty can be reduced. However, for many problems where uncertainty quantification is important, the acquisition of data is difficult or expensive. The epistemic uncertainty cannot be removed entirely, and so one needs modeling and computational techniques which can also accommodate this form of uncertainty.

Much of the work to date has focused on what one might call the propagation of uncertainty. Here, one characterizes uncertainty concerning one or more elements of a physical system via some probability distribution, and attempts to quantify how that uncertainty will be propagated throughout the system under its constitutive equations and laws. The simplest and most straightforward method to characterize this output distribution would be to sample from the input probability distribution, solve the system equations, and thereby produce output samples (i.e., standard Monte Carlo). Although this is very simple conceptually, it can be far from practical, especially when it is computationally intensive to construct the mapping that takes input variables to output. Straightforward schemes such as Monte Carlo would require that the system be solved for each sample of the random variable, a situation that is often not practical. In recent years efficient computational methods have been developed to calculate particular functionals of the induced distribution that may be required for a particular application (e.g., covariances or error probabilities). Most recently, polynomial chaos has become popular as an efficient method for approximating the distribution of the output variable.

We note that random variables with a well-defined and given distribution are often used in this context even when there is no justification for their use, as in the case of modeling error. In other words, it is (perhaps implicitly) assumed that epistemic uncertainty can be modeled by aleatoric uncertainty. One reason is that, at least as they have been developed to date, most computational techniques (e.g., polynomial chaos and Monte Carlo) are based on the assumption that the user can identify some "appropriate" distribution for each uncertain aspect of the system, regardless of the type of uncertainty, aleatoric or epistemic. If one is interested in just basic qualitative properties of the system then this may not be a central issue, since virtually any model of uncertainty will give information on the sensitivities of the system. However, when the intended use of uncertainty quantification is for regulatory assessment or some other application where performance measures are sensitive to distributional assumptions, the issue becomes much more important, and one should carefully distinguish how one accounts for the two types of uncertainty.

The aim of the present work is to describe an approach that (i) logically distinguishes those aspects of uncertainty that are treated as stochastic variability from other forms of uncertainty, (ii) in cases where a stochastic model is theoretically valid but for which determination of the distribution is not practical, gives bounds for performance measures that are valid for explicitly identified families of distributions, and (iii) is computationally feasible if ordinary uncertainty propagation is feasible. The intended audience is broad, including numerical analysts interested in robust performance bounds as well as applied probabilists for whom certain computational aspects of uncertainty quantification may be novel. Since a typical reader might not be familiar with the terminology and methods from both fields, we have included background material for both the probabilistic and numerical analysis approaches used to make the paper broadly accessible.

This organization of this work is as follows. In Chapter 2 we further discuss the distinction between aleatoric and epistemic uncertainty, and explain some of the limitations to achieving useful performance bounds in the presence of epistemic uncertainty. Chapter 3 introduces risk-sensitive performance measures and discusses their robust properties. Sev-

eral hybrid forms are introduced that are more useful than the simplest form when both aleatoric and epistemic uncertainties are present. Various properties of the risk-sensitive measures that are useful in applications (monotonicity, optimization) are also discussed. Chapter 4 presents numerical examples, reviews the computational methods used to evaluate the risk-sensitive performance measures, and finishes with a subsection of discussion and conclusions.

Chapter Two

Aleatoric and Epistemic Uncertainties

As noted in the Introduction, *uncertainty* can be divided into two categories, aleatoric and epistemic. From the perspective of mathematical formulation and modeling, aleatoric uncertainty is in some sense simpler. In contrast, epistemic uncertainty can mean different things in different contexts. *Lack of knowledge* is an ambiguous term that encompasses many different scenarios.

To illustrate, we consider an example. Consider a system described by some well-posed partial differential equation (PDE), but with an uncertain boundary condition. The boundary condition is expressed in terms of a random variable X (e.g., $u_x(0,t) = X$) with a particular fixed distribution, and the numerical analysis problem is to characterize the induced distribution of the solution to the PDE at some time and location (e.g., $u(x_0, t_0; X)$). Suppose we know that the boundary condition is properly modeled as a random variable, but we do not know the correct value of some parameter in that distribution. For example, based on a central limit type argument, one might claim the distribution is known to be Gaussian, but still unknown are the "true" mean and/or variance. In this case, the *lack of knowledge* is this missing information about the parameters of the distribution. The lack of information regarding these parameters is a form of epistemic uncertainty. Hence in this example aleatoric and epistemic uncertainty are mingled. Assuming that samples are available, one could use the empirical mean and variance from data to approximate the true mean and variance, and thus reduce this uncertainty. However, in the common situation where sampling is necessarily limited, some epistemic uncertainty is inherent in the model due to practical limitations on data acquisition and modeling.

Another type of epistemic uncertainty is the omission of important aspects of the system model. This form is possibly the most difficult to quantify. For example, there might be a hidden random variable in the true physical system. In the PDE example we have assumed the boundary condition is modeled via a random variable, but there may be other coefficients in the true physical system that are random, but are modeled as constants. It is also frequently true that the mathematical model used for the system is only approximately true. For example, the model might simplify the geometry of the true system, nonlinearities

may be approximated by linear relations, etc. In the context of the PDE example, boundary conditions of Dirichlet form were chosen, when in fact a more realistic model might use a mixed form (e.g., Robin boundary conditions).

For almost all forms of epistemic uncertainty there is little justification for the use of random variables to model the uncertainty. Nonetheless, it is common practice to do just that, and in practice randomness and random variables are used in many situations to account for errors in the physical model or to model ignorance. The reasons were mentioned previously–that basic distributional properties of the output (e.g., variance) will still provide a reasonable sensitivity analysis for the problem, and existing computational methods are largely based on aleatoric uncertainty. However, with little justification for the use of any particular distribution (or even the use of randomness at all), in more critical applications one may insist on rigorous bounds on performance that are valid for a particular family of distributions. In the extreme case where no probabilistic model is considered acceptable, one may wish to only prescribe bounds on certain parameters, and then obtain tight bounds on performance over all values of the parameters that satisfy the bounds.

Thus there are many different types of uncertainty that one should account for in an analysis of the effects of uncertainty. These include: (i) aleatoric with known distribution; (ii) aleatoric with partly known distribution (mingled aleatoric and epistemic); (iii) epistemic for which one is willing to model by a family of aleatoric uncertainties, and (iv) epistemic where one is only willing to place bounds on the uncertainties. As remarked in the Introduction, this work will introduce an approach that allows these uncertainties to be handled within a single framework that can exploit computational methods originally developed just for the treatment of aleatoric uncertainties.

Chapter Three

Relative Entropy and Duality

3.1 Duality for Exponential Integrals

Our development of performance measures that distinguish forms of uncertainty, and which in particular allow for robustness with respect to epistemic uncertainties, depends on a duality relation between exponential integrals and relative entropy. Its first use along these lines but with regard to estimation appears to be in [1], and for optimization in [8]. We first state the basic duality result, and then define two functionals which will allow aleatoric and epistemic uncertainties to be analyzed simultaneously but at the same time distinguished.

The general duality is stated for random variables that take values in a Polish metric space (i.e., a complete, separable metric space) $\mathcal{X}$. The associated σ-algebra is the Borel σ-algebra. A typical example of $\mathcal{X}$ for our purposes is a closed subset of some Euclidean space $\mathbb{R}^d$. Let $\mathcal{P}(\mathcal{X})$ denote the collection of probability measures on $\mathcal{X}$, and let $\mu \in \mathcal{P}(\mathcal{X})$. Given any $\nu \in \mathcal{P}(\mathcal{X})$ that is absolutely continuous with respect to μ, we define the relative entropy of ν with respect to μ by

$$R(\nu \,\|\mu) \doteq \int_{\mathcal{X}} \log\left(\frac{d\nu}{d\mu}(x)\right) \nu(dx)$$

whenever $\log(d\nu/d\mu(x))$ is integrable with respect to ν. In all other cases $R(\nu \,\|\mu)$ is defined to be ∞.

Relative entropy defines a mapping $(\nu, \mu) \to R(\nu \,\|\mu)$ from $\mathcal{P}(\mathcal{X})^2$ to $\mathbb{R} \cup \infty$. This mapping has a number of very attractive properties such as non-negativity, joint convexity, etc. (see A.1.1 for more detail). It is important to note, however, that relative entropy is not a true metric since it is not symmetric, nor does it satisfy the triangle inequality. A property of particular interest for our purposes is the following variational formula for exponential integrals, which can be considered as an infinite-dimensional version of the Legendre transform [7, Proposition 1.4.2].

Proposition 3.1. *Let $(\mathcal{X}, \mathcal{A})$ be a measurable space where $\mathcal{X}$ is a Polish metric space and*

$\mathcal{A}$ is the associated σ-algebra. Given any bounded and continuous function $F : \mathcal{X} \to \mathbb{R}$, a probability measure $\mu \in \mathcal{P}(\mathcal{X})$, and any $c \in (0, \infty)$, we obtain the following results.

a) We have the variational formula

$$\Lambda_c \doteq \frac{1}{c} \log \int_{\mathcal{X}} e^{cF(x)} \mu(dx) = \sup_{\nu \in \mathcal{P}(\mathcal{X})} \left[-\frac{1}{c} R\left(\nu \,\|\mu\right) + \int_{\mathcal{X}} F(x)\nu(dx) \right]. \tag{3.1}$$

b) Let $\nu^* \in \mathcal{P}(\mathcal{X})$ be absolutely continuous with respect to μ and let it satisfy

$$\frac{d\nu^*}{d\mu}(x) \quad \doteq \quad \frac{e^{cF(x)}}{\int_{\mathcal{X}} e^{cF(x)} \mu(dx)}.$$

Then the infimum in the variational formula (3.1) is uniquely attained at ν^*.

Proof. It suffices to consider the case when ν is absolutely continuous with respect to μ since $R(\nu\|\mu) < \infty$ and $R(\nu\|\mu) = \infty$ otherwise. Since ν^* is absolutely continuous with respect to μ and F is bounded and continuous, μ is absolutely continuous with respect to ν^*. Therefore, ν is absolutely continuous with respect to ν^*. We have

$$\begin{aligned}
&-\frac{1}{c} R(\nu\|\mu) + \int_{\mathcal{X}} F(x)\nu(dx) \\
&= \quad -\frac{1}{c} \int_{\mathcal{X}} \left(\log \frac{d\nu}{d\mu} \right) d\nu + \int_{\mathcal{X}} F(x)\nu(dx) \\
&= \quad -\frac{1}{c} \int_{\mathcal{X}} \left(\log \frac{d\nu}{d\nu^*} \right) d\nu - \frac{1}{c} \int_{\mathcal{X}} \left(\log \frac{d\nu^*}{d\mu} \right) d\nu + c \int_{\mathcal{X}} F(x)\nu(dx) \\
&= \quad -R(\nu\|\nu^*) + \frac{1}{c} \log \int e^{cF} d\mu.
\end{aligned}$$

Since $R(\nu\|\nu^*) \geq 0$ and $R(\nu\|\nu^*) = 0$ if and only if $\nu = \nu^*$, the proof for the variational formula (a) is complete and the measure that attains the supremum is ν^*. □

The duality continues to hold as stated if F is bounded from above, and also when bounded from below if one restricts the supremum to $\nu \in \mathcal{P}(\mathcal{X})$ for which $R\left(\nu \,\|\mu\right) < \infty$ [7, Proposition 4.5.1].

As an elementary example on how such a formula can be useful, suppose that F is a performance measure (e.g., variance or an error probability), and that μ is a model for some random phenomena. We consider μ to be the *nominal model*, e.g., our best guess[1]. However, we are uncertain if this is the correct model, and would like a measure of performance that also holds for alternative models, but with a penalty for deviation from the nominal model. Λ_c is just such a performance measure, since by (A.3), for any alternative model ν the bound

$$\int_{\mathcal{X}} F(x)\nu(dx) \leq \Lambda_c + \frac{1}{c} R\left(\nu \,\|\mu\right) \tag{3.2}$$

applies. Hence we obtain bounds on performance over a family of alternative models. The parameter c allows one to balance robustness with respect to possible model inaccuracies against tighter bounds. In particular, as c tends to 0, Λ_c converges to $\int_{\mathcal{X}} F(x)\mu(dx)$, which is the performance measure under the nominal model, but in the limit the bound is meaningful only when $\nu = \mu$.

Suppose that a bound on performance over a specific family of distributions is needed. Let R^* denote the maximum of relative entropy with respect to the nominal model over this family. Then the tightest possible bound is obtained by minimizing

$$\Lambda_c + \frac{1}{c} R^*$$

over $c > 0$. We show in Proposition 3.3 below that this function has only one local minimum over $c \in (0, \infty]$, and thus the global minimum is easy to compute.

Functionals of the exponential form that appears in the definition of Λ_c are sometimes called *risk-sensitive* because the exponential amplifies the effect of any large values of the performance measure F. In the next two sections we consider two hybrid risk-sensitive functionals that will allow aleatoric and epistemic variables to be combined and yet distinguished.

[1]This is similar to the prior distribution in the Bayesian framework.

3.2 Two Hybrid Forms

In this section and the next random variables with a *known* distribution will take values in a Polish space $\mathcal{X}$. Variables whose distribution is *not known* or are otherwise of the epistemic variety take values in the Polish metric space $\mathcal{Y}$.

The performance measure of interest for some given problem is assumed to be of the form

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} F(x,y)\lambda(dy)\nu(dx), \tag{3.3}$$

where ν (resp., λ) is a probability measure on $\mathcal{X}$ (resp., $\mathcal{Y}$). If X and Y are independent random variables with distributions ν and λ, then $F(X,Y)$ represents both the performance measure (e.g., a second moment) as well as the underlying physical or mechanical system that maps these aleatoric and epistemic inputs into outputs. Now, let μ and γ represent the nominal distribution of X and Y, resp., while ν and λ represent the alternative probability distributions of X and Y, resp. Applying (A.3) to the joint measures $\mu \times \gamma$ (nominal) and $\nu \times \lambda$ (alternative), we obtain bounds on the performance measure

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} F(x,y)\lambda(dy)\nu(dx) \leq \frac{1}{c} R\left(\lambda \,\|\gamma\right) + \frac{1}{c} R\left(\nu \,\|\mu\right) + \Lambda_c \tag{3.4}$$

where the ordinary risk-sensitive performance measure is

$$\Lambda_c = \frac{1}{c} \log \int_{\mathcal{X}} \int_{\mathcal{Y}} e^{cF(x,y)} \gamma(dy)\mu(dx). \tag{3.5}$$

Since the distribution of X is known, we can take $\nu = \mu$ so that (3.4) now becomes

$$\int_{\mathcal{X}} \int_{\mathcal{Y}} F(x,y)\lambda(dy)\nu(dx) \leq \frac{1}{c} R\left(\lambda \,\|\gamma\right) + \Lambda_c. \tag{3.6}$$

While (3.6) seems to give upper bounds on the performance measure under uncertainty in the distribution of Y only, the measure (3.5) does not distinguish the variables according to type (aleatoric or epistemic). In fact, the risk-sensitive performance measure (3.5) is the

same in both (3.4) and (3.6). Therefore, the use of a risk-sensitive form (3.5) is not well motivated for the aleatoric variables. Indeed, use of this measure will give bounds that are robust with respect to variations on a distribution that is known (as can be seen in (3.4)), and obviously such bounds are not as tight as possible.

The first form we consider for a hybrid measure is

$$\Lambda_c^1 = \frac{1}{c} \log \int_{\mathcal{Y}} e^{\int_{\mathcal{X}} cF(x,y)\mu(dx)} \gamma(dy). \tag{3.7}$$

Note that by Jensen's inequality (applied to the convex function exp), this is smaller than (and in general strictly smaller than) Λ_c. Using the relative entropy representation for exponential integrals given in (A.3) but with $\mathcal{Y}$ in place of $\mathcal{X}$ and letting $\int_{\mathcal{X}} cF(x,y)\mu(dx)$ be the cost function], it follows that for any distribution $\theta(dy)$

$$\int_{\mathcal{Y}} \int_{\mathcal{X}} F(x,y)\mu(dx)\theta(dy) \leq \frac{1}{c} R\left(\theta(dy) \,\|\gamma(dy)\right) + \Lambda_c^1. \tag{3.8}$$

This gives a bound on the performance measure for an arbitrary distribution of Y, but with the distribution of X equal to the known true distribution. The distributions thus play very different roles. In particular, we think of γ as a *nominal* distribution of Y, which should be distinguished from a possible *true* distribution. The risk sensitive functional Λ_c^1, whose numerical evaluation can be carried out by a variety of methods (including polynomial chaos expansion as discussed below), is based on the nominal distribution. Through the relative entropy duality, it will yield various bounds (depending on c) on a families of distributions, with the relative entropy distance the key metric. Ideally one would consider the smallest family that contains the "true" distribution (if it exists), and get the tightest bound by optimizing over c. With the hybrid risk-sensitive formulation, μ is allowed to be both the nominal (computational) distribution and also the true distribution.

The second form one could consider for a hybrid measure is

$$\Lambda_c^2 = \frac{1}{c} \int_{\mathcal{X}} \left[\log \int_{\mathcal{Y}} e^{cF(x,y)} \gamma(dy) \right] \mu(dx). \tag{3.9}$$

Note that by Jensen's inequality (applied now to the concave function log), this is again in general strictly smaller than Λ_c. The relative entropy representation for exponential integrals gives

$$\int_{\mathcal{Y}} F(x,y)\theta(dy|x) \le \frac{1}{c} R\left(\theta(dy|x) \,\|\gamma(dy)\right) + \frac{1}{c}\log \int_{\mathcal{Y}} e^{cF(x,y)}\gamma(dy).$$

Here $\theta(dy|x)$ is any stochastic kernel on $\mathcal{Y}$ given $\mathcal{X}$, which is essentially a conditional distribution on $\mathcal{Y}$ given $X = x$ (for the precise definition see [7, page 35]). Integrating over $\mathcal{X}$ gives

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} F(x,y)\theta(dy|x)\mu(dx) \le \frac{1}{c}\int_{\mathcal{X}} R\left(\theta(dy|x) \,\|\gamma(dy)\right)\mu(dx) + \Lambda_c^2.$$

In the case where $\theta(dy|x)$ is independent of x, we obtain

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} F(x,y)\theta(dy)\mu(dx) \le \frac{1}{c} R\left(\theta(dy) \,\|\gamma(dy)\right) + \Lambda_c^2.$$

Note that the second hybrid cost gives a more general bound, in that it allows the alternative distribution on Y (i.e., $\theta(dy|x)$) to depend on the value taken by X. This additional flexibility comes at a cost, and indeed we will show in the next subsection that the following inequalities hold (typically in a strict fashion):

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} F(x,y)\gamma(dy)\mu(dx) \le \Lambda_c^1 \le \Lambda_c^2 \le \Lambda_c. \tag{3.10}$$

Hence if one is concerned with performance bounds for distributions on Y that do not depend on the variable X, then the tightest bound is obtained using Λ_c^1.

Before studying properties of the risk-sensitive measures, we note some elementary generalizations that are possible. Suppose that the nominal distribution of X and Y is not of product form. In the setting of the first form, we could let γ be the marginal distribution of Y and let $\mu(dx|y)$ be the conditional distribution of X given $Y = y$. Using the extended definition

$$\bar{\Lambda}_c^1 = \frac{1}{c}\log \int_{\mathcal{Y}} e^{\int_{\mathcal{X}} cF(x,y)\mu(dx|y)}\gamma(dy),$$

we obtain the bound

$$\int_{\mathcal{Y}}\int_{\mathcal{X}} F(x,y)\mu(dx|y)\theta(dy) \leq \frac{1}{c}R\left(\theta(dy)\,\|\gamma(dy)\right) + \bar{\Lambda}_c^1.$$

For the second form, we let μ be the marginal distribution of X and let $\gamma(dy|x)$ be the conditional distribution of Y given $X = x$. With the definition

$$\bar{\Lambda}_c^2 = \frac{1}{c}\int_{\mathcal{X}}\left[\log\int_{\mathcal{Y}} e^{cF(x,y)}\gamma(dy|x)\right]\mu(dx),$$

we obtain

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} F(x,y)\theta(dy|x)\mu(dx) \leq \frac{1}{c}\int_{\mathcal{X}} R\left(\theta(dy|x)\,\|\gamma(dy|x)\right)\mu(dx) + \bar{\Lambda}_c^2.$$

It is indeed sometimes useful to allow the distribution of one type of variable to depend on the value taken by the other type of variable. As an elementary example, consider the situation where an aleatoric variable appears in a system, and while it is known that this variable has a Gaussian distribution with mean zero, the variance of the variable is however not known. Then the variance itself might be modeled as an uncertainty whose distribution is not known precisely, i.e., as an epistemic uncertainty. In this case $\bar{\Lambda}_c^1$ would be the relevant risk-sensitive formulation. An example of this sort is presented in Subsection 4.2.2.

Remark 3.1. *In problems of regulatory assessment epistemic uncertainty can pose a unique set of challenges. The collection of data and model validation are often particularly difficult or expensive, while at the same time stringent bounds on performance might be demanded. In these circumstances, the framework presented here, wherein tight bounds are obtained for a known family of alternative models, would seem very natural. Prior to the collection of data or testing of samples to determine compliance, all critical elements of the model, including selection of the nominal model, size of the family of alternative models allowed by the relative entropy bounds, and the performance measures that would be required to hold uniformly for this family, would all be determined by negotiation among the interested parties. This would allow various competing needs (e.g., expense of model*

validation versus adequate protection of consumers) to be balanced according to the interests of the participants.

3.3 Properties of the Risk-Sensitive Forms

In this section we prove properties of the two hybrid risk-sensitive forms. In particular we derive an inequality relating the two, and study the limit as $c \to \infty$. For the results to hold as stated we often need F be bounded from below. This is a mild assumption. Indeed, standard measures of performance such as second moments and error probabilities satisfy this condition.

The first proposition shows the inequality between the two hybrid forms.

Proposition 3.2. *Assume that F is bounded from below and consider the functionals defined in (3.3), (3.5), (3.7) and (3.9). Then the inequalities (3.10) hold.*

Proof. The only inequality which does not follow from Jensen's inequality is $\Lambda_c^1 \leq \Lambda_c^2$. The relative entropy duality as stated in (A.3) applied to Λ_c^1 gives

$$\Lambda_c^1 = \sup_{\theta \in \mathcal{P}(\mathcal{Y})} \left[\int_{\mathcal{Y}} \int_{\mathcal{X}} F(x,y)\mu(dx)\theta(dy) - \frac{1}{c} R\left(\theta(dy) \,\|\gamma(dy)\right) \right],$$

where $\mathcal{P}(\mathcal{Y})$ is the space of probability measures on $\mathcal{Y}$. The same representation applied to

$$\frac{1}{c} \log \int_{\mathcal{Y}} e^{cF(x,y)} \gamma(dy)$$

gives

$$\frac{1}{c} \log \int_{\mathcal{Y}} e^{cF(x,y)} \gamma(dy) = \sup_{\theta \in \mathcal{P}(\mathcal{Y})} \left[\int_{\mathcal{Y}} F(x,y)\theta(dy) - \frac{1}{c} R\left(\theta(dy) \,\|\gamma(dy)\right) \right].$$

The supremum operation in the last display can be done in such a way that an optimizing (or near optimizing) θ is a Borel measurable function of x, i.e., a stochastic kernel $\theta(dy|x)$ on $\mathcal{Y}$ given $\mathcal{X}$ [7]. Let $\mathcal{P}(\mathcal{Y}|\mathcal{X})$ denote the space of such stochastic kernels. Integrating the

last display with respect to μ gives

$$\Lambda_c^2 = \sup_{\theta \in \mathcal{P}(\mathcal{Y}|\mathcal{X})} \left[\int_{\mathcal{X}} \int_{\mathcal{Y}} F(x,y)\theta(dy|x)\mu(dx) - \frac{1}{c} \int_{\mathcal{X}} R\left(\theta(dy|x) \,\|\gamma(dy)\right) \mu(dx) \right].$$

Since $\mathcal{P}(\mathcal{Y}) \subset \mathcal{P}(\mathcal{Y}|\mathcal{X})$, $\Lambda_c^2 \geq \Lambda_c^1$ follows. □

The next proposition shows how to obtain bounds on the performance over a family of distributions defined in terms of a maximum relative entropy B.

Proposition 3.3. *Consider the functionals defined in (3.5), (3.7) and (3.9), and let* $D = \{c : \Lambda_c < \infty\}$ *(resp.,* $D^i = \left\{c : \Lambda_c^i < \infty\right\}, i = 1, 2$*). Assume that the interior of* D *(resp.,* D^i*) is nonempty. Then* Λ_c *(resp.,* Λ_c^i*) is differentiable on the interior of* D *(resp.,* D^i*). Assume also that* F *is bounded from below by zero. Then* $c \mapsto \Lambda_c$ *(resp.,* $c \mapsto \Lambda_c^i$*) is nondecreasing for* $c \geq 0$*. Let* $B > 0$ *be given. Then there is a unique* $c \in (0, \infty]$ *at which*

$$c \mapsto \frac{1}{c}B + \Lambda_c \quad \left(resp., \; \frac{1}{c}B + \Lambda_c^i \right)$$

attains a local minimum, where the statement that the minimum occurs at $c = \infty$ *means that* $\Lambda_c + B/c > \Lambda_\infty$ *for a well defined limit* Λ_∞ *and all* $c < \infty$*.*

Proof. To simplify we give the proof only for the case of Λ_c and omit the y variable. Thus we consider $\Lambda_c = \frac{1}{c} \log \int_{\mathcal{X}} e^{cF(x)} \mu(dx)$. Proofs for the other functionals are analogous. It is well known that $H(c) = \log \int_{\mathcal{X}} e^{cF(x)} \mu(dx)$ is a convex function taking values in $\mathbb{R} \cup \{\infty\}$, and strictly convex and infinitely differentiable on the interior of $\{c : H(c) < \infty\}$ (see, e.g., [38]).

Next we prove the monotonicity. Differentiating gives

$$H'(c) = \frac{\int_{\mathcal{X}} F(x) e^{cF(x)} \mu(dx)}{\int_{\mathcal{X}} e^{cF(x)} \mu(dx)}.$$

Since $F \geq 0$ the derivative is non-negative, and convexity implies it is non-decreasing.

Using

$$\frac{1}{c}H(c) = \frac{1}{c}\left(H(c) - H(0)\right) = \frac{1}{c}\int_0^c H^{'}(s)ds,$$

it follows that $\Lambda_c = H(c)/c$ is nondecreasing for $c \geq 0$.

First assume that $M \triangleq^{\circ} \sup D = \infty$, and observe that

$$\frac{d}{dc}\left[\frac{1}{c}B + \frac{1}{c}H(c)\right] = \frac{1}{c^2}\left[-B + cH^{'}(c) - H(c)\right].$$

Since H is strictly convex and $H^{'}(0) \geq 0$, the mapping $f(c) \mapsto cH^{'}(c) - H(c)$ is monotone increasing and takes the value 0 at $c = 0$ (see Figure 3.1).

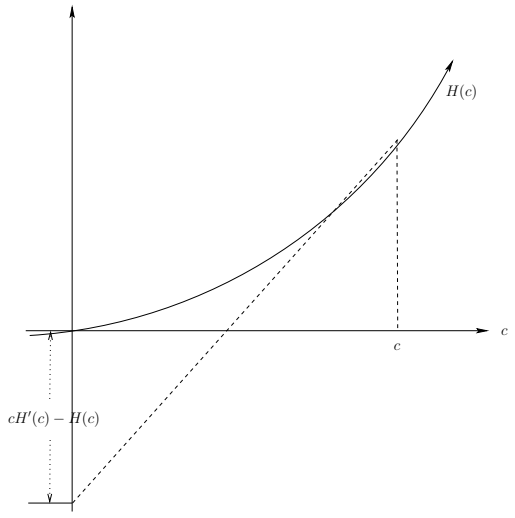

Figure 3.1: $f(c)$ is monotone increasing.

Let K denote the limit of $f(c)$ as $c \to \infty$. If $K = \infty$ then there is a unique solution to

$$cH^{'}(c) - H(c) = B,$$

and we are done. The same is true if $0 \leq B < K$. Hence the only case left is when $B \geq K$. In this case $(B + H(c))/c$ is monotone decreasing for all $c \in [0, \infty)$. Since $H(c)/c \geq H^{'}(0) \geq 0$, there is a well-defined limit Λ_∞ that is necessarily the minimum.

Next assume that $M \triangleq^{\circ} \sup D \in (0, \infty)$. We claim that in this case $cH^{'}(c) - H(c) \uparrow \infty$ as $c \uparrow M$, and therefore we can argue just as before. By monotone convergence it must be

that $H(c) \uparrow \infty$ as $c \uparrow M$, which implies that necessarily $H^{'}(c) \uparrow \infty$ as $c \uparrow M$. To prove the claim, let $0 < \beta < c < M$. Then since $H^{'}(s)$ is increasing

$$H(c) = \int_0^c H^{'}(s)ds \leq \beta H^{'}(\beta) + (c - \beta)H^{'}(c).$$

This implies

$$cH^{'}(c) - H(c) \geq \beta \left[H^{'}(c) - H^{'}(\beta) \right].$$

Letting $c \uparrow M$ and using that $\beta > 0$, the claim is proved, thus completing the proof of the theorem. □

Next, we consider the sense in which the infimum over c of right hand side of (A.3) provides a tight bound on the performance measure. To simplify we give the proof only for the case of Λ_c and omit the y variable. The proofs for the hybrid forms (e.g., (3.8)) are analogous.

Theorem 3.1. *Consider the functional defined in (A.3) and assume the conditions of Proposition 3.3. Suppose that $L \in \mathbb{R}$, $B \in [0, \infty)$ are given, and that $c^* < \infty$ minimizes*

$$c \mapsto \frac{1}{c} B_c + \Lambda_c$$

so that $\frac{1}{c^*} B + \Lambda_{c^*} < \infty$. Define $\mathcal{F}_B \triangleq \{\gamma \in \mathcal{P}(\mathcal{X}) : R(\gamma \| \mu) \leq B\}$, i.e., the family of alternative distributions on $\mathcal{X}$. Then for all $\nu \in \mathcal{F}_B$,

$$\int_{\mathcal{X}} F(x)\nu(dx) \leq L \Leftrightarrow \frac{1}{c^*} B + \Lambda_{c^*} \leq L.$$

Proof. If $\frac{1}{c^*} B + \Lambda_{c^*} \leq L$ then, by (A.3), $\int_{\mathcal{X}} F(x)\nu(dx) \leq L$ for all $\nu \in \mathcal{F}_B$. For the other direction, we will show that there exists a $\nu(dx) \in \mathcal{F}_B$ such that

$$\int_{\mathcal{X}} F(x)\nu(dx) = \frac{1}{c^*} B + \Lambda_{c^*}.$$

Hence, if $\frac{1}{c^*}B + \Lambda_{c^*} > L$, then, for at least one $\nu(dx) \in \mathcal{F}_B$, $\int_{\mathcal{X}} F(x)\nu(dx) > L$.

First, recall from (A.3) that for any given $c \in (0, \infty)$

$$\Lambda_c = \sup_{\nu \in \mathcal{P}(\mathcal{X})} \left[-\frac{1}{c} R(\nu \,\|\mu) + \int_{\mathcal{X}} F(x)\nu(dx) \right].$$

The supremum is achieved at $\nu^c(dx) = e^{cF(x)}\mu(dx)/Z(c)$, where $Z(c) = \int_{\mathcal{X}} e^{cF(x)}\mu(dx)$ (Proposition 3.1). Thus

$$\int_{\mathcal{X}} F(x)\nu^{c^*}(dx) = \frac{1}{c^*} R(\nu^{c^*} \,\|\mu) + \Lambda_{c^*}.$$

By the previous proposition, $c^* < \infty$ must lie in the interior of D. Therefore, we can differentiate $\frac{1}{c}B + \Lambda_c$ with respect to c, and we get that c^* satisfies

$$-\frac{1}{(c^*)^2} \log \int_{\mathcal{X}} e^{c^* F(x)}\mu(dx) + \frac{1}{c^*} \frac{1}{Z(c^*)} \int_{\mathcal{X}} F(x) e^{cF(x)}\mu(dx) - \frac{B}{(c^*)^2} = 0. \tag{3.11}$$

Also,

$$\begin{aligned} R(\nu^{c^*} \,\|\mu) &= \int_{\mathcal{X}} \frac{[c^* F(x) - \log Z(c^*)]}{Z(c^*)} e^{c^* F} \mu(dx) \\ &= \frac{1}{Z(c^*)} \int_{\mathcal{X}} c^* F(x) e^{c^* F(x)} \mu(dx) - \log Z(c^*), \end{aligned}$$

and so

$$\frac{1}{(c^*)^2} R(\nu^{c^*} \,\|\mu) = \frac{1}{Z(c^*)} \int c^* F e^{c^* F} d\mu - \frac{1}{(c^*)^2} \log Z(c^*). \tag{3.12}$$

Comparing (3.12) and (3.12) we have $R(\nu^{c^*} \,\|\mu) = B$ and therefore $\nu^{c^*} \in \mathcal{F}_B$. □

In the context of regulatory assessment, it may be required that the performance measure of interest be no greater than a particular value, e.g., L. We say that the performance measure will pass if it does not exceed L, i.e., $\int_{\mathcal{X}} F(x)\nu(dx) \leq L$, and fails if it exceeds L, i.e., $\int_{\mathcal{X}} F(x)\nu(dx) > L$. This theorem tells us is that the performance measure will pass the regulatory criteria for a given family of alternative distributions if and only if the

upper bound, $\frac{1}{c^*}B + \Lambda_{c^*}$, does not exceed L, i.e., $\frac{1}{c^*}B + \Lambda_{c^*} \leq L$. If $\frac{1}{c^*}B + \Lambda_{c^*}$ exceeds L, all we can say is that there exists at least one distribution in the given family of alternative distributions for which the performance measure fails. In this case, it may very well be that there are distributions within the family that actually pass the regulatory criteria, but, as a whole, the family fails to satisfy the requirement.

One of the situations of interest when aleatoric and epistemic variables appear simultaneously is the case where all that is known regarding the epistemic variables are bounds. It turns out that this problem is well posed (i.e., the relevant performance criteria are finite) essentially in those cases where $\Lambda^1_\infty < \infty$. In fact, when this is the case the value of Λ^1_∞ can be used to establish optimal bounds subject to the constraint on the epistemic variables. For simplicity, we assume that the set A appearing in the statement of the following theorem is bounded and that γ, the nominal distribution for the epistemic variables, is the uniform distribution (in the limit $c \to \infty$ the precise form of the nominal distribution is not important, and in fact it is only the support of the distribution that matters in the limit (see the remark after Theorem 3.2)). If A is not bounded one can use an appropriate distribution whose support is all of A (e.g., the exponential distribution could be used for $A = [0, \infty)$. The choice of γ as uniform when A is bounded is simplest and also one that is convenient for most uses. A related statement holds for Λ^2_∞, but it is not as useful (at least in this setting), since $\Lambda^1_\infty \leq \Lambda^2_\infty$.

Theorem 3.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be subsets of a finite dimensional Euclidean space. Suppose that $A \subset \mathcal{Y}$ is bounded and the closure of its interior and that γ is the uniform distribution on A. Assume that F is lower semicontinuous in y for each $x \in \mathcal{X}$ and bounded from below. Define the risk-sensitive functional Λ^1_c by (3.7) and let $\Lambda^1_\infty = \lim_{c\to\infty} \Lambda^1_c$. Then*

$$\sup_{y\in A} \int_{\mathcal{X}} F(x, y)\mu(dx) = \Lambda^1_\infty.$$

Proof. Since $c\Lambda^1_c$ is convex the limit of Λ^1_c is well defined, though it might take the value ∞. We give the proof for the case $\Lambda^1_\infty < \infty$. The extension to the case $\Lambda^1_\infty = \infty$ is

straightforward.

We first prove that the left side of the last display is bounded above by the right side. Fix any $\bar{y} \in A^\circ$, the interior of A. For small $\varepsilon > 0$ let B_ε be the ball in A about $\bar{y}$ of volume ε, and let M be the volume of A. Let $\theta(dy)$ be the measure whose density with respect to Lebesgue measure is $1/\varepsilon$ on B_ε, and zero elsewhere. Then

$$\begin{aligned} R\left(\theta(dy)\,\|\gamma(dy)\right) &= \int_{B_\varepsilon} \log\left[(1/\varepsilon)/(1/M)\right]\theta(dy) \\ &= (1/\varepsilon)\log\left[(1/\varepsilon)/(1/M)\right]. \end{aligned}$$

Then (3.8) gives

$$\int_{\mathcal{X}}\int_{B_\varepsilon} F(x,y)\theta(dy)\mu(dx) \le \frac{1}{c}\,(1/\varepsilon)\log\left[(1/\varepsilon)/(1/M)\right] + \Lambda_c^1. \tag{3.13}$$

Since $y \to F(x,y)$ is lower semicontinuous

$$\liminf_{\varepsilon\to 0}\int_{B_\varepsilon} F(x,y)\theta(dy) \ge F(x,\bar{y}).$$

Letting first $c \to \infty$ and then $\varepsilon \to 0$ in (3.13), Fatou's lemma gives $\displaystyle\int_{\mathcal{X}} F(x,\bar{y})\mu(dx) \le \Lambda_\infty^1$. Since $\bar{y} \in A^\circ$ is arbitrary, the lower semicontinuity and another use of Fatou give the bound for all $y \in A$.

For the reverse inequality, we use the fact that the minimizing measure in the definition of Λ_c^1 exists. In fact,

$$\int_{\mathcal{Y}}\int_{\mathcal{X}} F(x,y)\mu(dx)\theta_c^*(dy) = \frac{1}{c}R\left(\theta_c^*(dy)\,\|\gamma(dy)\right) + \Lambda_c^1$$

precisely at θ^* defined by

$$\frac{d\theta_c^*(\cdot)}{d\gamma(\cdot)}(y) = e^{\int_{\mathcal{X}} cF(x,y)\mu(dx)} \Big/ \int_A e^{\int_{\mathcal{X}} cF(x,y)\mu(dx)}\gamma(dy)$$

(see [7, Proposition 1.4.2]). By the non-negativity of relative entropy

$$\begin{aligned}\Lambda^1_\infty &= \lim_{c\to\infty} \Lambda^1_c \\ &\leq \limsup_{c\to\infty} \int_{\mathcal{Y}} \int_{\mathcal{X}} F(x,y)\mu(dx)\theta^*_c(dy) \\ &\leq \sup_{y\in A} \int_{\mathcal{X}} F(x,y)\mu(dx).\end{aligned}$$

□

Remark 3.2. *The set A need not be bounded, nor must the uniform distribution be used. Indeed, assume only that $\gamma(dy)$ has a density $f(y)$ such that $f(y) \geq \alpha > 0$ for $y \in B_\varepsilon$. Then*

$$\begin{aligned}R\left(\theta(dy)\,\|\gamma(dy)\right) &\leq \int_{B_\varepsilon} \log\left[(1/\varepsilon)/\alpha\right]\theta(dy) \\ &= (1/\varepsilon)\log\left[1/(\varepsilon\alpha)\right],\end{aligned}$$

and the result still holds.

Chapter Four

Examples and Computational Methods

To indicate how the robust performance bounds might be used in practice, we present some numerical examples, which illustrate the relationship between the different risk sensitive integrals and the relevant techniques and computational issues. We first review the polynomial chaos techniques that are used to compute the risk-sensitive integrals. The reader familiar with this material can skip to the next section.

Remark 4.1. *Although previously the random variables with known and unknown distributions were denoted by X and Y, in the rest of the paper these random variables will be denoted by Z_1 and Z_2. This is done to free up x and y to be spatial variables in various PDE and related equations that might be used to define the mapping F.*

4.1 Review of Polynomial Chaos Methods

This subsection reviews the polynomial chaos method that we use to compute the risk-sensitive integrals. The reader familiar with this material can skip to Section 4.2, or refer to [24] for a more comprehensive treatment. It should also be noted that one need not use polynomial chaos methods for the calculation of the risk-sensitive integrals. Various Monte Carlo type techniques can certainly be applied. However, for the examples in this work, we found that polynomial chaos methods worked better (higher accuracy with a smaller number of sample points) than straightforward Monte Carlo integration.

To calculate risk sensitive integrals, one could use Monte Carlo integration, which requires evaluating $F(Z_1, Z_2)$ for many replicas of (Z_1, Z_2). Because the order of convergence for Monte Carlo integration is $\mathcal{O}(N^{-1/2})$, where N is the number of sample points, this method can take a long time to converge (though not necessarily for the given examples), and so we utilize generalized polynomial chaos (gPC) methods to approximate $F(z_1, z_2)$ and evaluate various statistical properties of $F(Z_1, Z_2)$ (see, for example, [44]). We briefly recall the mathematical framework of polynomial chaos methods, along with a simple example in one dimension.

Let $(\Omega, \mathcal{A}, \mathcal{P})$ be a probability space where Ω is the sample space, $\mathcal{A}$ the associated σ-algebra, and $\mathcal{P}$ the probability measure on $\mathcal{A}$. Define a random vector $\mathbf{Z}(\omega) \stackrel{\circ}{=} (Z_1(\omega), \ldots, Z_N(\omega)) \in \mathbb{R}^N$ on this probability space. Consider a d-dimensional bounded domain $D \subset \mathbb{R}^d$ with boundary ∂D and let $t \in [0, T], T \in (0, \infty)$. We consider random fields $u(t, x; \mathbf{Z}(\omega)) : \bar{D} \times [0, T] \times \Omega \to \mathbb{R}$ that are defined by requiring that for a.e. $\omega \in \Omega$

$$\mathcal{L}(t, x, u; \mathbf{Z}(\omega)) = f(t, x; \mathbf{Z}(\omega)), \qquad (x, t, \omega) \in D \times [0, T] \times \Omega,$$

subject to the boundary and initial conditions

$$\begin{aligned} \mathcal{B}(t, x, u; \mathbf{Z}(\omega)) &= g(t, x; \mathbf{Z}(\omega)), \qquad (x, t, \omega) \in \partial D \times [0, T] \times \Omega, \\ u(t, x; \mathbf{Z}(\omega)) &= u_0(x; \mathbf{Z}(\omega)), \qquad (x, t, \omega) \in D \times \{0\} \times \Omega. \end{aligned}$$

Here $x = (x_1, ..., x_d) \in \mathbb{R}^d, \mathcal{L}$ is a linear or nonlinear differential operator, and $\mathcal{B}$ is a boundary operator. We assume for simplicity that a unique classical sense solution exists to the differential equation and/or boundary conditions. Note that for the first two examples of the last subsection the problem involves only time dependence, and hence there is neither a spatial variable nor a boundary condition.

We assume that $\{Z_i\}_{i=1}^N$ are independent random variables, with support $\{\Gamma_i\}_{i=1}^N$ and with probability density functions $\{\rho_i(z_i)\}_{i=1}^N$, respectively. Hence the joint density is $\rho(\mathbf{z}) = \prod_{i=1}^N \rho_i(z_i)$. Let Ω be the canonical space $\prod_{i=1}^N \Gamma_i$, in which case $\mathbf{Z}(\omega) = \omega$ and we identify ω with $\mathbf{z}$. Thus the equations above become

$$\mathcal{L}(t, x, u; \mathbf{z}) = f(t, x; \mathbf{z}), \qquad (x, t, \mathbf{z}) \in D \times [0, T] \times \Omega, \tag{4.1}$$

subject to the boundary and initial conditions

$$\mathcal{B}(t, x, u; \mathbf{z}) = g(t, x; \mathbf{z}), \qquad (x, t, \mathbf{z}) \in \partial D \times [0, T] \times \Omega, \tag{4.2}$$

$$u(t, x; \mathbf{z}) = u_0(x; \mathbf{z}), \qquad (x, t, \mathbf{z}) \in D \times \{0\} \times \Omega. \tag{4.3}$$

We consider approximating the mapping as a function of the stochastic variable, i.e., $\mathbf{z} \to u(t, x; \mathbf{z})$ at some particular (t, x), via a finite sum of orthogonal basis functions.

We first define finite dimensional subspaces for $L^2(\Gamma_i)$ according to

$$W_i^{P_i} = \left\{ v : \Gamma_i \to \mathbb{R} : v \in \text{span}\left\{\phi_{i,m}(z_i)\right\}_{m=0}^{P_i} \right\}, \qquad i = 1, ..., N.$$

Here P_i is the highest degree of the polynomial basis function, and $\{\phi_{i,m}(z_i)\}$ is a set of orthonormal polynomials with respect to the weight ρ_i, i.e., for $m \neq n$

$$\int_{\Gamma_i} \phi_{i,m}(z_i)\phi_{i,n}(z_i)\rho_i(z_i)dz_i = 0$$

and

$$\int_{\Gamma_i} \phi_{i,n}^2(z_i)\rho_i(z_i)dz_i = 1.$$

The orthonormal basis representation is determined by the probability density function ρ_i. For example, if the density is uniform or Gaussian, then Legendre or Hermite orthogonal polynomials, respectively, are used. A table of polynomial basis functions and their respective distributions can be found in the appendix in Section A.2 (see also [41]). A finite dimensional subspace for $L^2(\Gamma)$, where $\Gamma \triangleq \Gamma_1 \times \cdots \times \Gamma_N = \Omega$, can either be defined as

$$W_N^P = \bigotimes_{|\mathbf{P}| \leq P} W_i^{P_i}$$

where the tensor product is over all combinations of the multi-index $\mathbf{P} = (P_1, \ldots, P_N) \in \mathbb{N}_0^N$ with $|\mathbf{P}| = \sum_{i=1}^{N} P_i \leq P$, or

$$\tilde{W}_N^P = \bigotimes_{i=1}^{N} W_i^P.$$

Thus, W_N^P is the space of N-variate orthogonal polynomials of total degree at most P, whereas $\tilde{W}_N^P$ is the full tensor product of the one-dimensional polynomial spaces with each highest degree P. Note that $\dim(W_N^P) = \binom{N+P}{P}$ and $\dim(\tilde{W}_N^P) = (P+1)^N$, and that for large N, $\binom{N+P}{P} \ll (P+1)^N$. Since our examples only consider $N = 2$, we will use

the full tensor product space, $\tilde{W}_N^P$. Let $\Phi_j(\mathbf{z}), j = 1, \ldots, M$ denote the elements of $\tilde{W}_N^P$, where $M = \dim(\tilde{W}_N^P)$.

There are two standard methods for constructing gPC approximations, referred to as the stochastic Galerkin method [41] and the stochastic collocation method [42]. In the stochastic Galerkin method, we write the gPC approximation of the solution $u(t, x; z)$ as

$$v(t, x; \mathbf{z}) = \sum_{j=1}^{N_p} \hat{u}_j(t, x; \mathbf{z}_k)\mathbf{\Phi}_j(\mathbf{z}), \tag{4.4}$$

i.e., the projection of the solution onto the space $\tilde{W}_N^P$ (one can also project onto the space W_N^P), where the generalized Fourier coefficients are defined as

$$\hat{u}_j(t, z; \boldsymbol{z}_k) \quad = \quad \int_\Gamma v(t, x; \mathbf{z})\mathbf{\Phi}_j(\mathbf{z})\rho(\mathbf{z})d\mathbf{z}, j = 1, \ldots, N_p. \tag{4.5}$$

Classical approximation theory guarantees that this is the best approximation in the linear polynomial space of N-variate polynomials of degree at most P [41, Section 3.3]. In order to determine the coefficients $\hat{u}_j$ in (4.4), we require that the residuals

$$\begin{aligned} R(t, x; \mathbf{z}) &\doteq \mathcal{L}(t, x, v; \mathbf{z}) - f(t, x; \mathbf{z}) \\ R^{BC}(t, x; \mathbf{z}) &\doteq \mathcal{B}(t, x, v; \mathbf{z}) - g(t, x; \mathbf{z}) \\ R^{IC}(x; \mathbf{z}) &\doteq v(0, x; \mathbf{z}) - u_0(x; \mathbf{z}) \end{aligned} \tag{4.6}$$

be orthogonal to $\tilde{W}_N^P$, i.e.,

$$\begin{aligned} \int_\Gamma R(t, x; \mathbf{z})\mathbf{\Phi}_j(\mathbf{z})\rho(\mathbf{z})d\mathbf{z} &= 0, \text{for } j = 1, \ldots, N_p \\ \int_\Gamma R^{BC}\mathbf{\Phi}_j(\mathbf{z})\rho(\mathbf{z})d\mathbf{z} &= 0, \text{for } j = 1, \ldots, N_p \\ \int_\Gamma R^{IC}(x; \mathbf{z})\mathbf{\Phi}_j(\mathbf{z})\rho(\mathbf{z})d\mathbf{z} &= 0, \text{for } j = 1, \ldots, N_p. \end{aligned} \tag{4.7}$$

(4.7) now represents a set of coupled deterministic equations for $\hat{u}_j$ and standard numerical techniques can be applied (see the Example in [41, Section 4.2.1]). The stochastic Galerkin

approach is referred to as an *intrusive method*, because it requires a reformulation of the differential equations as a coupled system. The stochastic collocation method, however, does not suffer from this same inconvenience and it is for this reason we choose to perform our numerical experiments with this method.[1]

With stochastic collocation, we first consider an approximation to the solution $u(t, x; \mathbf{z})$ in terms of a Lagrange interpolant of the form

$$v(t, x; \mathbf{z}) = \sum_{k=1}^{N_p} v(t, x; \mathbf{z}_k) \mathbf{L}_k(\mathbf{z}), \tag{4.8}$$

where $\mathbf{L}_k(\mathbf{z})$ is a Lagrange polynomial of degree $\dim(W_N^P)$, satisfies $\mathbf{L}_k(\mathbf{z}_l) = \delta_{kl}$, and $\{\mathbf{z}_k\}_{k=1}^{N_p}$ are a set of prescribed nodes in the N-dimensional space Γ. We require that the residuals (4.6) vanish at the collocation points $\{\mathbf{z}_k\}_{k=1}^{N_p}$, i.e.,

$$\begin{aligned}
\mathcal{L}(t, x, v(t, x; \mathbf{z}_k); \mathbf{z}_k) - f(t, x; \mathbf{z}_k) &= 0, \text{for } k = 1, \ldots, N_p \\
\mathcal{B}(t, x, v(t, x; \mathbf{z}_k); \mathbf{z}_k) - g(t, x; \mathbf{z}_k) &= 0, \text{for } k = 1, \ldots, N_p \\
v(0, x; \mathbf{z}_k) - u_0(x; \mathbf{z}_k) &= 0, \text{for } k = 1, \ldots, N_p.
\end{aligned}$$

Thus, the stochastic collocation method produces a set of N_p decoupled equations, where each value of $v(t, x; \mathbf{z}_k)$ coincides with the exact solution $u(t, x; \mathbf{z})$ for the given $\mathbf{z}_k$, since both satisfy the same equations.

Once the values of $\{v(t, x; \mathbf{z}_k)\}_{k=1}^{N_p}$ have been determined at the collocation points, one can construct a gPC approximation in the orthogonal basis representation in the form

$$v(t, x; \mathbf{z}) = \sum_{j=1}^{M} \hat{v}_j(t, x) \Phi_j(\mathbf{z}), \tag{4.9}$$

which, under certain conditions (e.g., using Gauss quadrature points), will be equivalent to the Lagrange basis formulation. The coefficients $\{\hat{v}_j\}_{j=1}^{M}$ can be determined by inverting a

[1] In short, the stochastic Galerkin method determines $\hat{u}_j$ by solving a set of coupled differential equations (4.7), while the stochastic collocation approach determines $\hat{u}_j$ via an interpolatory quadrature (4.8).

Vandermonde matrix, where invertibility is dependent on the choice of collocation points. If one chooses quadrature or cubature[2] points for the collocation points, then invertibility is guaranteed. Furthermore, if the cubature rule is exact up to polynomials of degree $2M$, inversion of the Vandermonde matrix is not necessary[3], and the coefficients $\{\hat{v}_j\}_{j=1}^{M}$ can be computed by evaluating

$$\hat{v}_j(t,x) = \sum_{k=1}^{N_p} v(t,x;\mathbf{z}_k)\Phi_j(\mathbf{z}_k)w_k, \tag{4.10}$$

where $\{w_k\}_{k=1}^{N_p}$ are the respective weights according to the choice of quadrature or cubature points.

Statistical information is easy to obtain from the form (4.9). For example,

$$\mathbb{E}_{\mathbf{z}}[v(t,x;\mathbf{z})] = \int_{\Gamma} \left(\sum_{j=1}^{M} \hat{v}_j(t,x)\Phi_j(\mathbf{z}) \right) \rho(\mathbf{z})d\mathbf{z} = \hat{v}_1(t,x)$$

and

$$\mathbb{E}_{\mathbf{z}}[(v(t,x;\mathbf{z}))^2] = \int_{\Gamma} \left(\sum_{j=1}^{M} \hat{v}_j(t,x)\Phi_j(\mathbf{z}) \right)^2 \rho(\mathbf{z})d\mathbf{z} = \sum_{j=1}^{M} \hat{v}_j(t,x)^2$$

as a consequence of the orthonormality. Other statistical information, such as higher moments and sensitivity coefficients, are also easily calculated (see [41]). The ease with which one can calculate such moments explains why orthogonal representation (4.9) is preferred over the Lagrange form (4.8).

To illustrate the collocation approach, we look at a simplified version of Example 4.1. Let $g(z_2) = u_0$, so that the only uncertainty is in the decay rate k, which depends on the random variable Z. The simplified problem becomes

$$\frac{d}{dt}u(t) = -k(z)u(t), \qquad u(0) = u_0,$$

[2]Cubature points just refers to quadratures in more than one dimension.

[3]This is especially useful if the Vandermonde matrix is ill-conditioned.

where $u(t;z)$ denotes the solution parameterized by z, with z a point in the range of Z. Requiring that the residuals vanish at the collocation points $\{z_n\}_{n=1}^{N_p}$ gives

$$R(t;z_n) = \frac{d}{dt}v(t;z_n) + k(z_n)v(t;z_n) = 0$$

and

$$R^{IC}(z_n) = v(0;z_n) - u_0 = 0.$$

Note that there are indeed N_p decoupled equations, which are solved independently for each n. Here the solutions are exact:

$$v(t;z_n) = u_0 e^{-k(z_n)t}, for = 1, \ldots, N_p.$$

One can then obtain the approximation (4.9) with respect to the orthogonal basis, assuming that the appropriate collocation points and weights have been chosen.

We illustrate this example with $k(z) = z$ and $Z \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$. Here the $\{\Phi_j\}_{j=1}^{M}$ are the Hermite polynomials, which are orthonormal with respect to the weighting function $\rho(z) = e^{-(z-\mu)^2/2\sigma^2}/\sigma\sqrt{2\pi}$ with support $\Gamma = (-\infty, +\infty)$. Figure 4.1 shows the stochastic collocation approximation $v(1,z)$ at $t = 1$ as a function of z, with $N_p = 2, 3, 4, 5$, versus the exact solution $u_{exact}(1,z) = e^{-z}$. Note that the collocation approximation is an interpolation of the exact solution at N_p points.

Figure 4.2 shows the relative error between the exact and approximated mean and variance in a semilog plot. The linear decrease in relative error on the semilog plot implies an exponential decay in the error. Errors become a constant at the limits of the accuracy of the numerical scheme.

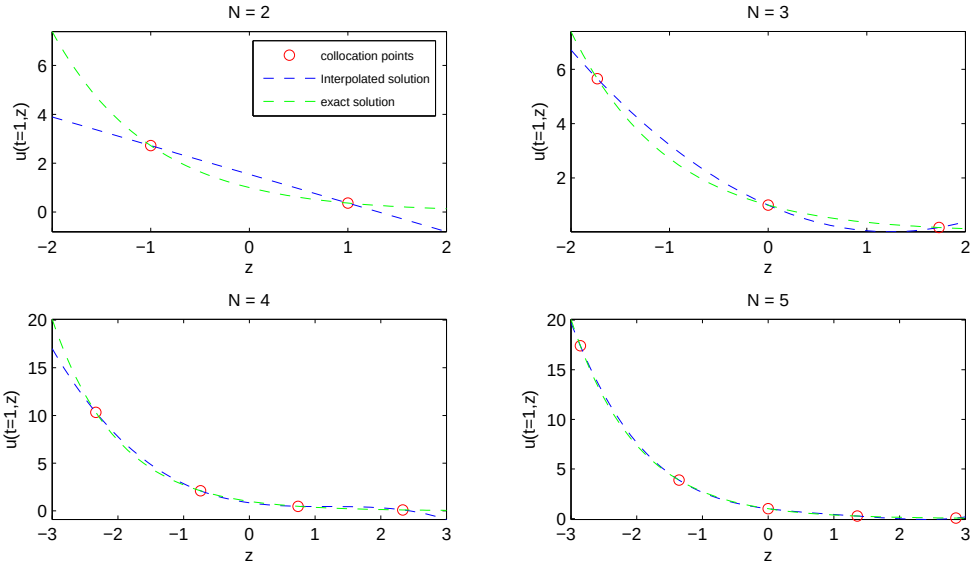

Figure 4.1: Interpolated solutions versus the exact solution at $t = 1$.

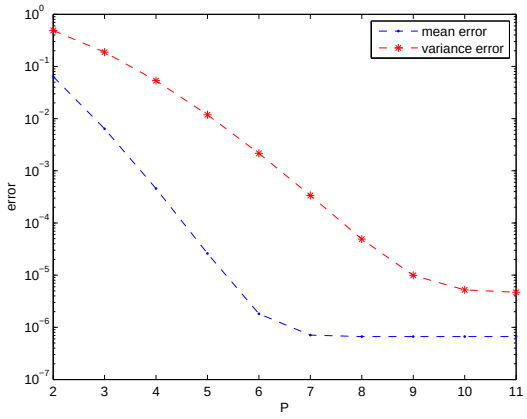

Figure 4.2: Relative errors.

4.2 Numerical Examples and Interpretation

We now discuss numerical calculation of the risk-sensitive integrals. For convenience, we recall the three types of risk-sensitive integrals introduced in Section 3.2, the ordinary risk

sensitive cost Λ_c and the two hybrid integrals Λ_c^1 and Λ_c^2:

$$\Lambda_c = \frac{1}{c}\log\int_{\mathcal{Z}_1}\int_{\mathcal{Z}_2} e^{cF(z_1,z_2)}\gamma(dz_2)\mu(dz_1), \tag{4.11}$$

$$\Lambda_c^1 = \frac{1}{c}\log\int_{\mathcal{Z}_2} e^{\int_{\mathcal{Z}_1} cF(z_1,z_2)\mu(dz_1)}\gamma(dz_2), \tag{4.12}$$

$$\Lambda_c^2 = \frac{1}{c}\int_{\mathcal{Z}_1}\left[\log\int_{\mathcal{Z}_2} e^{cF(z_1,z_2)}\gamma(dz_2)\right]\mu(dz_1), \tag{4.13}$$

where $\mathcal{Z}_1, \mathcal{Z}_2$ are the range spaces for the random variables Z_1, Z_2, respectively. Recall the relationships $\Lambda_c^1 \leq \Lambda_c^2 \leq \Lambda_c$. It is assumed that the distribution of Z_1 is known, but this is not true for Z_2. Instead, we assume there is a nominal distribution, or best guess, for Z_2, i.e, $\gamma(dz_2)$.

In the stochastic collocation approach, we replace $F(z_1, z_2)$ in (4.11) with a polynomial interpolant, $\tilde{F}(z_1, z_2)$, determined by the full tensor product of the one dimensional collocation points, $\{z_{1,i}\}_{i=1}^{N_{1,q}} \otimes \{z_{2,i}\}_{i=1}^{N_{2,q}}$ in $\mathcal{Z}_1 \times \mathcal{Z}_2$. Thus

$$\tilde{F}(z_1, z_2) = \sum_{j=1}^{M} \hat{F}_j(t, x)\Phi_j(z_1, z_2),$$

where $M = N_{1,q} \cdot N_{2,q}$, Φ_j is the full tensor product basis, and $\hat{F}_j(t, x)$ is computed using (4.10). From here, one can efficiently compute the risk sensitive integrals (3.5) via either Monte Carlo sampling or quadrature. For both methods, once the samples or quadrature points have been chosen, one can calculate the risk sensitive integrals for different values of c without resampling or choosing different quadrature points. Hence, it is computationally inexpensive to compute the risk sensitive integral for different values of c. Note that for higher dimensions, a sparse grid is preferred and, in most cases, necessary for these types of computations. It is typical to use the Smolyak algorithm to generate these sparse grids, which are based on a one dimensional quadrature rule (for example, see Section 4.1.2 in [40] for a more detailed explanation and further references).

In this work, we opt for the numerical quadrature approach. Since $\tilde{F}(z_1, z_2)$ can be evaluated very quickly, we can compute the risk sensitive integrals with high accuracy and

efficiency using a high order quadrature rule. The number of quadrature points chosen to compute the risk sensitive integrals need not be the same as the number of collocation points used in computing the coefficients of the gPC expansion, though the type of quadrature points is the same. For example, if we were to use Legendre-Gauss quadrature points to evaluate the gPC coefficients, we would again use Legendre-Gauss quadrature points to calculate the risk sensitive integrals. In the examples presented below a much larger number of quadrature points are used to compute the risk sensitive integrals. To avoid confusion, to the set of quadrature points used to evaluate the risk sensitive integrals is denoted by $\{\tilde{z}_{1,i}\}_{i=1}^{\tilde{N}_{1,q}} \otimes \{\tilde{z}_{2,i}\}_{i=1}^{\tilde{N}_{2,q}}$, i.e., the full tensor product of one dimensional quadrature points. In the examples, $\tilde{N}_{1,q}$ and$\tilde{N}_{2,q}$ equal 2^8, while $N_{1,q}$ and $N_{2,q}$ are 8 for smooth F and 12 when F is an indicator functions.

We also implemented the Monte Carlo approach, but observed that numerical quadrature gave a much better approximation for the same computational effort, at least for our examples. It should also be noted that a variation on traditional Monte Carlo is needed for the hybrid forms of the risk sensitive integrals. For the original risk-sensitive cost function (3.5), the standard approach is to use many independent samples of the pair (Z_1, Z_2). This produces an estimate which has a mean square error of $\mathcal{O}(N^{-1/2})$, where N is the total number of Monte Carlo samples. However, for the first hybrid form (3.7), one must approximate the inner integral $\int_{\mathcal{X}} cF(z_1, z_2)\mu(dz_1)$ for every fixed sample of Z_2. Thus one has to simulate more Z_1 samples than Z_2 samples. Similarly, to compute the second form of the hybrid (3.9), one must approximate the inner integral $\int_{\mathcal{Y}} e^{cF(z_1,z_2)}\gamma(dz_2)$ for every fixed Z_1. In particular, for both hybrid forms one does not simply choose independent samples of (Z_1, Z_2).

4.2.1 Independent Aleatoric and Epistemic Uncertainties

When the aleatoric and nominal epistemic variables are independent, the numerical quadrature approach would evaluate the risk sensitive integrals via the formulas

$$\Lambda_c \approx \frac{1}{c}\log\left(\sum_{i=1}^{\tilde{N}_{1,q}}\sum_{j=1}^{\tilde{N}_{2,q}} e^{cF(\tilde{z}_{1,i},\tilde{z}_{2,j})}\tilde{w}_{1,i}\tilde{w}_{2,j}\right) \tag{4.14}$$

$$\Lambda_c^1 \approx \frac{1}{c}\log\left(\sum_{j=1}^{\tilde{N}_{2,q}} e^{\left(\sum_{i=1}^{\tilde{N}_{1,q}} cF(\tilde{z}_{1,i},\tilde{z}_{2,j})\tilde{w}_{1,i}\right)}\tilde{w}_{2,j}\right) \tag{4.15}$$

$$\Lambda_c^2 \approx \frac{1}{c}\sum_{i=1}^{\tilde{N}_{1,q}}\left(\log\sum_{j=1}^{\tilde{N}_{2,q}} e^{cF(\tilde{z}_{1,i},\tilde{z}_{2,j})}\tilde{w}_{2,j}\right)\tilde{w}_{1,i}. \tag{4.16}$$

where $\{\tilde{z}_{1,i},\tilde{w}_{1,i}\}_{i=1}^{\tilde{N}_{1,q}}, \{\tilde{z}_{2,i},\tilde{w}_{2,i}\}_{i=1}^{\tilde{N}_{2,q}}$ are the pairs of quadrature points and weights chosen based on the underlying distribution.[4]

The first example is taken from [43] and involves a simple ODE. The example is convenient because analytic solutions to the risk-sensitive integrals can be obtained, and then compared with the numerical approximations described in the next section. The second example is a random oscillator, where the stochastic parameters are the spring and dampening coefficients. The third example involves a heat equation with random heat capacity and thermal conductivity.

Example 4.1. *We consider $F(Z_1, Z_2)$ defined in terms of the solution of the ODE*

$$\frac{d}{dt}u(t) = -k(z_1)u(t), \quad u(0) = g(z_2),$$

where k and g are known functions. Letting $u(t; z_1, z_2)$ denote the solution parameterized by (z_1, z_2), set $F(Z_1, Z_2) = h(u(t; Z_1, Z_2))$, where typical choices of h are $h(u) = u, h(u) = u^2, h(u) = \mathbf{1}\{a \le u \le b\}$.

[4] In some of the examples below, the function of interest, F, is more easy to evaluate than surrogate $\tilde{F}$. In these cases, the point is merely to illustrate the use and accuracy of gPC surrogate models.

Here we take $k(z_1) = z_1, g(z_2) = z_2$, let $Z_1 \sim U[0,1]$, $Z_2 \sim U[0,1]$, and compute risk sensitive integrals for $F(z_1, z_2) = [u(1; z_1, z_2)]^2$, and $F(z_1, z_2) = \mathbf{1}\{0.8 \leq u(1; z_1, z_2) \leq 1\}$. Hence the goal is to obtain robust bounds on the second moment and the probability that the solution to the stochastic ODE falls within the interval $[0.8, 1]$ at time $t = 1$, respectively. Note that the indicator function is not smooth, and thus more collocation points are required to obtain an accurate approximation. We choose a Legendre polynomial basis for both Z_1 and Z_2 and use the Legendre-Gauss quadrature points and weights for each dimension.

Figures 4.3 and 4.4 show the approximations as a function of c for $F(u) = u^2$ and $F(u) = \mathbf{1}\{0.8 \leq u \leq 1\}$, respectively. As expected Λ_c^1 gives the best bounds, while the original form Λ_c gives the worst upper bounds of the three.

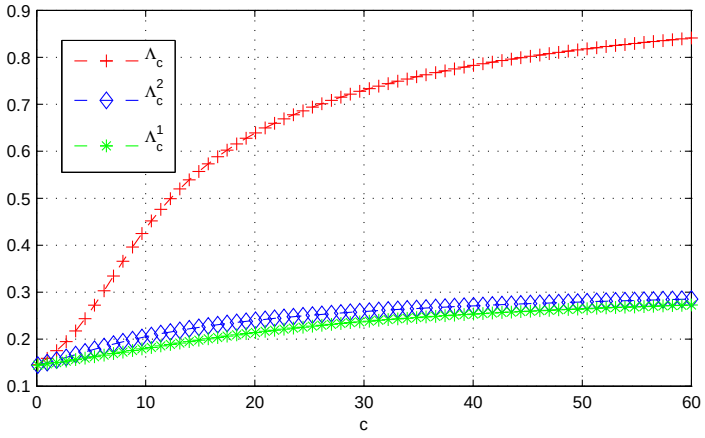

Figure 4.3: Example 4.1, $\Lambda_c, \{\Lambda_c^i\}_{i=1}^2$ for $F(u) = u^2$.

The plots also suggest what happens as $c \to \infty$. The measure Λ_c is expected to yield robust bounds if one is uncertain regarding the distributions of both random variables Z_1 and Z_2, and $\lim_{c\to\infty} \Lambda_c$ represents the tightest upper bound on performance if we know nothing about these random variables except for their support. In this case, the limit in Figure 4.4 appears to be 1, which means that for some choice of z_1 and z_2, $u(1; z_1, z_2) \in [0.8, 1]$. Hence without more information on the distributions of Z_1, Z_2 we do not obtain information other than this from this performance bound. On the other hand, $\lim_{c\to\infty} \Lambda_c^i$ for $i = 1, 2$ is strictly less than 1 and thus by Theorem 3.2 gives a meaningful performance bound when the only

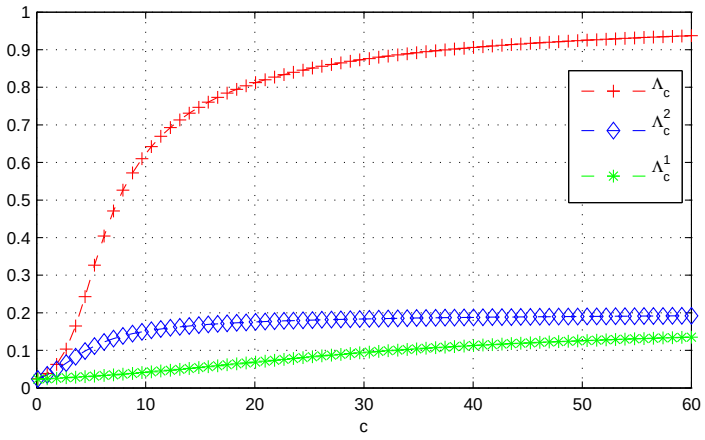

Figure 4.4: Example 4.1, $\Lambda_c, \{\Lambda_c^i\}_{i=1}^2$ for $F(u) = \mathbf{1}\{0.8 \leq u \leq 1\}$.

information available about Z_2 is its support. Note also that while Λ_c^1 and Λ_c^2 seem close for $F(u) = u^2$, the differ considerably for $F(u) = \mathbf{1}\{0.8 \leq u \leq 1\}$.

In Figures 4.5 and 4.6 we display the relative error $|\Lambda_c^i - \Lambda_{c,exact}^i| / \left|\Lambda_{c,exact}^i\right|$ between the three risk sensitive integrals and the exact solution at $c = 30$ and $c = 60$ as a function of the number of quadrature points, N^q. The exact solutions were calculated by reducing the risk-sensitive integrals to one-dimensional integrals, and then using an adaptive quadrature method. We note that the rate of convergence is significantly slower for $F(u) = \mathbf{1}\{0.8 \leq u \leq 1\}$ than $F(u) = u^2$. This is not surprising since the former is discontinuous. Also, the calculation of risk-sensitive integrals for $F(u) = \mathbf{1}\{0.8 \leq u \leq 1\}$ is not as sensitive to large c values when compared to $F(u) = u^2$. In the former case, because $0 \leq F(u) \leq 1$, large values of c do not signifcantly affect the growth of the exponents in the risk-sensitive integrals which could cause large round-off errors. However, with $F(u) = u^2$, this is not the case, and Figure 4.6 shows that there is a reduction in accuracy between $c = 30$ and $c = 60$ for the same number of quadrature points. Finally, Figure 4.6 also reveals which risk-sensitive integrals are are the slowest to converge, for a fixed c. In order from slowest to fastest convergence, we have Λ_c, Λ_c^2, and Λ_c^1. The reason that Λ_c^1 converges more quickly than Λ_c or Λ_c^2 is due to the summation of the exponential terms in (4.14) for large values of c. The calculation of Λ_c^1 involves half as many computations of exponential terms (since the summation is in the exponent), which is the main source of the loss in accuracy for

large values of c.

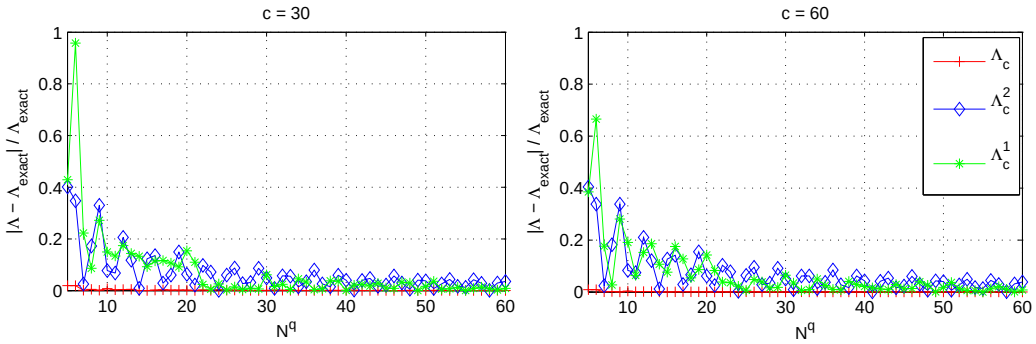

Figure 4.5: Example 4.1, relative error for $\Lambda_c, \{\Lambda_c^i\}_{i=1}^2$ with $F(u) = \mathbf{1}\{0.8 \le u \le 1\}$ for $c = 30$ (left) vs. $c = 60$ (right) for $N^q = 5, \ldots, 60$.

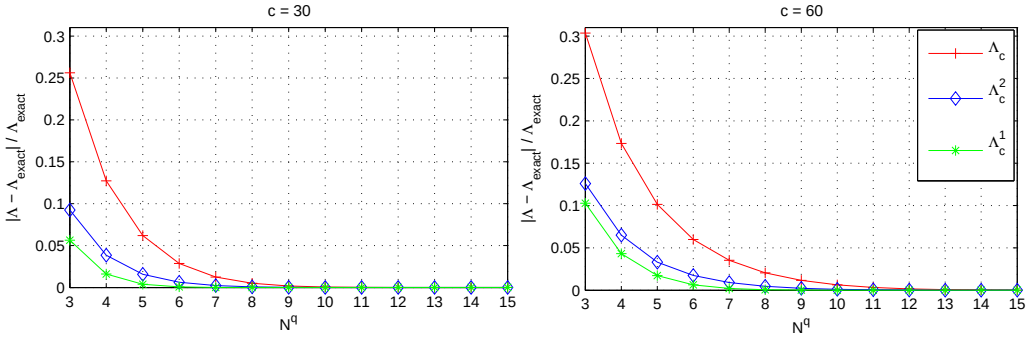

Figure 4.6: Example 4.1, relative error for $\Lambda_c, \{\Lambda_c^i\}_{i=1}^2$ for $F(u) = u^2$ for $c = 30$ (left) vs. $c = 60$ (right) for $N^q = 3, \ldots, 15$.

Next we consider the problem of computing the value of c which minimizes $c \to \frac{1}{c}B + \Lambda_c^i$ for $i = 0, 1, 2$.[5] Here B is a maximum relative entropy distance that will be allowed between the "true" distribution, $\theta(dz_2)$, and the nominal distribution, $\gamma(dz_2)$, and the minimization produces the tightest possible bounds. For the example, suppose that the family of distributions for which bounds are to be valid includes $\theta(dz_2) \sim \text{beta}(\alpha = 1.5, \beta = 1.5)$, which implies $B \approx 0.0484$. In this example $B \le f(c) \doteq c\frac{d}{dc}H^i(c) - H(c)$ for some c, where $H^i(c) \doteq c\Lambda_c^i$. As shown in Proposition 3.3, this implies there is a unique local minimum for $c \in (0, \infty)$, which is illustrated in Figures 4.7 and 4.8. Alternatively, if $B > f(c)$ for all c, then Proposition 3.3 implies the minimum is achieved in the limit as $c \to \infty$. From Figure 4.7 we find a minimum for $\frac{1}{c}B + \Lambda_c^1$ of approximately 0.04 at $c \approx 5.12$. Thus for all distributions on Z_2 whose relative entropy distance to $U[0, 1]$ is less

[5]Note that the risk sensitive integrals are infinite for $c = 0$. Thus in the numerical calculations, we actually compute c starting from 0.01.

than 0.0484, the probability that the solution to the random ODE falls between 0.8 and 1 at time 1 is less than 0.04.

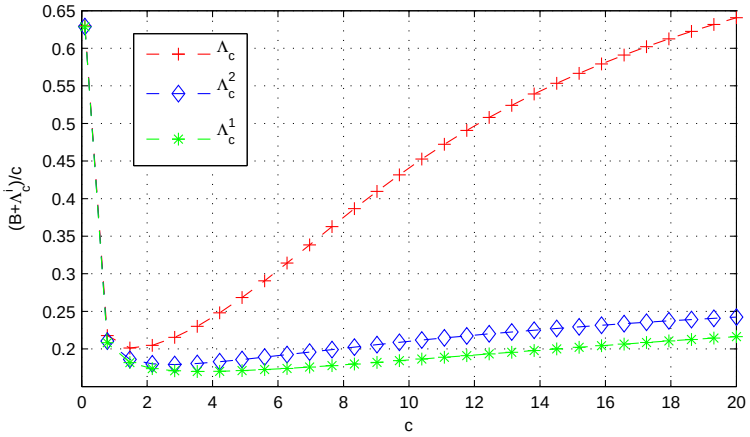

Figure 4.7: Example 4.1, $\frac{1}{c}(B+\Lambda_c)$, $\left\{\frac{1}{c}(B+\Lambda_c^i)\right\}_{i=1}^{2}$ with $F(u)=u^2$.

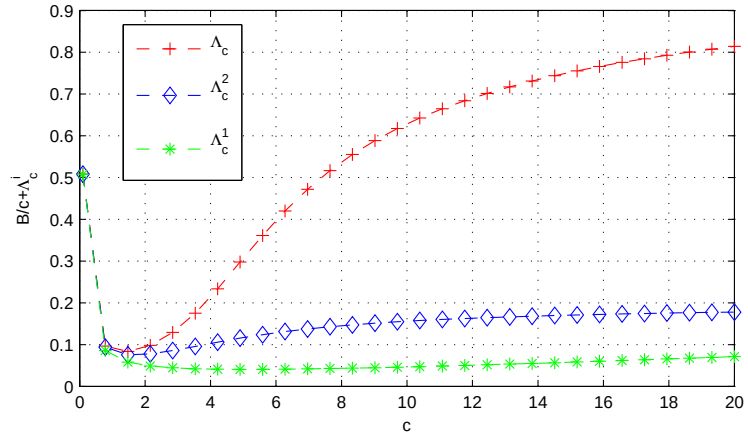

Figure 4.8: Example 4.1, $\frac{1}{c}(B+\Lambda_c)$, $\left\{\frac{1}{c}(B+\Lambda_c^i)\right\}_{i=1}^{2}$ with $F(u)=\mathbf{1}\{0.8\le u\le 1\}$.

Note that the minimum is at different values of c for the three different integrals, but the minimum is always the smallest for the first form of the hybrid. In the case when the minimum occurs at some $c<\infty$, the minimum is easily calculated, since this is a one-dimensional unconstrained minimization problem (for example, one can use a golden method search algorithm, such as MATLAB's fminbnd function). In the case when the minimum occurs in the limit as $c\to\infty$, in order to find the minimum, one can iterate Λ_c^i until the successive iterations differ by less than a prescribed tolerance.

The second example is a random oscillator, where the stochastic parameters are the spring and dampening coefficients.

Example 4.2. *We consider $F(Z_1, Z_2)$ defined in terms of the solution of the ODE*

$$\frac{d^2}{dt^2}u(t) + \gamma(z_2)\frac{d}{dt}u(t) + k(z_1)u(t) = f(t), \quad u(0) = u_0, \quad u'(0) = u_0',$$

where k and γ are known functions, which represent the spring and dampening coefficients, respectively. The outcome of interest is whether or not the position of the random oscillator falls within a specified range, $[a, b]$, at a specified time, $t_{critical}$. Letting $u(t; z_1, z_2)$ denote the solution parameterized by (z_1, z_2), set $F(Z_1, Z_2) = \mathbf{1}\{a \leq u(t_{critical}; Z_1, Z_2) \leq b\}$.

Here we consider the random oscillator with $k(z_1) = 4z_1, \gamma(z_2) = z_2/5$,and $f(t) = 10\cos(10t) + 3$, and choose $Z_1 \sim \text{beta}(\alpha = 5, \beta = 5), Z_2 \sim \text{beta}(\alpha = 5, \beta = 5)$, $t_{critical} = 4$, and $[a, b] = (-\infty, 2]$. Here we use the Gauss-Jacobi quadrature points and weights. In this example we are concerned with the position of the oscillator at a critical time. Figure 4.9 plots the risk sensitive integrals for Example 4.2 with $F(u) = \mathbf{1}\{u \leq 2\}$. Figure 4.10 depicts $c \mapsto \frac{1}{c}(B + \Lambda_c^i)$, where $B = R\left(\theta(dz_2) \,\|\gamma(dz_2)\right) \approx .0587$, with $\theta(dz_2) \sim \text{beta}(\alpha = 10, \beta = 10)$.

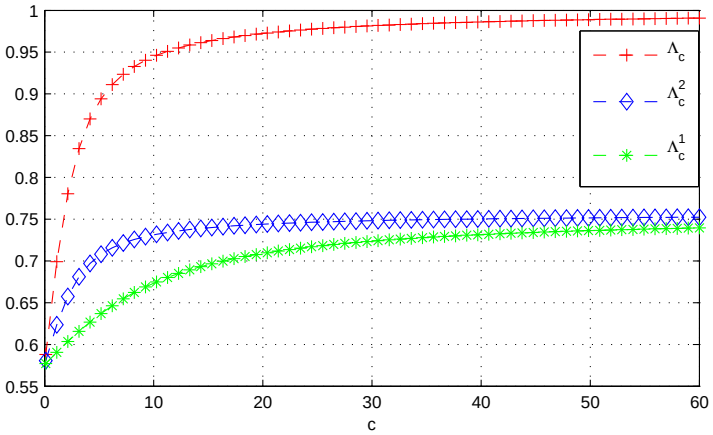

Figure 4.9: Example 4.2, $\Lambda_c, \{\Lambda_c^i\}_{i=1}^2$ for $F(u) = \mathbf{1}\{u \leq 2\}$.

Finally, the third example involves a heat equation with random heat capacity and thermal conductivity.

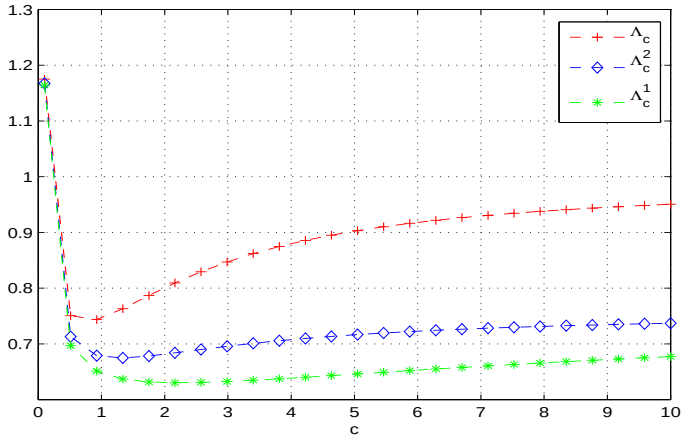

Figure 4.10: Example 4.2, $\frac{1}{c}(B+\Lambda_c)$, $\left\{\frac{1}{c}(B+\Lambda_c^i)\right\}_{i=1}^{2}$ for $F(u)=\mathbf{1}\{u\le 2\}$.

Example 4.3. *We consider $F(Z_1,Z_2)$ defined in terms of the solution of the PDE*

$$\frac{d}{dt}u(t,x)=\frac{k(u;z_1)}{m(z_2)}\frac{d^2}{dx^2}u(t,x),\quad u(0,x)=u_0(x)$$

on $x\in(0,L)$ with boundary conditions

$$-k(u;z_1)\frac{d}{dx}u(t,0)=q,\frac{d}{dx}u(t,L)=0.$$

Here k and m are the thermal conductivity and heat capacity, respectively. Note that in this example, the randomness affects the diffusivity and the boundary conditions. We are interested in the probability that the material exceeds a particular temperature, $T_{critical}$, at the time t_{final} and at some point $x^\star\in[0,L]$. Letting $u(t;z_1,z_2)$ denote the solution parameterized by (z_1,z_2), set $F(Z_1,Z_2)=\mathbf{1}\{u(t_{final},x^\star;Z_1,Z_2)\ge T_{critical}\}$.

Here we consider the random heat equation with (nonlinear) Neumann boundary conditions. We use $k(u,z_1)=z_1+(1.5\times10^{-7})u$, $m(z_2)=(10^{-6})z_2$, $q=0.35$, $L=1.90$, $u_0(x)=25$, and $T_{critical}=980$, $x^\star=0$, $t_{final}=1000$, and set $F(Z_1,Z_2)=\mathbf{1}\{u(t_{final},x^\star;Z_1,Z_2)\ge T_{critical}\}$. For the random variables, we introduce two independent random variables $\tilde{Z}_1\sim\text{beta}(\alpha=5,\beta=5),\tilde{Z}_2\sim\text{beta}(\alpha=5,\beta=5)$ and let $Z_1=(2\times10^{-3})\tilde{Z}_1+3\times10^{-3}$, $Z_2=(0.11)\tilde{Z}_2+0.30$. Figures 4.11 and 4.12 plot the risk sensitive integrals as a function of

c and illustrate $c \to \frac{1}{c}(B+\Lambda_c^i)$, where $B = R\left(\theta(dz_2)\,\|\gamma(dz_2)\right) \approx 0.0587$. and with $\theta(dz_2) \sim$ beta$(\alpha = 10, \beta = 10)$. Again, we use the Gauss-Jacobi quadrature points and weights.

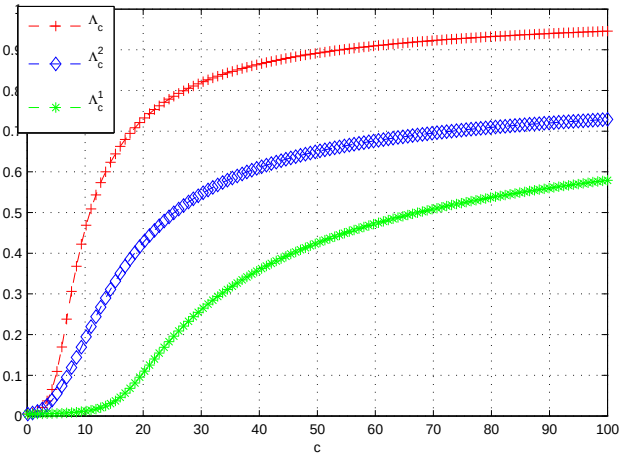

Figure 4.11: Example 4.3, $\Lambda_c, \{\Lambda_c^i\}_{i=1}^2$ for $F(u) = \mathbf{1}\{u(t_{final}, x^\star; Z_1, Z_2) \geq T_{critical}\}$.

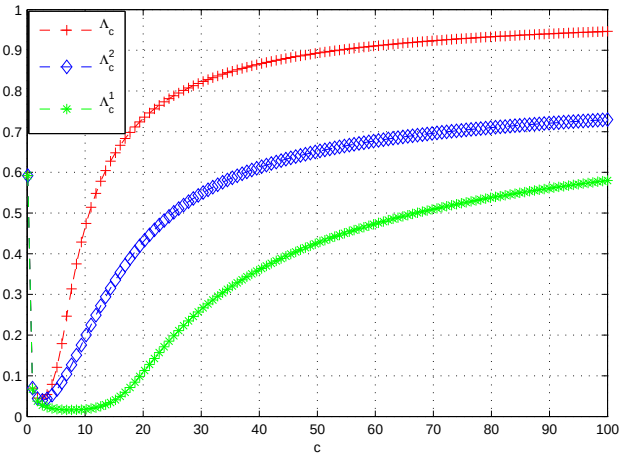

Figure 4.12: Example 4.3, $\frac{1}{c}(B+\Lambda_c)$, $\left\{\frac{1}{c}(B+\Lambda_c^i)\right\}_{i=1}^2$ for $F(u) = \mathbf{1}\{u(t_{final}, x^\star; Z_1, Z_2) \geq T_{critical}\}$.

4.2.2 Dependent Aleatoric and Epistemic Uncertainties

In this section we consider problems where the distribution of Z_1 can depend on the value of Z_2, or vice versa. In the case when Z_1 depends on Z_2, we use the following inequality,

which holds for the alternative of the first hybrid form (see the discussion in Section 3.2):

$$\int_{\mathcal{Z}_2}\int_{\mathcal{Z}_1} F(z_1,z_2)\mu(dz_1|z_2)\theta(dz_2) \le \frac{1}{c}R\left(\theta(dz_2)\,\|\gamma(dz_2)\right) + \bar{\Lambda}_c^1,$$

where

$$\bar{\Lambda}_c^1 = \frac{1}{c}\log\int_{\mathcal{Z}_2}\exp\left(\int_{\mathcal{Z}_1} cF(z_1,z_2)\mu(dz_1|z_2)\right)\gamma(dz_2).$$

Recall that the relative entropy term $\frac{1}{c}R\left(\theta(dz_2)\,\|\gamma(dz_2)\right)$ represents our lack of knowledge about the distribution of Z_2, i.e., the epistemic uncertainty where we want robustness, whereas Z_1 ∼represents aleatoric uncertainty. A question that now arises is whether or not it is still possible to use gPC methods to evaluate this integral. Consider the example with $Z_2 \sim U[0,1]$ and $Z_1 \sim N(Z_2,1)$, so that Z_1 is Gaussian with variance 1 and mean Z_2. Certainly Z_1 and Z_2 are not independent random variables. However, we can introduce an auxiliary normal random variable $Z \sim \mathcal{N}(0,1)$, and write $Z_1 \stackrel{\circ}{=} Z + Z_2$, where Z_2 and Z are independent random variables. Now we consider $G(Z,Z_2)$ instead of $F(Z_1,Z_2)$, and note that $G(Z,Z_2) = F(Z+Z_2,Z_2)$. Furthermore, $\bar{\Lambda}_c^1$ can be written as

$$\bar{\Lambda}_c^1 = \frac{1}{c}\log\int_{\mathcal{Z}_2}\exp\left(\int_{\mathcal{Z}} cG(z,z_2)\mu(dz)\right)\gamma(dz_2).$$

Hence, $\bar{\Lambda}_c^1$ can be calculated just like Λ_c^1 via a gPC approximation. Similar results hold for the case when Z_2 depends on Z_1, but the epistemic uncertainty is in Z_2.

In general, as long as we can find a smooth, invertible mapping from (Z_1,Z_2) to an independent pair of random variables, one can evaluate the integrals as before. Figures 4.13 and 4.14 show the performance of risk sensitive integrals for dependent random variables as above for Example 4.1 and with $F(u) = \mathbf{1}\{1/2 \le u \le 1\}$.

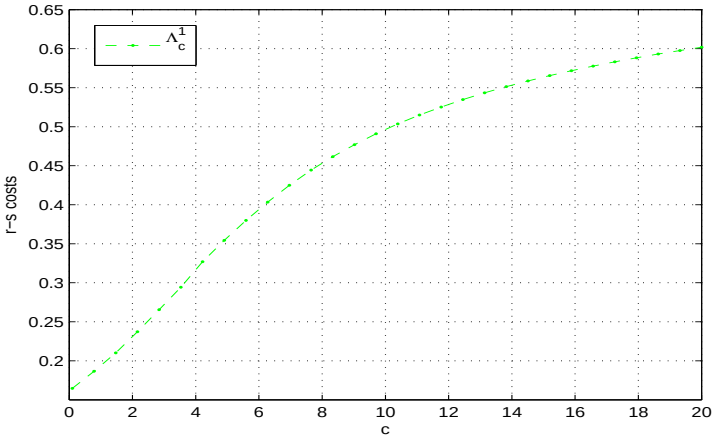

Figure 4.13: Example 4.1, $\overline{\Lambda}_c^1$ for $F(u) = \mathbf{1}\{1/2 \leq u \leq 1\}$.

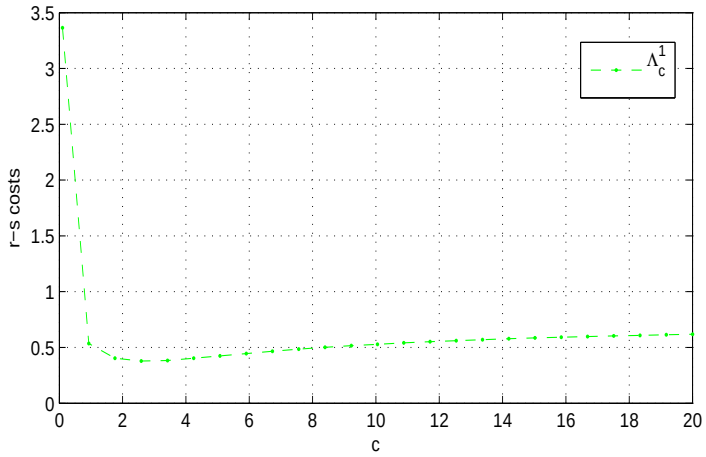

Figure 4.14: Example 4.1, $\dfrac{1}{c}(B + \overline{\Lambda}_c^1)$ for $F(u) = \mathbf{1}\{1/2 \leq u \leq 1\}$, and $\theta \sim \mathcal{N}(0.8, 1)$.

Chapter Five

Discussion and Concluding Remarks

There exist alternative methods for integrating aleatoric and epistemic variation in uncertainty quantification. The most straightforward techniques amount to treating the aleatoric and epistemic variables separately, and perform nested iterations, iterating over a finite collection of epistemic parameters on the outer loop, then sampling over aleatoric variables on the inner loop [9]. From this, one can plot the ensemble of cumulative distribution functions to visualize the family of distributions, sometimes called horsetail plots. For example, suppose the aleatoric random variable is Gaussian with an epistemic mean, where all we know about this epistemic variable is its support ($X \sim \mathcal{N}(Y, 1), Y \in [a, b]$). Then, for a set of N values contained within the known support (outer iteration), $\{Y_i = y_i : y_i \in [a, b]\}_{i=1}^{N}$, one can compute the performance (inner iteration) based on the corresponding set of input distributions, $\{X_i : X_i \sim \mathcal{N}(y_i, 1)\}_{i=1}^{N}$.

Alternatively, one can consider a Bayesian approach, and introduce hyperparameters (epistemic uncertainties) on known distributions (aleatoric uncertainties) in order to integrate both types into system modeling. For example, suppose the aleatoric random variable is Gaussian with an epistemic mean, where now we assume this epistemic variable takes on a uniform distribution ($X \sim \mathcal{N}(Y, 1), Y \sim U[a, b]$). In the parametric Bayesian approach, if one wants bounds for a collection of distributions on the epistemic variables each joint distribution must be sampled separately. In contrast to the nested iteration approach, the Bayesian hyperparameter approach produces a quantity that averages the epistemic uncertainty. Our approach gives tight upper bounds on performance measures under alternative distributions, without sampling under the various alternative joint distributions.

Most recently, in a paper by Xiu et. al. [43], an approach to dealing with epistemic uncertainty via polynomial chaos methods was developed. In [43], epistemic uncertainty meant that the true distribution of the underlying stochastic parameters was not known precisely. Here we compare our method of handling this type of epistemic uncertainty to theirs.

The approach taken in [43] is based on stochastic collocation approximation. Specifi-

cally, the differential equation of interest was solved on a generic collocation grid on the sample space of the random variables. This is in contrast to standard uses of stochastic collocation for uncertainty quantification, where the collocation grid is chosen in accordance with the (assumed known) underlying distribution. Of course the reason for the particular choice of basis polynomials and quadrature points is to obtain optimal (spectral) convergence with respect to the L^2 weighted error. In the approach of [43], regardless of what is the "true" (and assumed unknown) distribution, one chooses grid points with respect to the L^2 weighted norm with constant weight (e.g., Legendre-Gauss quadrature weights), even if the original space is unbounded. One then solves the differential equation at these collocation points and constructs the interpolating polynomial of the approximation. Hence, the gPC approximation converges optimally in the L^2 norm only if the underlying distribution is uniform. For non-uniform random variables the approximation will still converge point-wise as the number of collocation points increases (see Section 5 in [43]).

Once the interpolated approximation is obtained, one can determine the mean, variance, probabilities, and so on with respect to various underlying distributions on the parameters via Monte Carlo or other methods. The benefits are still large, in that one need not evaluate the differential equation for every Monte Carlo sample, but instead evaluates a much simpler approximation via a finite sum of polynomial basis functions. However, information can be obtained only by fixing a choice for the underlying distribution. Also, depending on which underlying distribution is used, the errors in the function approximation will have a greater or smaller effect.

In this work we have developed an approach in which performance bounds can be calculated for the mean, variance, probabilities, and so on with respect to all distributions that are within a particular relative entropy distance from the nominal distribution, and the bounds are optimal for that class of distributions. In addition to providing performance bounds on a class of distributions defined by their relative entropy distance, the limit $c \to \infty$ is particularly interesting in that it gives tight bounds on epistemic uncertainties

where one assumes nothing about the underlying distribution of certain variables except their support. In addition, the method one can efficiently handle both aleatoric and epistemic uncertainty simultaneously. More precisely, by computing certain hybrid forms of a risk-sensitive expectation we obtain tight performance bounds that distinguish between variables with known distribution and variables with unknown distribution or other forms of epistemic uncertainty. In particular, the scenarios to which the method applies include (i) aleatoric with known distribution; (ii) aleatoric with partly known distribution (mingled aleatoric and epistemic); (iii) epistemic for which one is willing to model by a family of aleatoric uncertainties, and (iv) epistemic where one is only willing to place bounds on the uncertainties. For all the examples we have considered (including those presented in the paper), approximation of the required risk-sensitive integrals can be done via gPC methods using roughly the same computational effort as would be required to compute the corresponding ordinary performance measures using the nominal probability distributions, although possibly at a lower rate.

Part II

Sparse Gradient Image Reconstruction from Incomplete Fourier Measurements and Prior Edge Information

Chapter One

Introduction

Magnetic Resonance Imaging (MRI) is a widely used medical imaging technique to visualize detailed soft tissue structures of the body, e.g., brain, muscles, heart, etc. Due to its ability to employ contrast weightings based on physiological function, e.g. perfusion, applications of MRI have been developed requiring the acquisition of time series information [28]. For example, functional Magnetic Resonance Imaging (fMRI) uses changes in local blood flow and blood oxygenation levels to reveal regions of neural activity. An important consideration in fMRI is the total scan time required to obtain an accurate representation of the the region of interest (which determines the temporal resolution of the time series). Temporal resolution must be adequate for rendering the time varying property of the region of interest (contrast change due to changes in local blood flow and oxygenation). For example, on a typical clinical scanner, the achievable temporal resolution using conventional echo-planar acquisitions with whole-brain coverage is on the order of 2-2.5 seconds. Thus, a reduction in total scan time and improved temporal resolution for fMRI is highly desirable, especially for functional connectivity studies. Because MRI and fMRI techniques acquire a set of Fourier or frequency space measurements for image reconstruction [23, 26], an obvious solution to this problem is to reduce the amount of frequency space information obtained for each image and time point.

One method of reducing the amount of Fourier or frequency space measurements is based on the theory of Compressed Sensing (CS). CS indicates that it is possible to reconstruct an image accurately, and sometimes even exactly, from random, incomplete Fourier measurements [4, 11]. The set of Fourier measurements are incomplete in the sense that one cannot simply perform a discrete Fourier transform (DFT) matrix inversion to recover the image. The Nyquist-Shannon sampling theorem tells us that in order to reconstruct the image exactly, we need to acquire the same number of Fourier measurements as the resolution of the image. However, CS suggests that accurate recovery can be achieved by sampling far fewer Fourier measurements than specified by Nyquist-Shannon.

Many types of images exhibit sparsity in some domain, e.g. Fourier, gradient, wavelet, etc. In this paper, we will focus on images that exhibit sparsity with respect to their

gradients. For example, consider the widely used Shepp-Logan phantom image [35] in Figure 1.1.

Figure 1.1: (Left) 256×256 Sparse Gradient Shepp-Logan phantom image. (Right) Nonzero horizontal and vertical discrete gradient information.

Using CS, it is possible to recover a sparse gradient image from random Fourier measurements via a convex ℓ_1-minimization problem. More precisely, an image $X \in \mathbb{C}^{N\times N}$ that is s-sparse with respect to the discrete gradient operator, i.e., X has only s nonzero discrete gradients, can be recovered exactly from only $m \geq \mathcal{O}(s \log^4 N^2)$ discrete Fourier measurements, chosen uniformly at random, via a convex ℓ_1-minimization problem [11, 33, 32, 4]. In practice this bound is much too conservative and it is conjectured that a theoretical lower bound of $\mathcal{O}(s \log N)$ is possible [30]. In comparison, the Nyquist-Shannon sampling theorem requires that all N^2 discrete Fourier measurements are required for exact reconstruction.

In this work, we propose a new method to further reduce the amount of Fourier measurements, while still obtaining an exact or stable reconstruction, by utilizing additional, prior edge information from a high-resolution reference image. In this context, the edge information of a reference image $Y \in \mathbb{C}^{N\times N}$ refers to both the indices for which the gradient is nonzero and the corresponding gradient magnitudes.[1] The reference image itself is not acquired directly, but rather the full sampled two-dimensional Fourier measurements $\hat{Y} \in \mathbb{C}^{N\times N}$ are obtained. The Fourier transform information of this reference image is typically acquired prior to an fMRI time series study and is used for grey/white matter segmentation. Under the assumption that both the reference image and sparse-gradient time series images share similar edge information and the same field of view, our proposed

[1] For an image $Y \in \mathbb{C}^{N\times N}$ the gradient refers to the horizontal and vertical first-order finite differences. See Section 2.3 for a precise definition.

algorithm seeks to reduce the number of Fourier measurement by combining CS techniques for sparse-gradient images with edge information from the Fourier measurements of the reference image. Figure 1.2 shows an example of a reference image (left) and a sparse gradient image at a single time point (right). These two images share almost identical edge information except for different magnitudes in their gradients. Figure 1.3 shows a cross-section of the two images to illustrate the difference between their gradients. This represents the case where the reference image is acquired at the same resolution as the time series images. The case where the reference image is acquired at higher spatial resolution than the time series is discussed subsequently.

Figure 1.2: (Left) Reference image from which we have fully sampled Fourier measurements taken prior to the fMRI study. (Right) Image at a single time point, which we are trying to reconstruct from incomplete Fourier measurements.

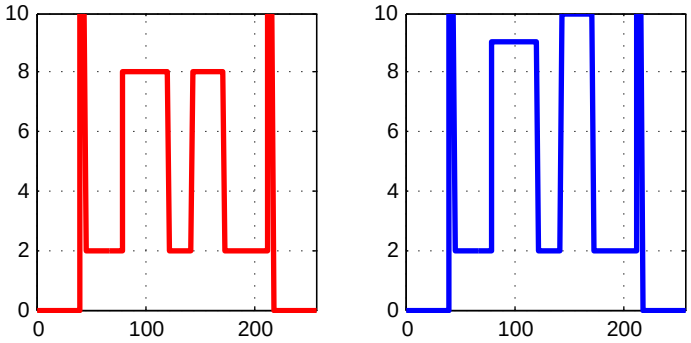

Figure 1.3: Comparison of center cross section of the two images in Figure 1.2. The red cross section (left) shows the reference image and the blue cross-section shows the image at a single time point.

In order to extract edge information from the Fourier measurements of the reference image, we will utilize the theory developed in a series of papers by Gelb and Tadmor on the detection of edges from Fourier coefficients [14, 13, 16]. This will allow us to determine approximate edge information about the sparse gradient time series. Then, by properly combing this approximate edge information with the Fourier measurements of the time series images via our proposed algorithm, we will show that the number of Fourier

measurements required for accurate, or even exact, reconstruction is significantly lower than for total-variation techniques alone.

This work is organized as follows. Chapter 2 introduces the theory of CS and extensions to sparse-gradient image recovery from random Fourier measurements. Chapter 3 introduces the basic concepts of edge detection from Fourier coefficient data, and explains how one can extract the relevant edge information from the reference image. Chapter 4 introduces the main algorithm which combines CS techniques for sparse-gradient images with edge information from the reference image. Finally, Section 4.3 presents three sets of numerical experiments for the reconstruction of a time series from its Fourier measurements and prior edge information using our proposed algorithm.

Chapter Two

Compressed Sensing and Extensions

Compressed Sensing (CS) gained considerable attention in recent years after Candes, Tao and Romberg published a seminal paper on exact signal recovery from incomplete, random frequency information [4], which built upon previous work by Donoho et al. on basis pursuit methods [5, 6]. At its core, the theory of CS is concerned with recovering sparse signals accurately, and sometimes even exactly, from incomplete, random measurements. Traditional approaches to signal or image reconstruction require that for an N length signal, one must obtain N frequencies, i.e., at the Nyquist-Shannon sampling rate, or N linearly independent measurements. However, through CS, the mathematical groundwork has been laid for obtaining accurate signal reconstruction from surprisingly far fewer measurements.

In this chapter we will introduce the reader to the basic theory of CS, the random partial Fourier matrix, extensions to sparse gradient image recovery, and numerical methods for ℓ_1-minimization. If the reader is familiar with the basics of CS theory, he or she may skip to Sections 2.3 and 2.4 where sparse gradient image recovery and numerical methods for ℓ_1-minimization are discussed. For a more thorough treatment see [11] and [33].

2.1 Basics of Compressed Sensing

2.1.1 Sparsity

Consider a complex vector $x \in \mathbb{C}^n$ and its associated ℓ_p-norm defined as

$$\begin{aligned} \|x\|_p &= \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}, \quad 0 < p < \infty, \\ \|x\|_\infty &\doteq \max_{1 \le j \le n} |x_j|. \end{aligned}$$

Note that for $0 < p < 1$, the ℓ_p-norm is only a quasi norm, e.g. the triangle inequality is not satisfied. Consider a complex matrix $A \in \mathbb{C}^{m \times n}$, where $m \ll n$ in typical CS applications.

Define the operator norm of a matrix to be

$$\|A\|_p \triangleq \max_{x\in\mathbb{C}^n, \|x\|_p=1} \|Ax\|_p,$$

where $Ax \in \mathbb{C}^m$ (see Section B.1 for properties of matrix and vector norms for $p = 1, 2, \text{and } \infty$). We say that $x \in \mathbb{C}^n$ is s-sparse if

$$\|x\|_0 \triangleq \{\#i : x_i \neq 0\} \leq s,$$

where the ℓ_0-norm counts the number of non-zero entries in the vector x. When x is not exactly s-sparse, the best s-sparse approximation is defined as

$$\sigma_s(x)_p \triangleq \inf\{\|x - z\|_1 : z \text{ is } s\text{-sparse}\}. \tag{2.1}$$

In many applications, x is not measured directly. Rather, one is given a set of $m \ll n$ linear measurements

$$y = Ax, \quad y \in \mathbb{C}^m,$$

where A is referred to as the measurement matrix.

Example 2.1. *In many imaging applications one acquires discrete Fourier measurements of a signal or image x. Define the discrete Fourier matrix $F \in \mathbb{C}^{n\times n}$ with entries*

$$F_{l,j} = \frac{1}{\sqrt{n}} e^{2\pi i(l-1)(j-1)/n},$$

for $l, j \in 1, \ldots, n$ and $i^2 = -1$. The discrete Fourier transform of a signal x is given by $\hat{x} \triangleq Fx \in \mathbb{C}^n$. In practice, one may only observe $m \ll n$ Fourier measurements out of $\hat{x}$. Let $\Omega \subset \{1, \ldots, n\}$ denote the sub-sampled frequencies or Fourier measurements of cardinality m, i.e., $|\Omega| = m$. More precisely, let Ω be the of frequencies $w_l = \omega_{l_j}$ for $1 \leq j \leq m, w_{l_j} \in \{1, \ldots, n\}$. The measurement matrix $F_\Omega \in \mathbb{C}^{m\times n}$ can be constructed by

taking the rows of F from the index set Ω, where

$$(F_\Omega x)_l \quad = \quad \hat{x}_{w_l},$$

for $l = 1, \ldots, m$.

The goal is to determine x from the m linear measurements y. One way to do this might be to solve the ℓ_0-minimization problem

$$\min_{z \in \mathbb{C}^n} \|z\|_0 \quad \text{subject to} \quad Az = y, \tag{2.2}$$

and hope that the minimizer to (2.2) is the s-sparsest solution. However, what guarantees do we have that the solution to (2.2) is unique and, moreover, equal to the solution x? The following two examples may provide some valuable insight into the answers.

First, recall Example 2.1 and let $\Omega = \{1, \ldots, n\}$ so that $m = n$. Then $A = F$ is unitary and the transformation Az is an isometry, which means that the measurements have the same ℓ_2 energy as the original signal, i.e. $\|y\|_2 = \|x\|_2$. In this case, x can be recovered by a simple inversion. This is the ideal situation, but it motivates an important concept that we will quantify later: the measurement matrix is preferred to be as close to an isometry as possible so that the *energy* of the measurements is similar or comparable to the *energy* of the exact solution.

Second, assume that every $2s$ set of columns of the measurement matrix A defines an injective mapping, i.e. all sets of $2s$ columns are linearly independent.[1] Under this assumption, we claim that the solution to (2.2) is unique. Consider two s-sparse vectors z_1 and z_2 such that $Az_1 = Az_2 = y$. Define $z_3 \stackrel{\circ}{=} z_1 - z_2$ so that $z_3 \in \text{Null}(A)$. Since z_1 and z_2 are s-sparse, z_3 is at worst $2s$-sparse. Because all $2s$ columns are linearly independent, $Az_3 = 0$ implies that $z_3 \equiv 0$. Hence, $z_1 \equiv z_2$ and uniqueness is guaranteed. In turns out that when A is the partial discrete Fourier matrix, as in Example 2.1, any $2s$ columns are

[1] Assume that $m \gg s$. Otherwise, if $2s > m$, it would be impossible to have $2s$ linearly independent m length vectors.

linearly independent, as long as $m > 2s$ and n is prime [4, Lemma 1.2].

These two examples hint at two very important properties of the measurement matrix: the null space property and the restricted isometry property. These two properties provide the foundation for CS theory and will be defined more precisely in the subsequent sections.

Uniqueness aside, is it even computationally feasible to solve (2.2)? The answer is no, even for modest-sized measurements. It turns out that (2.2) is an NP-hard problem and the only known way to solve this problem would be to consider all s subsets of n columns of A, and determine if the solution lies in the range of this s-dimensional sub-matrix [4]. The number of subsets scales exponentially and becomes computationally intractable very quickly. Thus, solving (2.2) is not realistic. At the other extreme, we could replace the ℓ_0-norm with the familiar ℓ_2-norm. Then the problem is clearly convex and one can even write down an analytic solution. However, there is no guarantee that the minimizer with respect to the ℓ_2-norm will be the sparsest solution.

A computational efficient strategy, that preserves the search for sparsity, is the convex ℓ_1-minimization problem

$$\min_{z \in \mathbb{C}^n} \|z\|_1 \quad \text{subject to} \quad Az = y \ . \tag{2.3}$$

One of the key developments in CS is the equivalence between the ℓ_0 and ℓ_1-minimization problem. But why should we even expect the solution to (2.3) to be sparse? Consider a simple two-dimensional example with $n = 2$ and $m = 1$, so that all feasible solutions lie on a line $Az = y$. Clearly, the sparsest solution lies on one of the axes. Figure 2.1 shows the feasible set $Az = y$ and compares the minimizer under different ℓ_p-norms for $p = 1, 2$, and $1/2$.

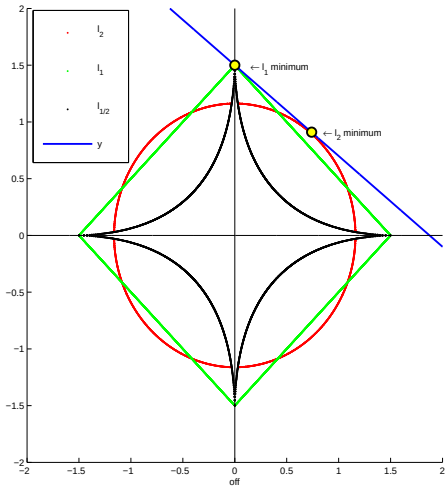

Figure 2.1: The blue line represents all feasible solution that satisfy $Az = y \in \mathbb{R}$. The green diamond is the ℓ_1-ball with radius equal to the sparsest solution, while the red circle represents the ℓ_2-ball with radius equal to the minimum ℓ_2 solution. The black curve represents the $\ell_{1/2}$-ball with radius also equal to the sparsest solution, but is not convex

2.1.2 Null Space Property (NSP)

In this section, we determine the conditions for which ℓ_1-minimization is equivalent to the ℓ_0-minimization problem, i.e., when (2.3) is equivalent to (2.2). Let z^* be the unique s-sparse minimizer to (2.3). Denote the index of sparsity by $\Omega \triangleq \{i : z_i^* \neq 0, 1 \leq i \leq n\}$ and let Ω^c refer to its complement, i.e., $\Omega^c \triangleq \{i : z_i^* = 0, 1 \leq i \leq n\}$. Then,

$$\sum_{i=1}^{n} |z_i^* + h_i| \quad > \quad \sum_{i=1}^{n} |z_i^*|$$

for all $h \neq 0, h \in \mathbb{C}^n$ and $Ah = 0$. The triangle inequality gives

$$\begin{aligned} \sum_{i=1}^{n} |z_i^* + h_i| &= \sum_{i\in\Omega} |z_i^* + h_i| + \sum_{i\in\Omega^c} |h_i| \\ &\geq \sum_{i\in\Omega} |z_i^*| - |h_i| + \sum_{i\in\Omega^c} |h_i| = \sum_{i\in\Omega} |z_i^*| + \left(\sum_{i\in\Omega^c} |h_i| - \sum_{i\in\Omega} |h_i| \right). \end{aligned}$$

Hence, z^* is the unique minimizer to (2.3) if

$$\sum_{i\in\Omega^c} |h_i| > \sum_{i\in\Omega} |h_i|, \quad \forall h \in \mathbb{C}^n, h \neq 0, Ah = 0.$$

In other words, while z^* may be concentrated on Ω, the vectors that lie in the null space of the measurement matrix *cannot* be concentrated on Ω. In fact, the above inequality says that $\sum_{i\in\Omega}|h_i| \leq \frac{1}{2}\|h\|_1$ for $h \in \text{Null}(A)$. This motivates the following definition and resulting theorem, which guarantees sparse recovery from ℓ_1-minimization [34, Theorem 2.3].

Definition 2.1. *A matrix $A \in \mathbb{C}^{m\times n}$ satisfies the Null Space Property (NSP) of order s if for all subsets $\Omega \subset \{1,\dots,n\}$ with $|\Omega| = s$, it holds*

$$\sum_{i\in\Omega}|h_i| \quad < \quad \sum_{i\in\Omega^c}|h_i|, \quad \forall h \in \text{Null}(A)\backslash\{0\}.$$

Theorem 2.1. *Let $A \in \mathbb{C}^{m\times n}$ and $x \in \mathbb{C}^n$ be a s-sparse vector. Suppose we are given measurements $y \in Ax, y \in \mathbb{C}^m$. Then x is the unique solution to the ℓ_1-minimization problem*

$$\min_{z\in\mathbb{C}^n}\|z\|_1 \quad \textit{subject to} \quad Az = y$$

if and only if A satisfies the NSP of order s.

The NSP is often times difficult to show. Instead, we consider an equivalent property known as the Restricted Isometry Property (RIP). The connection between the RIP and the NSP is not too surprising following our discussion in Section 2.1.1.

2.1.3 Restricted Isometry Property (RIP)

The definition of the restricted isometry constant is as follows.

Definition 2.2. *The restricted isometry constant δ_s of a matrix $A \in \mathbb{C}^{m\times n}$ is defined as the smallest δ_s such that*

$$(1-\delta_s)\|x\|_2^2 \quad \leq \quad \|Ax\|_2^2 \quad \leq \quad (1+\delta_s)\|x\|_2^2 \tag{2.4}$$

for all s-sparse $x \in \mathbb{C}^n$.

A matrix satisfies the restricted isometry property (RIP) if the restricted isometry constant(RIP constant) is reasonably small for a reasonably large value of s. The definition will be made more precise later. As an example, if $A \in \mathbb{C}^{n\times n}$ is the full discrete Fourier matrix (see Example 2.1) then A is an isometry and $\delta_s = 0$ for all s-sparse vectors. Ideally one would like δ_s as small as possible, i.e., as close to an isometry as possible.

Remark 2.1. *The RIP constant must be valid for* all *s-sparse vectors in $\mathbb{C}^n$. It is not dependent on any one particular s-sparse vector. If we define $\delta_s(x)$ be the constant that satisfies (2.4) for a particular s-sparse vector x, then the restricted isometry constant must satisfy*

$$\delta_s \quad = \quad \sup_{x\in\mathbb{C}^n, \|x\|_0\leq s} \delta_s(x).$$

The following are some properties of the RIP constant [34].

Proposition 2.1. *Let $A \in \mathbb{C}^{m\times n}$ with RIP constants δ_s. Then*

a) The RIP constants are ordered $\delta_1 \leq \delta_2 \leq \delta_3 \leq \cdots$.

b) It also holds that

$$\begin{aligned} \delta_s &= \max_{\Omega\subset\{1,\ldots,n\}, |\Omega|\leq s} \|A_\Omega^* A_\Omega - I\|_2 \\ &= \sup_{x\in T_k} |\langle A^*A - I)x, x\rangle| \end{aligned}$$

where $A_\Omega \in \mathbb{C}^{m\times s}$ is a sub-matrix of A created by concatenating the columns in the index set Ω of A, and $T_k = \{x \in \mathbb{C}^N, \|x\|_2 = 1, \|x\|_0 \leq s\}$.

c) Let $u, v \in \mathbb{C}^n$ with disjoint supports, i.e. $\operatorname{supp}(u) \cap \operatorname{supp}(v) = \emptyset$. Let $s = \operatorname{supp}(u) + \operatorname{supp}(v)$. Then,

$$|\langle Au, Av\rangle| \quad \leq \quad \delta_s \|u\|_2 \|v\|_2.$$

Let us take a closer look at part (b) from the above Proposition. If A satisfies the RIP constant of δ_s, then

$$|\langle A^*A - I)x, x\rangle| \quad \leq \quad \delta_s$$

for any s-sparse vector $x \in \mathbb{C}^n$. For $\Omega \subset \{1, \ldots, n\}, |\Omega| \leq s$, this can be re-written as

$$|\langle A_\Omega^* A_\Omega - I)x_\Omega, x_\Omega \rangle| \quad \leq \quad \delta_s \ ,$$

where $x_\Omega = \{f \in \mathbb{C}^{|\Omega|} : f_i = z_{i_k}, 1 \leq i, k \leq |\Omega|, i_k \in \Omega\}$ and $A_\Omega \in \mathbb{C}^{m \times s}$ is a sub-matrix created by concatenating the columns in the index set Ω of A. This inequality implies that the eigenvalues of the Hermitian operator $A_\Omega^* A_\Omega - I$ are bounded by δ_s. Denote $\lambda(A_\Omega^* A_\Omega)$ to be the eigenvalues of $A_\Omega^* A_\Omega$. Then part (b) says

$$1 - \delta_s \quad \leq \quad \lambda(A_\Omega^* A_\Omega) \quad \leq \quad 1 + \delta_s \ .$$

which implies that the singular values of A_Ω, denoted by $\sigma(A_\Omega)$, must satisfy

$$\sqrt{1 - \delta_s} \quad \leq \quad \sigma\left(A_\Omega\right) \quad \leq \quad \sqrt{1 + \delta_s} \ . \tag{2.5}$$

Thus, the RIP constant provides a bound on the condition number of A_Ω, i.e., the ratio between the largest and smallest singular values of A_Ω. For the case when $\delta_s \in (0, 1)$, (2.5) shows that the singular values are necessarily non-zero. Hence, the matrix A_Ω must be full rank, i.e., the columns must be linearly independent.[2] This result is summarized in the following corollary.

Corollary 2.2. *Let $A \in \mathbb{C}^{m \times n}$ with isometry constant δ_s. Then, the condition number of A_Ω is bounded by $\sqrt{\dfrac{1 + \delta_s}{1 - \delta_s}}$, with $|\Omega| \leq s$. In particular, if $\delta_s \in (0, 1)$, then any Ω columns of A are linearly independent.*

2.1.4 ℓ_1 Recovery

In this section we introduce the two main theorems regarding RIP constants for the exact (noiseless measurements) and stable (noisy measurements) recovery of a s-sparse signal via ℓ_1-minimization. The proofs are presented in Section B.2.

[2] Recall the discussion in Section 2.1.1 about the linear independence of columns and the uniqueness of the ℓ_0-minimization problem.

Theorem 2.3. *Suppose the RIP constant δ_{2s} of a matrix $A \in \mathbb{C}^{m\times n}$ satisfies*

$$\delta_{2s} \quad \leq \quad \frac{1}{3}.$$

Then, the null space property (NSP) of order s is satisfied. In particular, every s-sparse vector $x \in \mathbb{C}^n$ is recoverable from measurements Ax via ℓ_1-minimization (2.3).

Theorem 2.4. *Assume that the restricted isometry constant δ_{2s} of the matrix $A \in \mathbb{C}^{m\times n}$ satisfies*

$$\delta_{2s} \quad < \quad 2(3-\sqrt{2})/7 \quad \approx \quad .4531.$$

Then, the following holds for all $x \in \mathbb{C}^n$. Let noisy measurements $y = Ax+\varepsilon$ be given with $\|\varepsilon\|_2 \leq \eta$ and let x^* be the solution of

$$\min_{z\in\mathbb{C}^N} \quad \|z\|_1 \quad \textit{subject to} \quad \|Az-y\| \leq \varepsilon. \tag{2.6}$$

Then,

$$\begin{aligned} \|x-x^*\|_1 &\leq C_1\sigma_s(x)_1 + D_1\sqrt{s}\varepsilon, \\ \|x-x^*\|_2 &\leq C_1\frac{\sigma_s(x)_1}{\sqrt{s}} + D_2\varepsilon, \end{aligned}$$

where $\sigma_s(x)_1$ is defined in (2.1) and for some constants $C_i, D_i > 0$ that depend only on δ_{2s}. In particular, if x is s-sparse and $\varepsilon \equiv 0$, then we obtain exact recovery.

Note that if $\delta_{2s} < 1/3$ we can also guarantee stable recovery with noisy measurements. Theorems 2.3 and 2.4 allow us to make very powerful statements about the equivalence between ℓ_0 and ℓ_1-minimization from the RIP constants. The next section introduces the concept of coherence which can be used to measure the RIP constant.

2.1.5 Coherence

Denote the columns of $A \in \mathbb{C}^{m\times n}$ by $\{a_j\}_{j=1}^n$. For orthonormal columns, the coherence is defined as

$$\mu \triangleq \max_{j\neq k} |\langle a_j, a_k\rangle|.$$

For non-normalized columns, the coherence can be defined as

$$\mu \triangleq \max_{j\neq k} \frac{|\langle a_j, a_k\rangle|}{\|a_j\|_2\|a_k\|_2}.$$

A small coherence is desirable. This means that the measurements are as *different* from each other as possible. Ideally, one would like the columns to be orthogonal or orthonormal. For example, consider again the full discrete Fourier matrix F as in Example 2.1. Since this matrix is unitary, the coherence is zero.

If $x \in \mathbb{C}^{n\times n}$ is s-sparse, then $Ax \in \mathbb{C}^m$ is effectively determined by some combination of s or less columns of A. Therefore, let us introduce a refinement of the coherence function known as the 1-coherence function or Babel function

$$\mu_1(s) := \max_{\ell\in\{1,\ldots,n\}} \max_{S\subset\{1,\ldots,n\}\setminus\{\ell\},|\Omega|\leq s} \sum_{j\in\Omega} |\langle a_j, a_l\rangle|.$$

Below are some properties of μ and μ_1 in relation to the restricted isometry constant [34].

Proposition 2.2. *Let $A \in \mathbb{C}^{m\times n}$ with unit norm columns for simplicity, coherence μ, 1-coherence function $\mu_1(s)$ and restricted isometry constant δ_s. Then,*

a) $\mu = \delta_2$,

b) $\mu_1(s) = \max_{\Omega\subset\{1,\ldots,n\},|\Omega|\leq s+1} \|A_\Omega^* A_\Omega - I\|_1$,

c) $\delta_s \leq \mu_1(s-1) \leq (s-1)\mu$,

where $A_\Omega \in \mathbb{C}^{|\Omega|\times n}$ is the sub-matrix of $|\Omega|$ columns of A taken from the index set Ω.

Theorem 2.3 guarantees exact recovery via ℓ_1-minimization provided that $\delta_{2s} < 1/3$. Combining this result with part (c) above, exact recovery can be guaranteed provided that the coherence or Babel function satisfies $(2s-1)\mu < 1/3$ or $\mu_1(2s-1) < 1/2$, respectively. While it may seem tempting to use the coherence to place bounds on the RIP constant, the example below shows that doing so leads to unsatisfactory results.

Example 2.2. *Consider the measurement matrix $A = (I|F) \in \mathbb{C}^{m\times 2m}$ which is a concatenation of the identity matrix $I \in \mathbb{C}^{m\times m}$ and the unitary discrete Fourier matrix $F \in \mathbb{C}^{m\times m}$. Clearly, the coherence is $\mu = 1/\sqrt{m}$. By part (c) of Proposition 2.2 and Theorem 2.3, exact recovery is possible with $(2s-1)/\sqrt{m} < 1/3$ or*

$$m \;\geq\; 9(2s-1)^2.$$

This might seem like a satisfactory result, but as s grows, m grows quadratically and this result becomes less satisfactory. Unfortunately, it turns out that one cannot improve upon this quadratic dependence when using the coherence alone to bound the RIP constant. This is due to a lower bound on the coherence

$$\mu \;\geq\; \sqrt{\frac{n-m}{m(n-1)}}$$

which scales like $m^{-1/2}$ when $n \gg m$, so that $m \geq \mathcal{O}(s^2)$ at best [36].

2.1.6 Sparsity in Another Basis

Suppose $x \in \mathbb{C}^n$ is sparse with respect to the basis spanned by the linearly independent columns of the matrix $\Psi \in \mathbb{C}^{n\times n}$. That is, $x = \sum_{i=1}^{n} \psi_i c_i$ where $\{\psi_i\}_{i=1}^n$ are the orthonormal columns of Ψ and $c = (c_1, \ldots, c_n)^T \in \mathbb{C}^n$ is s-sparse. Given m measurements $y \triangleq \Phi x \in \mathbb{C}^m$ where $\Phi \in \mathbb{C}^{n\times m}$ is the measurement matrix, the ℓ_1-minimization problem (2.3) can be re-written as

$$\min_{z\in\mathbb{C}^n} \|z\|_1 \quad \text{subject to} \quad \Phi\Psi z = y. \tag{2.7}$$

Then, Theorem 2.3 or 2.4 gurantees exact or stable recovery, respectively, if the RIP constant for $\Phi\Psi$ satisfies the requisite bound.

In a slightly different scenario, suppose x is s-sparse after a linear transformation under Ψ, i.e., $\Psi x \in \mathbb{C}^n$ is a s-sparse vector. Again, given m measurements $y = \Phi x$, the ℓ_1-minimization problem (2.3) can be re-written as

$$\min_{f \in \mathbb{C}^n} \|f\|_1 \quad \text{subject to} \quad \Phi\Psi^{-1} f = y, \tag{2.8}$$

where the substitution $f = \Psi z$ was made. In this scenario, the RIP constant for $\Phi\Psi^{-1}$ will determine exact or stable recovery.

2.2 Random Fourier Measurement Matrices

In many imaging applications such as Magnetic Resonance Imaging (MRI) [26] and Synthetic Aperture Radar (SAR) [29] a set of Fourier measurements of an unknown image is acquired, from which the image must be reconstructed. Ideally, one would like to reconstruct the image or signal from the fewest possible set of Fourier measurements, every single one of which has a potential time and/or power cost. CS can be used to obtain accurate or even exact recovery from a relatively small set of random Fourier frequencies.

Recall the discrete Fourier matrix $F \in \mathbb{C}^{n \times n}$ with entries

$$F_{l,k} \quad = \quad \frac{1}{\sqrt{n}} e^{2\pi i (l-1)(k-1)/n},$$

for $l, k = 1, \ldots, n$, which was first introduced in Example 2.1. It is well known that the columns of this matrix are orthonormal, so that F is a unitary matrix.[3] For a vector

[3] F is also symmetric, but not Hermitian.

$x \in \mathbb{C}^n$, let $\hat{x} \in \mathbb{C}^n$ be its discrete Fourier transform, where

$$\hat{x}_j \quad = \quad \frac{1}{\sqrt{n}} \sum_{l=0}^{n-1} x_l e^{-2\pi i (j-1) l/n}, \qquad j = 1, \ldots, n.$$

This matrix vector product can be computed in $\mathcal{O}(n \log n)$ operations rather than $\mathcal{O}(n^2)$ using the Fast Fourier Transform (FFT) algorithm [37].

Let us define the random partial Fourier matrix as follows. Let m be the number of Fourier samples or measurements. Choose a subset $\Omega \subset \{1, \ldots, n\}$ of size m uniformly at random from all sets of this size; i.e. each of the $\binom{n}{m}$ possible subsets are equally likely. Define the random discrete partial Fourier matrix $F_\Omega \in \mathbb{C}^{m \times n}$ as the m randomly chosen rows of F from the index set Ω. The operation $F_\Omega x : \mathbb{C}^n \to \mathbb{C}^m$ gives us the m randomly chosen Fourier transforms corresponding to the index set Ω, i.e.,

$$(F_\Omega x)_l \quad = \quad \hat{x}_{j_l}, \quad 1 \leq l \leq m, j_l \in \Omega.$$

The discrete Fourier transform matrix can also be extended to any higher dimension $d \in \mathbb{N}, d \geq 1$, where the discrete basis functions are of the form $z \cdot e^{2\pi i \mathbf{l} \cdot \mathbf{k}/\mathbf{n}}$, where $\mathbf{l} = (l_1, \ldots, l_d)$, $\mathbf{k} = (k_1, \ldots, k_d)$, $\mathbf{n} = (n_1, \ldots, n_d)$, and $z = (\prod_{l=1}^{d} N_l)^{-1}$. The results in this section are analogous to the higher-dimensional setting [34].

Sufficient bounds on the RIP constant for exact recovery cannot be guaranteed with any deterministic choice of indices in Ω, due to possible worst-case scenarios (see Section 2.2.1). However, one can guarantee bounds on the RIP constant for F_Ω with high probability by choosing a random set Ω. The proof is quite involved so we refer the reader to [10] or [34] (the latter reference provides a proof for general bounded orthonormal systems).

Theorem 2.5. *Let $A \in \mathbb{C}^{m \times n}$ be the random partial Fourier just described. Then the restricted isometry constant of the rescaled matrix $\sqrt{\frac{n}{m}} A$ satisfies $\delta_s \leq \delta$ with probability*

at least $1 - n^{-\gamma \log^3 n}$ *provided that*

$$m \geq C\delta^{-2} s \log^4 n.$$

The constants $C, \gamma > 1$ *are universal and do not depend on the* m *or* n*.*

Theorem 2.3 says that the RIP constant of $\delta_{2s} < 1/3$ is needed to guarantee exact recovery via ℓ_1-minimization. Combining this with Theorem 2.5, exact recovery for all s-sparse $x \in \mathbb{C}^n$ is possible, with high probability, if

$$m \geq \tilde{C} s \log^4 n.$$

This is the best known result so far for random partial Fourier matrices. However, it has been observed by many researchers in practice that one can get away with far fewer measurements and still obtain exact reconstruction. In fact, it is even conjectured that the lower bound on m may in fact be $\mathcal{O}(s \log n)$ [30, 25]. But, why should one even expect $\mathcal{O}(s \log n)$ as a theoretical lower bound? The answer has to do with uncertainty principles.

2.2.1 Uncertainty Principles, Random Sampling, and Optimality

The reason that Fourier measurements work so well to reconstruct sparse signals is due to an uncertainty principle between a signal and its frequency information. Generally speaking, the more sparse the signal is, the less sparse its Fourier transform must be. Define the normalized discrete Fourier transform and its inverse as

$$\begin{aligned} \hat{f}_l &= \frac{1}{\sqrt{n}} \sum_{j=0}^{n} f_j e^{-i2\pi lj/n}, \\ f_j &= \frac{1}{\sqrt{n}} \sum_{l=0}^{n} \hat{f}_l e^{i2\pi lj/n}. \end{aligned}$$

Then the discrete uncertainty principle is

$$\text{supp}(f) \times \text{supp}(\hat{f}) \quad \geq \quad n, \tag{2.9}$$

where $\text{supp}(f) = \{\#j : f_j \neq 0\}$ (see Section B.3 for the proof of (2.9)). What does (2.9) have to do with Compressed Sensing and the optimal number of measurements?

Consider a non-zero s-sparse signal $x \in \mathbb{C}^n$ and its discrete Fourier transform $\hat{x} \in \mathbb{C}^n$. Suppose all m Fourier measurements are zero. This can happen if both the signal and its frequency information are sparse. Clearly, the sparsest solution to either the ℓ_0 or ℓ_1-minimization problem will be a signal that is zero everywhere. The discrete uncertainty principle gives us limits on the combined sparsity of a signal and its frequency. For example, if $x \in \mathbb{C}^n$ and $\|x\|_0 = n/4$, it is not possible to have $\|\hat{x}\|_0 < 4$. At worst, (2.9) allows us to have a signal that is $\sqrt{n}$-sparse in both the time and frequency domain, i.e., $\|x\|_0 = \|\hat{x}\|_0 = \sqrt{n}$. The discrete Dirac Comb signal exhibits this level of sparsity and we examine it in more detail below.

Example 2.3. *(Discrete Dirac Comb). Suppose that n is a perfect square and define a signal which consists of unit spikes at a uniform spacing of $\sqrt{n}$. This signal is invariant under the discrete Fourier transform. Figure 2.2 shows the signal (left) and its discrete Fourier transform (right).*

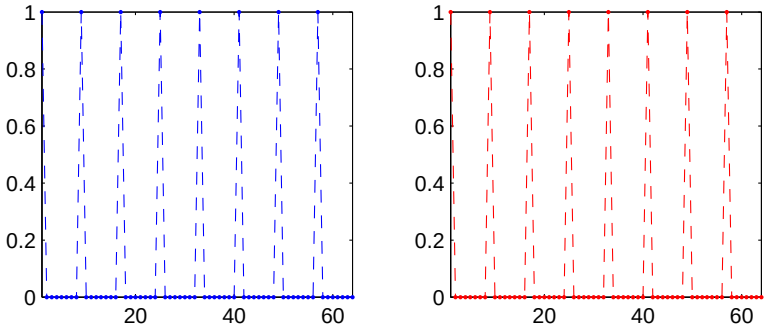

Figure 2.2: Dirac comb function on the left and its discrete Fourier transform on the right. The signal is invariant under the Fourier transform, and keeps the same sparsity.

Let us attempt to reconstruct the discrete Dirac comb signal $x \in \mathbb{R}^n$ from its Fourier

measurements $\hat{x} \in \mathbb{C}^n$. If we choose any of the $n - \sqrt{n}$ zero-valued Fourier measurements, ℓ_0 or ℓ_1-minimization will recover a signal that is identically zero. For example, if $n = 100$, one could take 90 Fourier measurements and still only recover the zero signal! To have any chance of recovering the Dirac comb signal, we need to capture at least one of the $\sqrt{n}$ non-zero frequencies. What if the frequencies are chosen randomly?

Let Ω be the set of measurements with cardinality m, chosen uniformly from the set of subsets of size $\binom{n}{m}$, as described in Section 2.2. The total number of ways we can obtain m zero-valued frequency measurements is $\binom{n-\sqrt{n}}{m}$. Define $W \triangleq \text{supp}(\hat{x})$. Then, the probability of choosing $m > \sqrt{n}$ measurements that contain zero-valued frequency information is

$$P(\Omega \cap W = \emptyset) \quad = \quad \frac{\binom{n-\sqrt{n}}{m}}{\binom{n}{m}} \quad \geq \quad \left(1 - \frac{2m}{n}\right)^{\sqrt{n}},$$

where the inequality can be shown using Stirling's approximation [4]. To ensure that $P(\Omega \cap W = \emptyset)$ is smaller than n^{-m}, i.e., that the probability of choosing zero-valued frequency information is small, it must be true that $\log P \leq -m \log n$ so that

$$\sqrt{n} \log \left(1 - \frac{2m}{n}\right) \quad \leq \quad -m \log n.$$

Assume that $m < n/2$ and use Taylor's approximation to write $\log \left(1 - \frac{2|\Omega|}{n}\right) \approx -\frac{2|\Omega|}{n}$. It follows that

$$m \quad \geq \quad \frac{1}{2} m \sqrt{n} \log n.$$

Thus, due to existence of the Dirac comb signal, any algorithm must have $m \sim sm \log n$. This example illustrates the necessity for random Fourier measurements and why one should expect lower bounds of the form $\mathcal{O}(s \log n)$.

2.2.2 RIP for 2D Fourier Measurement Matrices

Let $X \in \mathbb{C}^{N\times N}$ be a discrete image and $n = N^2$. The ℓ_p norm for $0 < p < \infty$ is defined as

$$\|X\|_p \quad = \quad \left(\sum_{i=1}^{N} \sum_{j=1}^{N} |X_{i,j}|^p \right)^{1/p} \tag{2.10}$$

For $p = 0$ let us define the quasi-norm

$$\|X\|_0 \quad = \quad \{\#(i,j) \quad : \quad X_{i,j} \neq 0\}$$

which counts the number of nonzero elements in X. Define X to be s-sparse if $\|X\|_0 = s$. One can also consider the image X as an n length vector. More precisely, consider the vectorized form $X(:) \in \mathbb{C}^n$ of X constructed by stacking the columns of X on top of each other. Note that $\|X(:)\|_0 = \{\#i \quad : \quad X(:)_i \neq 0\} = \|X\|_0$.

We define the discrete two-dimensional Fourier operation in two ways: as an operator on the image $X \in \mathbb{C}^{N\times N}$ or as an operator on $X(:) \in \mathbb{C}^n$. For the former, let $F \in \mathbb{C}^{N\times N}$ be the one-dimensional DFT matrix. Then,

$$\hat{X} \quad \triangleq \quad \mathcal{F}X \quad = \quad FXF$$

where $\mathcal{F} : \mathbb{C}^{N\times N} \to \mathbb{C}^{N\times N}$ is the two-dimensional discrete Fourier transform operator, and $\hat{X} \in C^{N\times N}$ is the two-dimensional discrete Fourier transform of X. Likewise, define the two-dimensional discrete Fourier transform operator on $X(:)$ by

$$\hat{X}(:) \quad \triangleq \quad \mathcal{F}(:)X(:) \quad = \quad (F \otimes F)X(:),$$

where $\hat{X}(:) \in \mathbb{C}^n$ is the vectorized form of $\hat{X}$, $\otimes$ denotes the Kronecker product, and $\mathcal{F}(:) = F \otimes F \in \mathbb{C}^{n\times n}$. Thus, $\mathcal{F}$ operates on X, while $\mathcal{F}(:)$ operates on $X(:)$. The reason for both forms will be clear in Section 2.3. In practice, the matrix $F \otimes F$ is never computed. Instead, one performs a two-dimensional FFT on the image X at the cost of $\mathcal{O}(n \log N)$

operations.

The two-dimensional random partial Fourier transform operator is now constructed as follows. Define the a set of m two-dimensional frequencies $w_i = (\omega_{1,i}, \omega_{2,i})$ for $1 \leq i \leq m$, chosen uniformly at random from all sets of size m, from the frequencies $\{0, \ldots, N-1\}^2$. Then, the random partial Fourier transform operator $\mathcal{F}_\Omega : \mathbb{C}^{N \times N} \to \mathbb{C}^m$ is defined as

$$(\mathcal{F}_\Omega X)_i \quad = \quad (\mathcal{F}X)_{\omega_i}, \quad 1 \leq i \leq m. \tag{2.11}$$

Alternatively, let us choose the set $\Omega' \subset \{1, \ldots, n\}$ with $|\Omega'| = m$, chosen uniformly at random from all sets of this size. Define the two-dimensional random discrete partial Fourier matrix as the m randomly chosen rows of $F \otimes F$ from the index set Ω'. Denote this sub-matrix by $\mathcal{F}(:)_{\Omega'} \in \mathbb{C}^{m \times n}$. There is a one-to-one correspondence between the indices in Ω' and Ω, since the former are the indices with respect to the vectorized form. Therefore,

$$\mathcal{F}(:)_{\Omega'} X(:) \quad = \quad \mathcal{F}_\Omega X. \tag{2.12}$$

The following is the two-dimensional version of Theorem 2.5.

Corollary 2.6. *Let $\mathcal{F}_\Omega : \mathbb{C}^{N \times N} \to \mathbb{C}^m$ be the random two-dimensional partial Fourier operator just described. Let $n = N^2$. Then, the restricted isometry constant of the rescaled operator $\sqrt{\frac{n}{m}}\mathcal{F}_\Omega$ satisfies $\delta_s \leq \delta$ with probability at least $1 - n^{-\gamma \log^3 n}$ provided that*

$$|\Omega| \quad = \quad m \quad \geq \quad C\delta^{-2} s \log^4 n.$$

The constants $C, \gamma > 1$ are universal and do not depend on the m or n.

Combining the above result with Theorem 2.3, a s-sparse image $X \in \mathbb{C}^{N \times N}$ can be recovered, with high probability, via the ℓ_1-minimization problem

$$\min_{Z \in \mathbb{C}^{N \times N}} \|Z\|_1 \quad \text{subject to} \quad \mathcal{F}_\Omega Z = \mathcal{F}_\Omega X, \tag{2.13}$$

or

$$\min_{z\in\mathbb{C}^n} \|z\|_1 \quad \text{subject to} \quad \mathcal{F}(:)_\Omega z = \mathcal{F}(:)_\Omega X(:), \tag{2.14}$$

with

$$m \quad \geq \quad \tilde{C} s \log^4 n.$$

This is summarized in the following corollary.

Corollary 2.7. *Let $\mathcal{F}_\Omega$ be the scaled random two-dimensional partial Fourier operator defined above. Then, every s-sparse image $X \in \mathbb{C}^{N\times N}$ is recoverable from measurements $\mathcal{F}_\Omega X$ with probability at least $1 - n^{-\gamma \log^3 n}$ provided that*

$$|\Omega| \quad \geq \quad \tilde{C} s \log^4 n.$$

The constants $\tilde{C}, \gamma > 1$ are universal.

2.3 Total-Variation Sparsity

This section extends the results of Theorem 2.5 and Corollary 2.7 for sparse-gradient images. The results in this section will be used in the main algorithm in Chapter 4.

2.3.1 1D Case

Define the one dimensional difference operator $D \in \mathbb{R}^{n\times n}$ on $x \in \mathbb{C}^n$ as

$$Dx \quad \triangleq \quad \begin{cases} x_j - x_{j-1} & 1 < j \leq n \\ x_1 - x_n & j = 1 \end{cases}$$

where we have adopted the convention $x_0 = x_n$. The discrete total-variation operator for a one-dimensional signal can be defined as

$$(\mathrm{TV}(x))_j \quad = \quad |(Dx)_j|,$$

and the discrete total-variation semi-norm of x is

$$\|x\|_{\mathrm{TV}} \quad \triangleq \quad \|\,\mathrm{TV}(x)\|_1 \quad = \quad \sum_{i=1}^{n} |x_{j+1} - x_j| \ .$$

Define the ℓ_1 total-variation quasi norm

$$\|\,\mathrm{TV}(x)\|_0 \quad \triangleq \quad \{\#i : |x_j - x_{j-1}| \neq 0\}.$$

We say that x is s-sparse in the total-variation or gradient sense if $\|\,\mathrm{TV}(x)\|_0 = s$. For example, if x is a piecewise constant function with s nonzero jumps, then $\|\,\mathrm{TV}(x)\|_0 = s$. Let $A \in \mathbb{C}^{m\times n}$ be the random partial Fourier matrix and $y \triangleq Ax \in \mathbb{C}^m$ be the corresponding m random discrete Fourier measurements, where $\Omega \subset \{1, \ldots, n\}$ denotes the index set of randomly selected rows of the discrete Fourier matrix. Is it possible to recover x from random Fourier measurements y, when x is s-sparse in gradient, via ℓ_1-minimization? Consider the total-variation minimization problem

$$\min_{z\in\mathbb{C}^n} \|z\|_{\mathrm{TV}} \quad \text{subject to} \quad Az = y. \tag{2.15}$$

Define the first order finite difference $\delta_j \triangleq x_{j+1} - x_j$. The discrete Fourier transform shift theorem gives

$$\hat{\delta}_j \quad = \quad (1 - e^{-2\pi i(j-1)/n})\hat{x}_j, \quad j = 1, \ldots, n,$$

where $\hat{\delta}_j$ and $\hat{x}_j$ denote the discrete Fourier transform of δ and x, i.e. $\hat{\delta} = F\delta$ and $\hat{x} = Fx$. Note that $\sum \delta_j = 0$ so we have $\hat{\delta}_1 = 0$. Define $v_j = (1 - e^{-2\pi i(j-1)/n})^{-1}$ for $1 < j \leq n$ and $v_1 = 0$. Then, the total-variation minimization problem (2.15) is equivalent to

$$\min_{z\in\mathbb{C}^n} \|z\|_1 \quad \text{subject to} \quad \hat{z}_j = v_j\hat{x}_j, \quad \text{for } j \in \Omega.$$

Because $\hat{v}_1 = 0$ we loose information at the zero frequency. Thus, one can obtain exact recovery provided that the solution z^* to (2.15) is adjusted so that $\sum z_j^* = \sum x_j$. We summarize the result in the following corollary [4, Section 1.5].

Corollary 2.8. *Let $x \in \mathbb{C}^n$ be total-variation or gradient s-sparse, where $s = \{\#i : |x_j - x_{j-1}| \neq 0\}$. Under the assumptions of Theorem 2.3, the minimizer to (2.15) is unique and equal to x with high probability.*

2.3.2 2D Case

For an image $X \in \mathbb{C}^{N \times N}$, define the horizontal and vertical difference operators as

$$(D_h X)_{i,j} \triangleq \begin{cases} X_{i,j} - X_{i-1,j} & 1 < i \leq n \\ X_{1,j} - X_{N,j} & i = 1 \end{cases}$$

$$(D_v X)_{i,j} \triangleq \begin{cases} X_{i,j} - X_{i,j-1} & 1 < j \leq n \\ X_{i,1} - X_{i,N} & j = 1 \end{cases}$$

and the discrete gradient operator

$$(DX)_{i,j} = \begin{pmatrix} (D_h X)_{i,j} \\ (D_v X)_{i,j} \end{pmatrix}.$$

For the two-dimensional case, let us introduce two types of total-variation discretizations: isotropic and anisotropic [18]. For the isotropic discretization, define the discrete total variation operator

$$(\mathrm{TV}_{\mathrm{iso}}(X))_{i,j} \triangleq \| (DX)_{i,j} \|_2 = \sqrt{|(D_h X)_{i,j}|^2 + |(D_v X)_{i,j}|^2}.$$

and

$$\|X\|_{\mathrm{TV}_{\mathrm{iso}}} \triangleq \| \mathrm{TV}_{\mathrm{iso}}(X) \|_1 = \sum_{i,j} \| (DX)_{i,j} \|_2.$$

For the anistropic total variation discretization, define

$$(\mathrm{TV}_{\mathrm{aniso}}(X))_{i,j} \triangleq \| (DX)_{i,j} \|_1 = |(D_h X)_{i,j}| + |(D_v X)_{i,j}|.$$

and

$$\|X\|_{\mathrm{TV}_{\mathrm{aniso}}} \triangleq \| \mathrm{TV}_{\mathrm{aniso}}(X)\|_1 = \|D_h X\|_1 + \|D_v X\|_1.$$

We say that X is s-sparse in the gradient or total variation sense if $\| \mathrm{TV}_{\mathrm{iso}}(X)\|_0 = s$ or $\| \mathrm{TV}_{\mathrm{aniso}}(X)\|_0 = s$, with respect to the isotropic or anisotropic discretizations, respectively.[4]

The ℓ_1-minimization problem for the isotropic and anisotropic total-variation discretization is defined as follows. First, consider an image $X \in \mathbb{C}^{N\times N}$ that is s-sparse in the isotropic total-variation or gradient sense, and m random Fourier measurements $y \triangleq \mathcal{F}_\Omega X \in \mathbb{C}^m$, where $m = |\Omega|$ and Ω is the index set of randomly selected Fourier measurements. The isotropic total-variation minimization problem is

$$\min_{Z\in\mathbb{C}^{N\times N}} \|Z\|_{\mathrm{TV}_{\mathrm{iso}}} \quad \text{subject to} \quad \mathcal{F}_\Omega Z = y. \tag{2.16}$$

Similarly, for $X \in \mathbb{C}^{N\times N}$ that is s-sparse in the anisotropic total-variation or gradient sense, the anisotropic total-variation minimization problem is

$$\min_{Z\in\mathbb{C}^{N\times N}} \|Z\|_{\mathrm{TV}_{\mathrm{aniso}}} \quad \text{subject to} \quad \mathcal{F}_\Omega Z = y. \tag{2.17}$$

Exact recovery for gradient sparse images can now be stated as follows.

Theorem 2.9. *Let $X \in \mathbb{R}^{N\times N}$ be s-sparse in the isotropic total-variation or gradient sense, i.e., $\|X\|_{\mathrm{TV}_{\mathrm{iso}}} = s$, and let $n = N^2, m = |\Omega|$. Then, for*

$$m \geq \tilde{C} s \log^4 n,$$

[4]Note that $\| \mathrm{TV}_{\mathrm{aniso}}(X)\|_0 \geq \| \mathrm{TV}_{\mathrm{iso}}(X)\|_0$.

the solution to (2.16) is unique and equal to X with probability $1-n^{-\gamma \log^3 n}$, where $\tilde{C}, \gamma > 1$ are universal constants.

Proof. The proof is similar to the one-dimensional case studied in Section 2.3.1. Define $\delta_h \triangleq D_h X$ and $\delta_v \triangleq D_v X$ to be the horizontal and vertical differences, then $\mathcal{F}_\Omega \delta_h$ and $\mathcal{F}_\Omega \delta_v$ can be determined by a simple linear transformation of the measurements $\mathcal{F}_\Omega X$. Now, let us define $\delta \triangleq \delta_h + i\delta_v$ where $i^2 = -1$. Then the isotropic total variation minimization problem can be restated as

$$\min_{Z \in \mathbb{C}^{N \times N}} \|Z\|_1 \quad \text{subject to} \quad \mathcal{F}_\Omega Z = \mathcal{F}_\Omega \delta.$$

□

For the anistropic recovery, we can prove the analogous result.

Theorem 2.10. *Let $X \in \mathbb{C}^{N \times N}$ be s-sparse in the anisotropic total-variation or gradient sense, i.e., $\|X\|_{\mathrm{TV}_{\mathrm{aniso}}} = s$, and let $n = N^2, m = |\Omega|$. Then, for*

$$m \quad \geq \quad \tilde{C} s \log^4 n,$$

the solution to (2.17) is unique and equal to X with probability $1-n^{-\gamma \log^3 n}$, where $\tilde{C}, \gamma > 1$ are universal constants.

Proof. Again, define $\delta_h \triangleq D_h X$ and $\delta_v \triangleq D_v X$ to be the horizontal and vertical differences, respectively. Again, $\mathcal{F}_\Omega \delta_h$ and $\mathcal{F}_\Omega \delta_v$ can be determined by a simple linear transformation of the measurements $\mathcal{F}_\Omega X$ using the discrete Fourier transform shift theorem. Now, define $\delta(:) \in \mathbb{C}^{2n}$ as

$$\delta(:) \quad = \quad \begin{pmatrix} \delta_h(:) \\ \delta_v(:) \end{pmatrix},$$

i.e., $\delta(:)$ is the vectorized form of the concatenation of δ_h and δ_v. Then, the random Fourier

measurements, $\mathcal{F}_\Omega \delta_h$ and $\mathcal{F}_\Omega \delta_v$, can be written in matrix-vector form as

$$\tilde{\mathcal{F}}_\Omega(:)\delta(:),$$

where $\tilde{\mathcal{F}}_\Omega(:) \in \mathbb{C}^{2n \times 2n}$ is a block diagonal matrix given by

$$\tilde{\mathcal{F}}_\Omega(:) \triangleq \begin{pmatrix} \mathcal{F}_\Omega(:) & \mathbf{0} \\ \mathbf{0} & \mathcal{F}_\Omega(:) \end{pmatrix}.$$

(2.17) can now be re-written as

$$\min_{Z \in \mathbb{C}^{2n}} \|Z\|_1 \quad \text{subject to} \quad \tilde{\mathcal{F}}_\Omega(:) Z = \tilde{y}.$$

The only thing left to show is that the RIP constant for $\tilde{\mathcal{F}}_\Omega(:)$ is strictly less than $1/3$ for $m \geq \tilde{C} s \log^4 n$. We know that

$$(1 - \delta_{2s})\|x\|_2^2 \quad \leq \quad \|\mathcal{F}_\Omega(:)x\|_2^2 \quad \leq \quad (1 + \delta_{2s})\|x\|_2^2$$

for all $2s$-sparse $x \in \mathbb{C}^n$ with $\delta_{2s} < 1/3$. Then,

$$(1 - \delta_{2s})(\|x_1\|_2^2 + \|x_2\|_2^2) \quad \leq \quad \|\mathcal{F}_\Omega(:)x_1\|_2^2 + \|\mathcal{F}_\Omega(:)x_2\|_2^2 \quad \leq \quad (1 + \delta_{2s})(\|x_1\|_2^2 + \|x_2\|_2^2)$$

is true for any $x_1, x_2 \in \mathbb{C}^n$ such that $\|x_1\|_0 + \|x_2\|_0 \leq 2s$ (both vectors combined are $2s$-sparse). The above inequality is equivalent to

$$(1 - \delta_{2s})\|x\|_2^2 \quad \leq \quad \|\tilde{\mathcal{F}}_\Omega(:)x\|_2^2 \quad \leq \quad (1 + \delta_{2s})\|x\|_2^2,$$

which means that the RIP constant for $\tilde{\mathcal{F}}_\Omega(:)$ satisfies $\delta_{2s} < 1/3$. □

Results for total-variation reconstruction under noisy measurements are analogous to Theorem 2.4 by the same reasoning used in the above proofs.

2.4 Numerical Methods

This section reviews common numerical methods for solving the ℓ_1-minimization problem (2.3). Section 2.4.1 shows how one can transform (2.3) and (2.6) into a linear program or a second-order cone program, respectively. Section 2.4.2 introduces the least-squares regularization method to solve the ℓ_1 constrained minimization problem (2.3) or (2.6), which is the primary numerical method used in this work.

2.4.1 LPs and SOCPs

The ℓ_1-minimization problem

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{subject to} \quad Az = y \tag{2.18}$$

in the real case is equivalent to the linear program (LP)

$$\min_{z,v\in\mathbb{R}^n} \sum_{j=1}^{n} v_i \quad \text{subject to} \quad \begin{cases} v_i \geq z_i, & 1 \leq i \leq n, \\ v_i \geq -z_i, & 1 \leq i \leq n, \\ Az = y, \end{cases}$$

where the constraints $v_i \geq z_i$ and $v_i \geq -z_i$ for $i = 1, \ldots, n$ are equivalent to $\|v\|_1 \geq \|z\|_1$.

For the noisy ℓ_1-minimization problem

$$\min_{z\in\mathbb{R}^n} \|z\|_1 \quad \text{subject to} \quad \|Ax - y\|_2 < \varepsilon \tag{2.19}$$

we can recast this problem as a second order cone program (SOCP)

$$\min_{z,t\in\mathbb{R}^n} \sum_{j=1}^{n} t_i \quad \text{subject to} \quad \begin{cases} t_i \geq z_i, & 1 \leq i \leq n \\ t_i \geq -z_i, & 1 \leq i \leq n \\ \|Az - y\|_2^2 - \varepsilon^2 < 0, \end{cases}$$

where $\|Az - y\|_2^2 - \varepsilon^2 < 0$ is a second order conic constraint. LPs and SOCPs have been studied extensively in the literature and there are various algorithms, e.g., simplex or log barrier, that can be used to solve these problems (see [3] or [2, Section 4.4.2]).

Remark 2.2. *In order to perform minimization over complex vector-valued variables, one can divide the problem into the real and imaginary parts, effectively doubling the dimension of the problem. An equivalent, but much less intrusive approach is discussed in [20, Chapter 3].*

2.4.2 ℓ_1-regularization

In this work, the preferred method of solving the constrained ℓ_1-minimization problem (2.18) and (2.19) is to use a Lagrange multiplier type method. For $\lambda > 0$ let us consider the ℓ_1-regularized least squares functional

$$\Lambda_\lambda(x) \quad \triangleq \quad \|Ax - y\|_2^2 + \lambda\|x\|_1 \tag{2.20}$$

and let x_λ^* be the minimizer of $\Lambda_\lambda(x)$. It can be shown that the limit of x_λ^* as $\lambda \to 0$ is equal to the minimizer x^* of (2.18). This result is summarized in the proposition below. For a detailed proof, see Section B.5.

Proposition 2.3. *Assume that (2.18) has a unique solution x^*. Then,*

$$\lim_{\lambda \to 0} x_\lambda^* \quad = \quad x^*.$$

More generally, if the minimizer of (2.18) is not unique, then the x_λ^* are bounded and every limit point of x_λ is a minimizer of (2.18).

In practice, due to noisy measurements, the choice of λ is not so simple. One option is to solve (2.20) for different values of λ and choose the x_λ^* that minimizes the measurement error, i.e., $\|Ax_\lambda^* - y\|_2$. However, this can be computationally expensive. A less expensive,

but similar approach, is known as cross-validation. In this approach, one still solves (2.20) multiple times, but each sub-problem is of lower-dimension (see Section B.4 for more details).

Chapter Three

Fourier Edge Detection

The purpose of this chapter is to introduce the reader to the theory and numerical techniques of edge detection from Fourier measurement data. These techniques will be useful in extracting the relevant edge information from the reference image mentioned in Chapter 1. The ideas in this chapter are based on a series of papers by Gelb and Tadmor [14, 13, 16]. Section 3.1 - 3.3 provides the theoretical framework for edge detection and Section 3.4 discusses the numerical implementation necessary for our main algorithm.

3.1 Conjugate Fourier Sum

Define the conjugate Fourier partial sum as

$$\tilde{S}_N[f](x) \triangleq \sum_{k=1}^{N} a_k \sin kx - b_k \cos kx \tag{3.1}$$

where the Fourier coefficients are given by

$$\begin{Bmatrix} a_k \\ b_k \end{Bmatrix} = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \begin{Bmatrix} \cos kt \\ \sin kt \end{Bmatrix} dt.$$

We can also write this sum in terms of complex exponentials

$$\tilde{S}_N[f](x) \triangleq \sum_{k=-N}^{N} i\,\mathrm{sgn}(k) \hat{f}_k e^{ikx} \tag{3.2}$$

where $i^2 = -1$ and the Fourier coefficients are

$$\hat{f}_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx \ .$$

This latter formulation will be convenient for the numerical implementation introduced in Section 3.4. The conjugate partial Fourier sum can be also be written as

$$\tilde{S}_N[f](x) = f * \frac{1}{\pi} \tilde{D}_N(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t) \tilde{D}_N(x-t) dt \tag{3.3}$$

where $\tilde{D}_N$ is the conjugate Dirichlet kernel

$$\tilde{D}_N(t) \quad = \quad \sum_{k=1}^{N} \sin kt \quad = \quad \frac{\cos(t/2) - \cos(N+1/2)t}{2\sin(t/2)}. \tag{3.4}$$

The equivalence is easily shown by direct substitution. The following theorem tells us that the conjugate partial Fourier sum converges to the jumps of $f(x)$ [14, Theorem 2.1].

Theorem 3.1. *Let $f(x)$ be a 2π-periodic piece-wise smooth function with a single discontinuity at $x = \xi$. Then,*

$$f * \frac{-1}{\log N}\tilde{D}_N(x) \quad \overset{N\to\infty}{\to} \quad [f](x)\delta_\xi(x) \quad \hat{=} \quad \begin{cases} [f](\xi), & x = \xi, \\ 0, & \text{else}\,. \end{cases}$$

or equivalently

$$-\frac{\pi}{\log N}\tilde{S}_N[f](x) \quad \overset{N\to\infty}{\to} \quad [f](x)\delta_\xi(x),$$

More precisely,

$$-\frac{\pi}{\log N}\tilde{S}_N[f](x) \quad = \quad [f](x)\delta_\xi(x) + \mathcal{O}\left(\frac{1}{\log N}\right)$$

for each $x \in [-\pi, \pi]$ where the convergence is point-wise.

Example 3.1. *Consider the saw tooth function defined*

$$\varphi_\xi(x) \quad = \quad \begin{cases} -\dfrac{\pi + x}{2}, & -\pi \le x < \xi \\ -\dfrac{x - \pi}{2}, & \xi \le x \le \pi. \end{cases}$$

This is a piece-wise smooth function with a single discontinuity at $x = \xi$. The conjugate Fourier partial sum is then

$$\tilde{S}_N[\varphi_\xi](x) \quad = \quad -\sum_{k=1}^{N} \frac{\cos k(x-\xi)}{k}.$$

We illustrate the scaled conjugate Fourier partial sum $-\dfrac{\pi}{\log N}\tilde{S}_N[f](x)$ for different values of N.

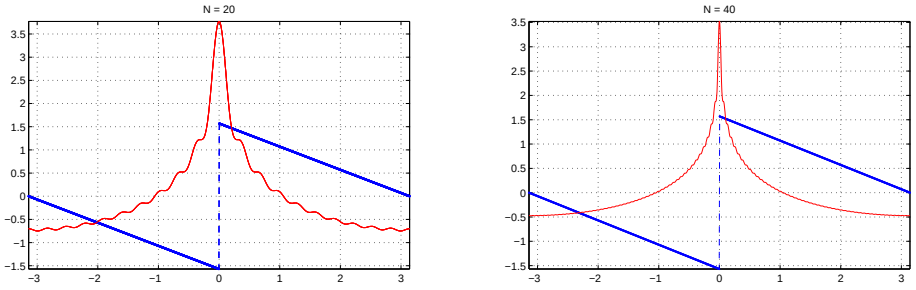

Figure 3.1: Conjugate partial Fourier sum for the saw tooth function for $N = 20$ and 100.

In general, the conjugate Dirichlet kernel $\tilde{D}_N(t)$ is not the only function that, when convolved with $f(x)$, exhibits the property described in Theorem 3.1. We introduce a class of conjugate kernels that, when convolved with the piece-wise smooth function with a finitely many jumps, can be used to detect edges [14, Definition 2.1].

Definition 3.1. *(Admissible Kernels) A conjugate kernel $\tilde{K}_N(t)$ is admissible if it satisfies the following four properties:*

1. $\tilde{K}_N(t)$ *is odd;*
2. $\lim_{N\to\infty} \int_0^\pi \tilde{K}_N(t)dt = -1$*;*
3. $\tilde{K}_N(t) = C \cdot \dfrac{\cos(N+1/2)t}{2\pi \sin(t/2)} + \tilde{R}_N(x)$*, where* $\|\tilde{R}_N\|_{L^1} \doteq \int_{-\pi}^{\pi} |\tilde{R}_N(t)|dt \leq$ *Const.;*
4. $\lim_{N\to\infty} \sup_{|t|>\delta>0} |\tilde{R}_N(t)| = 0$ *for all fixed* $\delta > 0$.

Now we will introduce a generalization to Theorem 3.1 that applies to all admissible kernels and for functions that are piece-wise smooth with finitely many jump discontinuities.

Theorem 3.2. *Let $f(x)$ be a piecewise smooth function with finitely many jump discontinuities and let $J = \{\xi\}$ denote the set of jump discontinuities. Also, assume the function $f(x)$ is Lipschitz continuous on each of the smooth pieces. Consider the generalized conjugate partial sum*

$$\tilde{S}_N^K[f] \doteq f * \tilde{K}_N = \int_{-\pi}^{\pi} f(t)\tilde{K}_N(x-t)dt$$

where $\tilde{K}_N$ is an admissible kernel satisfying properties in Definition 3.1. Then,

$$\tilde{S}_N^K[f](x) \quad \to \quad [f](x)\delta_\xi(x) \quad \triangleq \quad \begin{cases} [f](\xi), & x = \xi \in J \\ 0, & \text{else}\,. \end{cases}$$

More precisely,

$$\tilde{S}_N^K[f](x) \quad = \quad [f](x)\delta_\xi(x) + \mathcal{O}\left(\frac{1}{\log N}\right)$$

for each x so that the convergence is point-wise.

In the previous Theorem, we showed that $f * \dfrac{-1}{\log N}\tilde{D}_N(x)$ converges to the jump function of $f(x)$. Let us show that $\tilde{K}_N = \dfrac{-1}{\log N}\tilde{D}_N(x)$ is indeed an admissible kernel, as per Definition 3.1. First, clearly $\tilde{K}_N$ is odd since it is the sum of finitely many odd functions. Second, the integral $\displaystyle\int_0^\pi \tilde{K}_N(t)dt$ has an explicit form, for which the limit as $N \to \infty$ is -1.[1] Lastly, let $C \equiv 0$ and $\widetilde{R}_N(t) \triangleq \dfrac{-1}{\log N}\tilde{D}_N(x)$ in part (3) of Definition 3.1. Then,

$$\begin{aligned} \frac{-1}{\log N}\int_{-\pi}^{\pi} |\tilde{D}_N(t)|dt \quad &\leq \quad \frac{-2}{\log N}\left(\int_0^{2/N} |\tilde{D}_N(t)|dt + \int_{2/N}^{\pi} |\tilde{D}_N(t)|dt\right) \\ &\leq \quad \text{Const} \end{aligned}$$

where $|\tilde{D}_N(t)| < N$ when $t < 2N^{-1}$ and $|\tilde{D}_N(t)| < 2|t|^{-1}$ when $t > 2N^{-1}$ using the estimate $|\tilde{D}_N(t)| \leq \min\left(N, 2|t|^{-1}\right)$. For part (4), note that for a fixed δ and $N > 2\delta^{-1}$ we have $|\tilde{D}_N(t)| < 2|t|^{-1}$ so that

$$\frac{1}{\log N}|\tilde{D}_N(t)| \quad \leq \quad \frac{2}{|t|\log N}.$$

For the general class of admissible kernels, the convergence rate is still only $\mathcal{O}(1/\log N)$. In the next section we discuss how to accelerate the convergence by adding a concentration

[1] We have that $\displaystyle\int_0^\pi \sin ktdt = (1 - \cos \pi k)/k = 2/k$ for k odd and zero elsewhere.

factor to the conjugate partial sum.

3.2 Concentration Factors

3.2.1 Intuition

When $f(x)$ is piecewise smooth, it is well known that the decay of the Fourier coefficients suffers [19, Chapter 9] as compared to the case of a smooth function. In fact, without the existence of even the first derivative, we cannot guarantee that the Fourier coefficients satisfy a decay rate of $\mathcal{O}(k^{-1})$, i.e., $\hat{f}_k \propto k^{-1}$.[2] Moreover, the global rate of convergence scales poorly, i.e., $\|f(x) - \sum_{k=-N/2}^{N/2} \hat{f}_k e^{ikx}\|_{L^2[0,2\pi]} = \mathcal{O}(N^{-1/2})$. This leads one to believe that the higher modes or Fourier coefficients may provide some information about the discontinuities of the functions. However, the Fourier basis is a global one, and the Fourier coefficients do not provide local information (there is no spatial dependence on the Fourier coefficients).

Consider again a piecewise smooth function $f(x)$ with a single discontinuity at $x = \xi$. Let us write the conjugate Fourier sum expressed in terms of the complex Fourier basis

$$\tilde{S}_N[f](x) \quad = \quad -\sum_{k=-N}^{N} i\,\mathrm{sgn}(k)\hat{f}_k e^{ikx}.$$

Using the idea that the higher Fourier modes may contain relevant edge information, let us introduce a *concentration function* $\sigma(\eta), \eta \in (0,1)$, to the conjugate partial Fourier sum,

$$\tilde{S}_N^{\sigma}[f](x) \quad = \quad \sum_{k=-N}^{N} i\sigma\left(\frac{|k|}{N}\right)\mathrm{sgn}(k)\hat{f}_k e^{ikx}$$

where $\sigma(\eta)$ is non-decreasing for larger values of η, i.e., η closer to 1. For example, let

[2] Although analysis cannot guarantee that $\hat{f}_k \propto k^{-1}$, the Fourier coefficients of the sawtooth function actually satisfies this decay property (see [19, Example 9.2]).

$\sigma(\eta) = \eta$ so that the value of $\sigma(\eta)$ grows linearly as a function of η. Thus, $\sigma(\eta)$ puts more weight on the higher Fourier modes. It turns out that this is exactly what we need in order to accelerate the conjugate partial Fourier sum so that convergence is faster than $\mathcal{O}(1/\log N)$. More generally, for an arbitrary twice differentiable function $\sigma(\eta), \eta \in [0,1]$, that is nondecreasing and satisfies $\lim_{N\to\infty} \frac{1}{\pi} \int_{1/N}^{1} \frac{\sigma(x)}{x} dx = -1$, $\tilde{S}_N^{\sigma}[f](x)$ will converge to the jumps of $f(x)$ at a rate of at least $\mathcal{O}(N^{-1})$. In fact, one can obtain even faster convergence by using a class of polynomial concentration functions of the form $\sigma_r(\eta) = -\pi\eta^r$.

3.2.2 Generalized Conjugate Sum

Let $f(x)$ be a piecewise smooth function with a single discontinuity at $x = \xi$ as in the previous sections. We introduce the concept of a *generalized* conjugate partial sum

$$\tilde{S}_N^{\sigma}[f](x) \quad \triangleq \quad \sum_{k=1}^{N} \sigma_{k,N} \left(a_k \sin kx - b_k \cos kx \right), \tag{3.5}$$

where $\sigma(\eta), \eta \in (0,1)$ denotes the concentration function and $\sigma_{k,N} \triangleq \sigma(k/N)$ for $k = 1, \ldots, N$. In complex exponential form we have

$$\tilde{S}_N^{\sigma}[f](x) \quad = \quad \sum_{k=-N}^{N} i\sigma\left(\frac{2|k|}{N}\right) \mathrm{sgn}(k) \hat{f}_k e^{ikx} \tag{3.6}$$

where $i^2 = -1$.[3] We say that $\sigma(\eta)$ has a *concentration property* if

$$\tilde{S}_N^{\sigma}[f](x) \quad \overset{N\to\infty}{\to} \quad [f](x)\delta_{\xi}(x).$$

From Theorem 3.1 it is clear that $\sigma_{k,N} \triangleq -\pi/\log N$ is a concentration factor that satisfies the concentration property. In this section we will explore alternative concentration factors which exhibit faster convergence to the jump function $[f](x)$. It will be convenient to write (3.5) as

$$\tilde{S}_N^{\sigma}[f](x) \quad = \quad f * \tilde{K}_N^{\sigma}(x)$$

[3] In the complex exponential case, define $\sigma_{k,N} \triangleq (|k|/N)$ where case when $k = -N, \ldots, N$.

where $\tilde{K}_N^\sigma(t)$ is called the *generalized* conjugate kernel,

$$\tilde{K}_N^\sigma \triangleq \frac{1}{\pi}\sum_{k=1}^{N}\sigma\left(\frac{k}{N}\right)\sin kt.$$

The main result for the convergence of the generalized conjugate sum is presented below [14].

Theorem 3.3. *Consider the generalized conjugate kernel $\tilde{K}_N^\sigma$ defined above. Let $\sigma(\eta)$ be $C^2[0,1]$, i.e. a twice differentiable function on $[0,1]$ with continuous second derivative, such that $|\sigma\left(N^{-1}\right)| \leq C_0/\log N$ where $C_0 \in \mathbb{R}^+$ is some positive constant. Assume that*

$$\begin{aligned}\lim_{N\to\infty}\frac{1}{\pi}\int_{1/N}^{1}\frac{\sigma_N(\eta)}{\eta}d\eta &= -1,\\ \lim_{N\to\infty}\sum_{j=1}^{N}\frac{|\sigma(\frac{j}{N})|}{j^2} &= 0.\end{aligned}$$

Then, $\tilde{K}_N^\sigma(t)$ is an admissible kernel (see Definition 3.1) so that $\tilde{S}_N^\sigma[f](x) = f * \tilde{K}_N^\sigma$ satisfies the concentration property, i.e.

$$\tilde{S}_N^\sigma[f](x) \overset{N\to\infty}{\to} [f](x)\delta_\xi(x).$$

Moreover, if f is a piecewise $C^2[0,2\pi]$ function, then

$$\frac{1}{\log N}\tilde{S}_N^K[f](x) = [f](x)\delta_\xi(x) + \mathcal{O}\left(\frac{1}{N}\right).$$

The result is analogous to piecewise smooth functions with finitely many jumps. It is also possible to design concentration factors that exhibit faster convergence than $\mathcal{O}(N^{-1})$. In this paper, we will focus on one class of concentration factors that are of the polynomial type given by

$$\sigma^r(r) \triangleq -r\pi x^r.$$

It is easy to see that $\sigma^r(x)$ for $r \in \mathbb{N}, r \geq 1$ is admissible by Theorem 3.3. Moreover, using this class of concentration factors, one can obtain convergence of order $\mathcal{O}(N^{-r})$ [14].

Several other types of concentration factors are studied in [16].

3.3 Two-dimensional Edge Detection

The one-dimensional edge detection methods can be easily applied for a two-dimensional 2π-periodic function $f(x,y) : [0,2\pi]^2 \mapsto \mathbb{R}$ using a line-by-line approach. We start by considering the two-dimensional Fourier coefficients

$$\hat{f}_{k_x,k_y} = \frac{1}{4\pi^2}\int_{-\pi}^{\pi}\int_{-\pi}^{\pi} f(x,y)e^{-ik_x x}e^{-ik_y y}dxdy$$

for $k_x, k_y = -N,\ldots,N$. In order to apply the edge detection methods, one needs to determine the one-dimensional Fourier coefficients. The one-dimensional Fourier coefficients can computed be for each fixed slice in the orthogonal Cartesian directions as

$$\tilde{f}_{k_x}(y_\lambda) = \sum_{k_y=-N}^{N} \hat{f}_{k_x,k_y}e^{ik_y y_\lambda}, \qquad \text{and} \quad \tilde{f}_{k_y}(x_\nu) = \sum_{k_x=-N}^{N} \hat{f}_{k_x,k_y}e^{ik_x x_\nu},$$

where $y_\lambda = -\pi + \frac{\pi\lambda}{N}$ and $x_\nu = -\pi + \frac{\pi\nu}{N}$ with $\lambda,\nu = 0,1,\ldots,2N$. Then, $\{\tilde{f}_{k_x}(y_\lambda)\}_{k_x=-N}^{N}$ are the Fourier coefficients of $f(x,y_\lambda)$ along the x-direction, and $\{\tilde{f}_{k_y}(x_\nu)\}_{l=-N}^{N}$ are the Fourier coefficients of $f(x_\nu,y)$ along the y-direction. Now, the generalized conjugate Fourier sum can be applied to $f(x,y_\lambda)$, giving us edge data along the x-direction, and $f(x_\nu,y)$, giving us edge data along the y-direction. In summary, we can compute

$$\tilde{S}_N^\sigma[f](x,y_\lambda) \quad \text{and} \quad \tilde{S}_N^\sigma[f](x_\nu,y),$$

for $\lambda,\nu = 0,1,\ldots,2N$.

Remark 3.1. *If one has Fourier coefficient information, it may be tempting to reconstruct the function $f(x,y)$ via the Fourier partial sum and then approximate the edge information directly from the reconstructed image. However, determining the edges directly from the raw Fourier data is more cost efficient, and may also be more accurate, since reconstruction*

algorithms may introduce spurious artifacts and inaccuracies (e.g. Gibbs phenomenon, aliasing error, etc.) [16, p.32].

3.4 Numerical Implementation

This section introduces the discrete analogue to the generalized conjugate Fourier sum $\tilde{S}_N^\sigma$. The formulation for the discrete setting is almost identical to the continuous setting, except for the addition of a filter, which reduces spurious oscillations.

3.4.1 Adding a Filter

Let $f(x)$ be a piecewise smooth function with two jump discontinuities

$$f(x) \quad = \quad \begin{cases} \dfrac{1}{2}\cos 3x + \dfrac{1}{2}, & -\pi \le x < -\dfrac{\pi}{2}, \\ \dfrac{4}{1+2x^2} - 2, & -\dfrac{\pi}{2} \le x \le \dfrac{\pi}{2}, \\ \sin x \cos \dfrac{7x}{2}, & \dfrac{\pi}{2} \le x < \pi. \end{cases}$$

Consider the original un-accelerated concentration function $\sigma_1(x) \equiv -\pi/\log N$, and first-order polynomial concentration factor $\sigma_2(x) = -\pi x$. We can compute the generalized Fourier sum via the complex exponential form (3.6)

$$\tilde{S}_N^\sigma[f](x) \quad = \quad \frac{1}{\log N} \sum_{k=-N}^{N} i\sigma\left(\frac{|k|}{N}\right) \operatorname{sgn}(k) \hat{f}_k e^{ikx} \tag{3.7}$$

since the sum is easily computed by the Fast Fourier Transform (FFT). Figure 3.2 shows the performance of both concentration factors, $\sigma_1(x)$ on the left and $\sigma_2(x)$ on the right.

It is clear that the generalized conjugate partial sum using the polynomial concentration factor $\sigma_2(x)$ approximates the jump function more accurately than the original un-accelerated concentration function $\sigma_1(x)$. However, the concentration function $\sigma_2(x)$

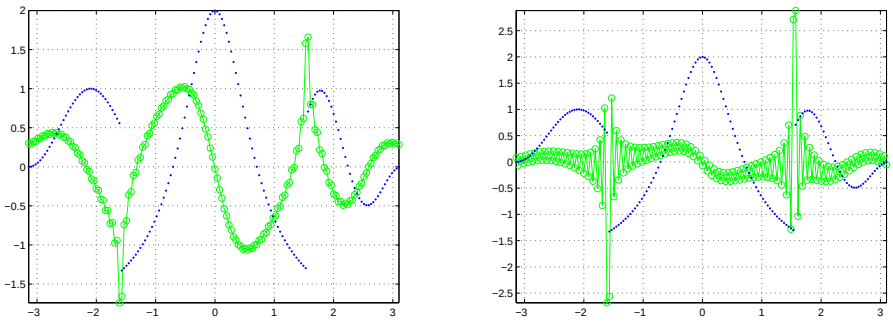

Figure 3.2: (Left) $f(x)$ (blue), and $\tilde{S}_N^{\sigma_1}[f](x)$ with $\sigma_1(x)$ (green). (Right)$f(x)$ (blue), and $\tilde{S}_N^{\sigma_2}[f](x)$ with $\sigma_2(x)$ (green).

introduces a new concern: spurious oscillations near the jumps. Fortunately, these oscillations can be removed by the addition of a filter on the Fourier coefficients [16, Section 3]. The filter essentially dampens the Fourier coefficients at the higher frequencies. In this work, we will use an exponential filter of the form

$$\sigma^{filt}(\eta) \quad = \quad \begin{cases} 1 & |\eta| \leq \eta_c \\ \exp\left(-\alpha\left(\dfrac{\eta-\eta_c}{1-\eta_c}\right)^p\right) & \eta > \eta_c \end{cases} ,$$

where $\eta \in (0,1]$, α is a measure of the how strong the higher frequencies should be filtered, and p is the order of the filter (for different types of filters see [19, Chapter 9]). With the addition of a filter, (3.7) becomes

$$\tilde{S}_N^{\sigma}[f](x) = \frac{1}{\log N}\sum_{k=-N}^{N} i\,\mathrm{sgn}(k)\sigma\left(\frac{|k|}{N}\right)\sigma^{filt}\left(\frac{|k|}{N}\right)\hat{f}_k e^{ikx} \quad . \tag{3.8}$$

Figure 3.3 shows the generalized conjugate partial sum after the exponential filter has been added using (3.8) with $\alpha = 5, p = 5$, and $\eta_c = .15$. Notice that the oscillations are virtually eliminated.[4]

[4]See [15] for a more detailed approach on handing discrete data. Note that a factor of $e^{i\pi k}$ must be added to the discrete Fourier transform coefficients calculated via the FFT.

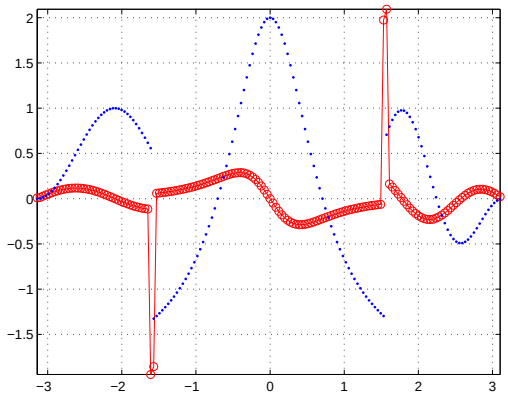

Figure 3.3: Generalized conjugate partial sum with exponential filter.

3.4.2 Edge Detection Operator in 1D

In this section we provide the matrix-vector analogue to (3.8) using the discrete Fourier transform (DFT) matrix. This will be necessary in order to define the edge measurement matrix. Consider a signal $x \in \mathbb{C}^N$ and its discrete Fourier transform $\hat{x} = Fx \in \mathbb{C}^N$, computed using the one-dimensional discrete Fourier matrix $F \in \mathbb{C}^{N\times N}$, where the entries are given by

$$F_{l,k} \quad = \quad \frac{1}{\sqrt{N}} e^{2\pi i(l-1)(k-1)/N}, \quad l,k = 1,\ldots,N.$$

(3.8) applied to our discrete Fourier coefficients $\hat{x} \in \mathbb{C}^N$ gives us

$$\tilde{S}_N^\sigma[x]_l \doteq \frac{1}{\log(N/2)} \sum_{k=-N/2+1}^{N/2} i\,\mathrm{sgn}(k)\sigma\left(\frac{2|k|}{N}\right)\sigma^{filt}\left(\frac{2|k|}{N}\right)\hat{x}_k \frac{e^{2\pi ikl}}{\sqrt{N}}, \tag{3.9}$$

for $l = 0,\ldots,N-1$. Define the vector $K \doteq (-N/2+1,\ldots,N/2)^T \in \mathbb{R}^N$ and a shifted version $\tilde{K} \doteq (0,\ldots,N/2,-N/2+1,\ldots,-1)^T \in \mathbb{R}^N$. Furthermore, define a diagonal matrix $\Lambda \in \mathbb{C}^{N\times N}$ with

$$\Lambda_{j,j} \quad \doteq \quad \frac{1}{\log(N/2)} i\,\mathrm{sgn}(\tilde{K}_j)\sigma\left(\frac{2|\tilde{K}_j|}{N}\right)\sigma^{filt}\left(\frac{2|\tilde{K}_j|}{N}\right), \quad \text{for } j = 1,\ldots,N\ .$$

Then, (3.9) can be written as

$$\tilde{S}_N^{\sigma}[x]_l \;=\; (\hat{E}\hat{x})_l, \quad l = 1, \ldots, N,$$

where the one-dimension edge detection operator $\hat{E} \in \mathbb{C}^{N\times N}$ is defined as

$$\hat{E} \;\triangleq\; F^{-1}\Lambda \tag{3.10}$$

Note that we can also write the matrix vector product $\tilde{S}_N^{\sigma}[x]_l$ as Ex where $E \triangleq F^{-1}\Lambda F$.

3.4.3 Edge Detection Operator in 2D

Consider an image $X \in \mathbb{C}^{N\times N}$ and its two-dimensional Fourier transform $\hat{X} \in \mathbb{C}^{N\times N}$. In order to extract the edge information from $\hat{X}$, we need to determine the one-dimensional Fourier transform information along the columns and rows, respectively. This can be done easily by pre-multiplying or post-multiplying $\hat{X}$ by the one-dimensional inverse DFT matrix, after which we can apply the discrete operator $\hat{E}$, defined in the previous section. This gives us the two-dimensional edge detection operator $\hat{\mathcal{E}} : \mathbb{C}^{N\times N} \to \mathbb{C}^{2N\times N}$

$$\hat{\mathcal{E}}\hat{X} \;\triangleq\; \begin{pmatrix} \hat{E}\hat{X}F^{-1} \\ F^{-1}\hat{X}\hat{E}^T \end{pmatrix}. \tag{3.11}$$

The edge detection matrix $\hat{\mathcal{E}}\hat{X}$ contains jump information along each fixed column, i.e., $\hat{E}\hat{X}F^{-1}$, and along each fixed row, i.e., $F^{-1}\hat{X}\hat{E}^T$.

In practice, one does not require the edge information at all points across the columns and rows of X. For this reason, we introduce the concept of the restricted edge operator, which contains only relevant jump information. In order to determine the relevant edge data, define a threshold parameter $\tau > 0$ and determine the indices of $\hat{\mathcal{E}}\hat{X}$ for which the individual pixel values in $\hat{\mathcal{E}}\hat{X}$surpass this threshold. The threshold parameter τ can be empirically determined so that it is small enough to capture the most significant jumps.

We summarize the steps to construct a restricted edge operator below.

1. Empirically estimate $\tau > 0$ so that τ is small enough to capture the relevant jumps.

2. After computing $\hat{\mathcal{E}}\hat{X} \in \mathbb{C}^{2N \times N}$, determine the indices for which the pixel values are larger than τ, in absolute magnitude. More precisely, let

$$\Omega'_\tau \;\stackrel{\circ}{=}\; \left\{ \begin{array}{lll} (i,j) & : & |(\hat{\mathcal{E}}\hat{X})_{i,j}— \geq \tau, \\ & & 1 \leq i \leq 2N, \quad 1 \leq j \leq N \end{array} \right\}$$

   where $m_\tau \stackrel{\circ}{=} |\Omega'_\tau|$. Define Ω_τ to be set of m_τ pairs $(i_k, j_k) \in \Omega'_\tau$ for $1 \leq k \leq m_\tau$.

3. Finally, define the restricted edge operator $\hat{\mathcal{E}}_{\Omega_\tau} : \mathbb{C}^{N \times N} \to \mathbb{C}^{m_\tau}$ where

$$(\hat{\mathcal{E}}_{\Omega_\tau}\hat{X})_k \;=\; (\hat{\mathcal{E}}\hat{X})_{i_k,j_k}, \quad 1 \leq k \leq m_\tau.$$

The edge operator $\hat{\mathcal{E}}_{\Omega_\tau}$ will be a critical component of our main algorithm, introduced in the next chapter.

Chapter Four

Algorithm and Examples

4.1 General Setup

Consider an image $X \in \mathbb{C}^{N_1 \times N_1}$ that is s-sparse in gradient, with respect to the anisotropic total-variation discretization[1], as discussed in Chapter 2, and Fourier transform information, $\hat{Y} \in \mathbb{C}^{N_2 \times N_2}$, from a reference image $Y \in \mathbb{C}^{N_2 \times N_2}$, where $N_2 \geq N_1$.[2] Recall from Section 2.3 that X is recoverable from random Fourier measurements $\mathcal{F}_\Omega X$ with high probability provided that $|\Omega| \geq \mathcal{O}(s \log^4 N_1^2)$ via the total-variation minimization problem (2.17). It is conjectured that this lower bound may be reduced to $\mathcal{O}(s \log N_1^2)$ and, in fact, many numerical tests show that accurate reconstruction can be obtained with only $3s$ to $5s$ measurements [30].

Let the edge information in Y refer to both the indices and magnitudes for which $\mathrm{TV}_{\mathrm{aniso}}(Y)$ is nonzero. This edge information can be approximated from the Fourier transform information $\hat{Y}$ using the edge detection operator $\hat{\mathcal{E}}$ introduced at the end of the previous chapter. In this chapter, we will provide a new algorithm for the reconstruction of X from Fourier measurements $\mathcal{F}_\Omega X$ and edge information $\hat{\mathcal{E}}\hat{Y}$, under the assumption that both X and Y share similar edge information. By combining both Fourier and edge measurements, we will show that the number of Fourier measurements required for accurate and even exact reconstruction is significantly lower than CS total-variation minimization techniques alone without edge information. We also introduce a modified algorithm for inexact edge information and noisy measurements.

4.2 The Main Algorithm

When $N_2 = N_1$ one can apply the information in $\hat{\mathcal{E}}\hat{Y}$ to X directly since the dimensions of the two images are the same. Consider the restricted edge operator $\hat{\mathcal{E}}_{\Omega_\tau}$ and edge

[1] From here on we will only use the anisotropic total-variation discretization, but analogous results apply to the isotropic formulation.

[2] In general, we will consider a discrete time series of images $\{X(t)\}_{t=1}^{T}, T \in \mathbb{N}$, but for simplicity the algorithm and setup will only be presented for a single image at single time point. One can then apply the algorithm for each image in the time series analogously.

measurements $y_e \triangleq \hat{\mathcal{E}}_{\Omega_\tau}\hat{Y} \in \mathbb{C}^{m_\tau}$ (see the end of Section 3.11). Then, for $\lambda, \gamma > 0$ we solve the ℓ_1-regularization problem

$$\min_{Z\in\mathbb{C}^{N\times N}} \lambda\|Z\|_{\mathrm{TV}_{\mathrm{aniso}}} + \|\mathcal{F}_\Omega Z - y\|_2^2 + \gamma\|\hat{\mathcal{E}}_{\Omega_\tau}\mathcal{F}Z - y_e\|_2^2 \tag{4.1}$$

where $\gamma > 0$ is a parameter that penalizes or restricts the edge information of desired solution. The penalization parameters, λ and γ, can be determined via a cross-validation approach (see B.4).

When $N_2 > N_1$, one cannot directly apply the information in $\hat{\mathcal{E}}\hat{Y}$ to X because $\hat{\mathcal{E}}\hat{Y}$ provides edge information on a higher resolution image than our image X. A simple solution to this problem is to perform a bilinear interpolation on $\hat{\mathcal{E}}\hat{Y} \in \mathbb{C}^{2N_2\times N_2}$ to obtain edge information on the lower resolution grid (see B.6). Let $\mathcal{I}_N(\hat{\mathcal{E}}\hat{Y}) \in \mathbb{C}^{2N_1\times N_1}$ be the interpolated low-resolution edge information. Then, the restricted edge operator can be computed by replacing $\hat{\mathcal{E}}\hat{Y}$ with $\mathcal{I}_N(\hat{\mathcal{E}}\hat{Y})$ in Step 2 of the algorithm presented in Section 3.11. Denote the corresponding edge measurements by $\tilde{y}_e \in \mathbb{C}^{m_\tau}$. In this case, because the interpolation introduces inaccuracies in the edge information, we introduce a modification of (4.1).

The modified algorithm is based on the EdgeCS (Edge Guided Compressed Sensing) algorithm presented in [18]. In EdgeCS, the idea is to introduce iteratively adapted weights to the total variation ℓ_1 term in (4.1). These weights correspond to the jumps of the recovered image. More precisely, the weights are set to zero wherever the most significant jumps are detected. This removes the insistence of sparsity, which is controled by the ℓ_1 term, at the locations we know are not sparse. The modified algorithm is as follows.

1. Set the iteration number $r = 1$, the weights $w_i = 1$ for $i = 1, \ldots, n$, and let $X^{(0)}$ be the initial guess.

2. Solve the weighted anisotropic total variation minimization problem with weights w_i

and initial guess $X^{(0)}$:

$$X^{(r)}(w) = \arg \min_{Z \in \mathbb{C}^{N_1 \times N_1}} \quad \lambda \sum_{i=1}^{n} w_i \left(|(D_h Z)_i| + |(D_v Z)_i| \right) + \|\mathcal{F}_\Omega X - y\|_2^2 + \gamma \|\mathcal{E}_{\Omega_\tau} X - \tilde{y}_e\|_2^2 \tag{4.2}$$

3. Detect the jumps of the reconstructed image $X^{(r)}(w)$:

$$I^{(r)} \quad \triangleq \quad \{i : |(D_h Z)_i| + |(D_v Z)_i| > 2^{-k} \max \{|(D_h Z)_i| + |(D_v Z)_i|)\Big\}, i = 1, \ldots, N\}$$

   where $I^{(r)}$ is the index set of the detected jumps.

4. Update the weights: $w_i = 0$ for all $i \in I^{(r)}$ and $w_j = 1$ for all $j \notin I^{(r)}$.

5. Set $X^{(0)} = X^{(r)}(w)$, $r = r + 1$ and $\gamma = 0$ and repeat Steps 2 - 5 until the stopping criteria are met.

For $r = 1$, the solution $X^{(r)}$ is equivalent to the solution of (4.1). For $r = 2$, the modified algorithm sets the weights to zero (Step 4 of the previous iteration) wherever the jumps, which exceed half the magnitude of the largest jump, are detected (Step 3 of the previous iteration), and (4.2) is solved once again. The process repeats until all the significant jumps are correctly detected. For the numerical examples in this work, four to six iterations of the modified algorithm are sufficient for exact or stable reconstruction.

In order to solve the ℓ_1-regularization problem (4.1) and (4.2), we use a non-linear conjugate gradient method, similar to [23, Section 4.5.2,] but with gradient restarts [27, Section 5.2] (see B.7 for the precise algorithm). Because the ℓ_1-norm is non-differentiable at zero, we approximate the ℓ_p-norm (2.10) with

$$\|X\|_1 \quad \approx \quad \sum_{i=1}^{N} \sum_{j=1}^{N} \sqrt{X_{i,j} + \mu} \quad .$$

where μ is a smoothing parameter. The choice of μ can vary, but numerical experiments show that $\mu \in [10^{-12}, 10^{-10}]$ works fairly well. Also, $\lambda \approx \gamma \approx 10^{-2}$ seems to work well for

the choice of regularization parameters. While these parameters were determined experimentally, a cross-validation approach is suggested in B.4.

4.3 Numerical Examples

In this section we present three sets of numerical experiments. The first set illustrates the reconstruction of a time series from incomplete Fourier measurements and edge information using (4.1). Uniform, Gaussian, and radial sampling patterns are used, and the reference image provides us with exact edge locations for the times series. The next set of experiments show reconstruction of the same time series, but with inexact edge measurements and a Gaussian sampling pattern for the Fourier measurements, using the modified algorithm. Noise is also introduced in to the measurements. Finally, the third set of experiments examines the reconstruction for a different time series under a Gaussian sampling pattern, inexact edge information, and noisy measurements. In this particular time series, very small changes are occurring between time points, which more accurately simulates an actual fMRI study. The inexact edge information examples correspond to the case in which the reference image dataset is acquired at higher spatial resolution than the time series dataset.In this last set of experiments, we also show results for three dimensional image reconstruction.

4.3.1 Reconstruction with Exact Edge Data

Example Setup

The first set of experiments examines the reconstruction of time series images $\{X(t)\}_{t=1}^{4}$, $X(t) \in \mathbb{C}^{64\times 64}$, from incomplete Fourier measurements and prior edge information from the Fourier transform information $\hat{Y} \in \mathbb{C}^{64\times 64}$ of a reference image $Y \in \mathbb{C}^{64\times 64}$. The reference image contains the same edge or jump locations as the time series. The time

series images are Shepp-Logan phantom images with ellipse intensities varying by 1%. Figure 4.1 shows the reference image and Figure 4.2 shows the time series images we are trying to reconstruct.

Figure 4.1: Reference image $Y \in \mathbb{C}^{64\times 64}$. We are given the two-dimensional Fourier transform information of this image, $\hat{Y}$.

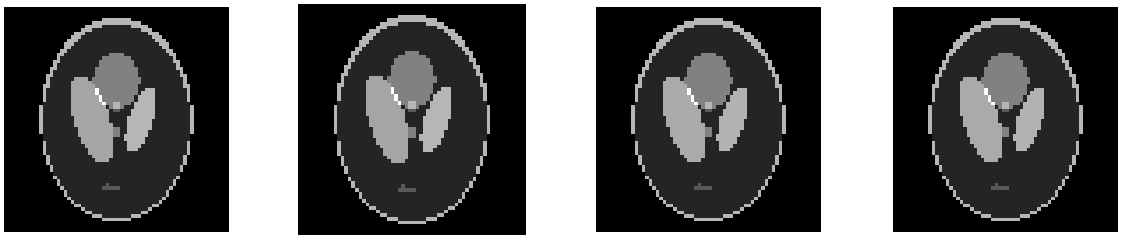

Figure 4.2: $\{X(t)\}_{t=1}^{4}$, $X(t) \in \mathbb{C}^{64\times 64}$, $t = 1, \ldots, 4$ from left to right with varying ellipse intensities.

The reference image, Y, and $X(t)$ are the same except for the ellipse intensities of the two center ellipses, $e_l(t), e_r(t) \in \mathbb{R}$. The table below compares the values of the two center ellipses between the reference image Y, and the times series images $X(t)$.

	$e_l(t)$	$e_r(t)$
X_1	9.0	10.0
X_2	9.1	9.9
X_3	9.2	9.8
X_4	9.3	9.7
Y	12	12

Table 4.1: Ellipse intensities, $e_l(t)$ and $e_r(t)$, for the time series images

In these examples, the magnitude of the jumps in the reference image are within 20% to 30% of the exact images $X(t)$. Figure 4.3 shows a cross sectional view of the reference image versus the image at the first time point, $X(1)$. Both have similar edge information, i.e. the same location of edges, but different jump magnitudes.

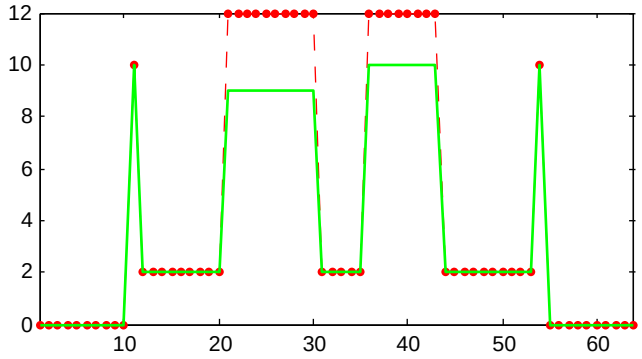

Figure 4.3: Cross sectional view of the two center ellipses of the reference image Y (shown in red) vs $X(1)$, shown in blue.

The goal of the algorithm will be to reconstruct $X(t)$ from a small number of Fourier measurements and edge information from $\hat{Y}$. Reconstruction of $\{X(t)\}_{t=1}^4$ using (4.1) is presented below under different sampling patterns for the Fourier measurements, i.e. different choices of Ω. Results indicate that a sampling pattern that places a higher number of samples near the zero frequency work extremely well.[3] In all the examples, we take $\lambda = \gamma = .01$, $\tau = .1$ and solve (4.1).

Uniform Sampling

Let us take m uniformly sampled Fourier measurements from the time series images $\{X(t)\}_{t=1}^4$ shown in Figure 4.2. Figure 4.4 shows the sampling pattern with $m = 410$, or roughly 10% of the Fourier coefficients of the image. In an actual MRI acquisition such sampling can be approximated using a non-Cartesian Fourier space trajectory.

By utilizing the edge information, we can obtain near exact reconstruction. Figure 4.5 compares the reconstruction with and without the edge data for all images in the times series $\{X(t)\}_{t=1}^4$, for $t = 1, \ldots, 4$ viewed along the center cross-section with the 10% uniformly sampled Fourier coefficients and edge data, using $\lambda = \gamma = .01, \tau = .1$. Figure 4.6 shows a close up of the reconstruction in Figure 4.5 (left) where the ellipse intensities are changing, and Figure 4.7 shows the same reconstruction shown in Figure 4.5 (left), but

[3] This is not surprising since most of the signals energy, i.e. in the ℓ_2 sense, is contained in the lower spatial frequencies.

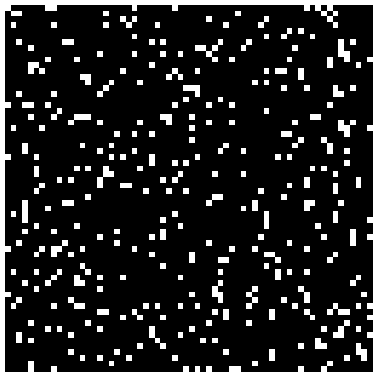

Figure 4.4: Fourier domain sampling pattern with uniform sampling. White pixels denote the indices of the Fourier measurement samples.

on individual plots.

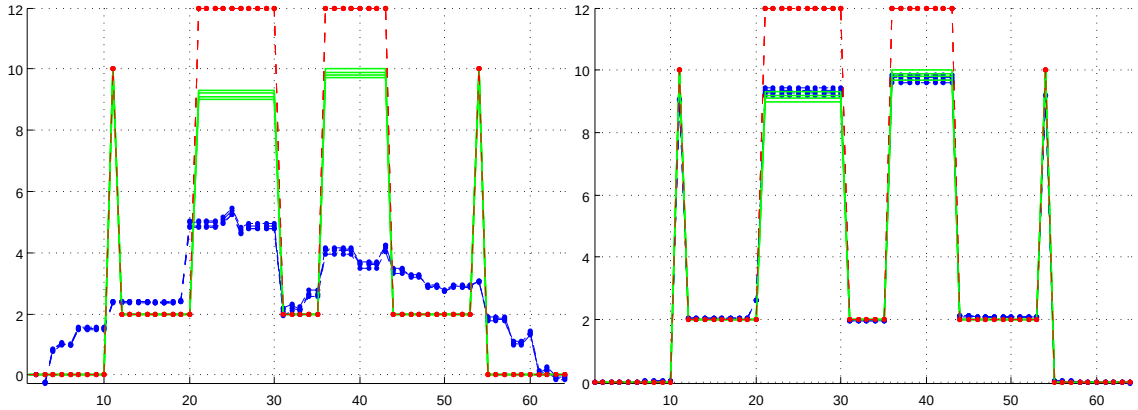

Figure 4.5: (Left) Reconstruction along the center cross-section with the same 10% uniformly sampled Fourier coefficients and no edge data, i.e. $\gamma = 0$. (Right) Reconstruction of the times series $\{X(t)\}_{t=1}^{4}$viewed along the center cross-section with the 10% uniformly sampled Fourier coefficients and edge data, with $\lambda = \gamma = .01, \tau = .1$. The blue dotted line represents the reconstruction of the time series, the green line represents the true values, $\{X(t)\}_{t=1}^{4}$, and the red line represents the values for the reference image Y.

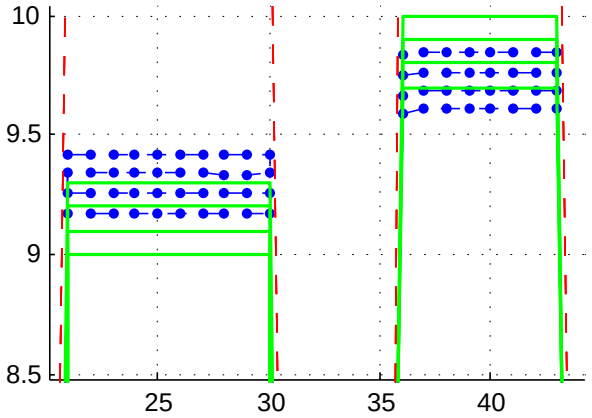

Figure 4.6: Close up of Figure 4.5 (right) at the location of the changing ellipse intensities.

Experiments indicate that one would require roughly 30% uniformly sampled Fourier

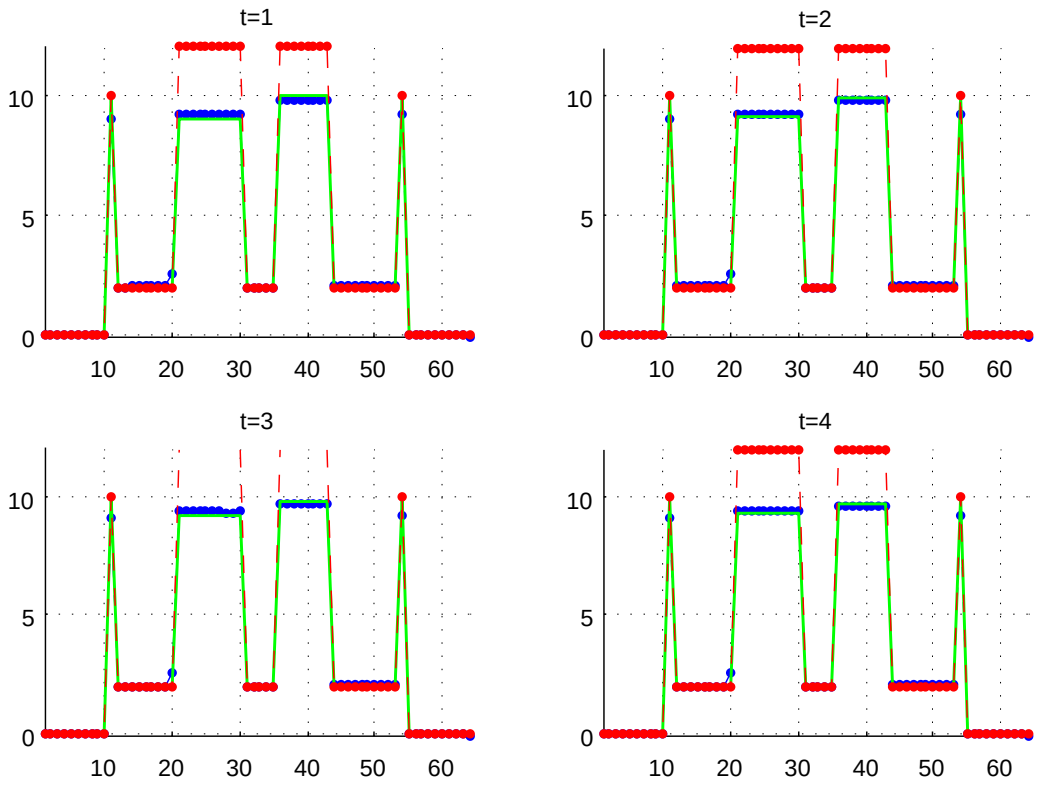

Figure 4.7: Figure 4.5 (right) on separate plots for $t = 1, \ldots, 4$.

measurements to obtain near exact reconstruction using total-variation techniques alone, without any edge data. This means that an overall reduction by a factor of 3.3, compared with full Fourier measurement acquisition, can be achieved using total-variation techniques without edge data. In comparison, using our approach, one can achieve an overall reduction by a factor of 10, compared with full Fourier measurement acquisition. The additional factor of 3 reduction comes from using the edge information. Thus, if the goal of an fMRI experiment is to find sites of activation where exact knowledge of the "ideal" intensity change is of secondary importance, then reductions well below 30% are feasible.

Let $X_1^*(t)$ be the solution to (4.1) using the 10% uniformly sampled Fourier measurements and edge information, $X_2^*(t)$ be the solution to (4.1) using the same 10% Fourier measurements and no edge data, and $X_3^*(t)$ be the solution to (4.1) using 30% uniformly sampled Fourier measurements and no edge data. Table 4.2 shows the relative errors between the exact solution and the time series reconstructions $X_i^*(t)$, $i = 1, 2$ and 3.

t	RelErr($X_1^*(t)$)	RelErr($X_2^*(t)$)	RelErr($X_3^*(t)$)
1	2.38%	53.47%	0.17%
2	2.24%	53.54%	0.16%
3	2.11%	52.34%	0.16%
4	1.99%	53.38%	0.16%

Table 4.2: Comparison of relative errors between the exact time series and the recovered images. The relative error is defined as $\text{RelErr}(X) := \|X^{\text{exact}}(t) - X\|_2 / \|X_{\text{exact}}(t)\|$, where $X_{\text{exact}}(t)$ is the exact solution at time t. $\text{RelErr}(X_1^*)$ is the reconstruction error with edge data, $\text{RelErr}(X_2^*)$ is the reconstruction error without edge data, and $\text{RelErr}(X_3^*)$ is reconstruction without edge data with more Fourier samples

Gaussian Sampling

Next, let us show reconstruction results with m Gaussian sampled Fourier coefficients taken from the sequence of images $\{X(t)\}_{t=1}^4$ shown in Figure 4.2. Figure 4.8 shows the sampling pattern with $m = 320$, or roughly 8% of the Fourier coefficients of the image.

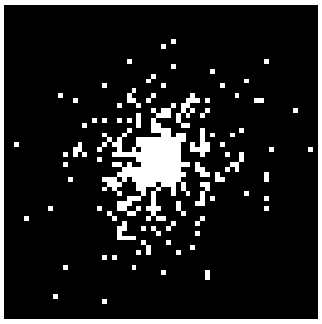

Figure 4.8: Fourier domain sampling pattern with Gaussian sampling. White pixels denote the indices of the Fourier measurement samples.

Figure 4.9 compares the reconstruction with and without the edge data. Figure 4.10 shows a close up of the reconstruction in Figure 4.9 (right), where the ellipse intensities are changing, and Figure 4.11 shows the same reconstruction shown in Figure 4.9 (right), but on individual plots.

Experiments indicate that one would require roughly 20% Gaussian sampled Fourier measurements, using the same density, to obtain near exact reconstruction using total-variation techniques alone, without any edge data. This means that an overall reduction by a factor of 5, compared with full Fourier measurement acquisition, can be achieved using total-variation techniques without edge data. In comparison, using our approach, one can

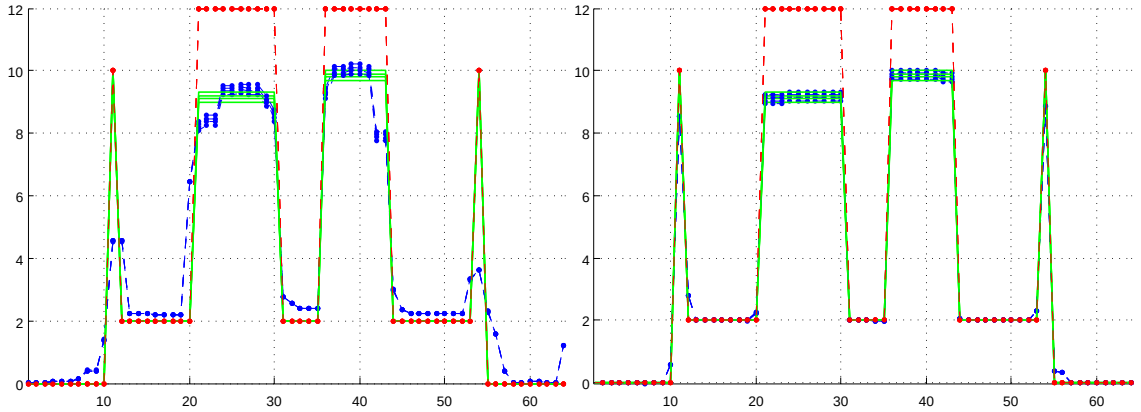

Figure 4.9: (Left) Reconstruction along the center cross-section with 8% Gaussian sampled Fourier coefficients and no edge data, i.e. $\gamma = 0$. (Right) Reconstruction viewed along the center cross-section with 8% Gaussian sampled Fourier coefficients and edge data, i.e. $\lambda = \gamma = .01, \tau = .1$. The blue dotted line represents the reconstruction, the green line represents the true values, $\{X(t)\}_{t=1}^{4}$, and the red line represents the values for the reference image Y.

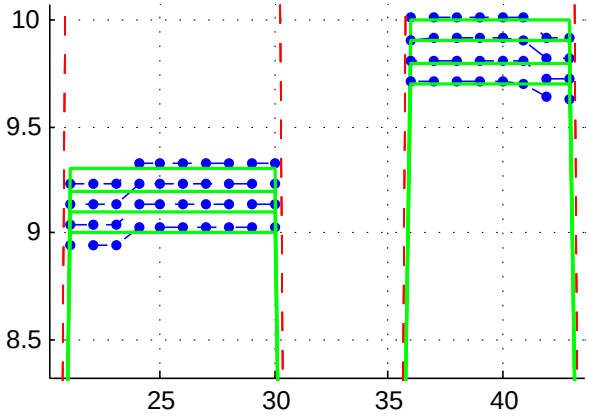

Figure 4.10: Close up of the left illustration in Figure 4.9 (left) at the location of the changing ellipse intensities.

achieve an overall reduction by a factor of 12.5, compared with full Fourier measurement acquisition. The additional factor of 2.5 reduction comes from using the edge information.

Let $X_1^*(t)$ be the solution to (4.1) using the 8% Gaussian sampled Fourier measurements and edge information, $X_2^*(t)$ be the solution to (4.1) using the same 8% Fourier measurements and no edge data, and $X_3^*(t)$ be the solution to (4.1) using 20% Gaussian sampled Fourier measurements and no edge data. Table 4.3 shows the relative errors between the exact solution and the time series reconstructions $X_i^*(t)$, for $i = 1, 2$ and 3.

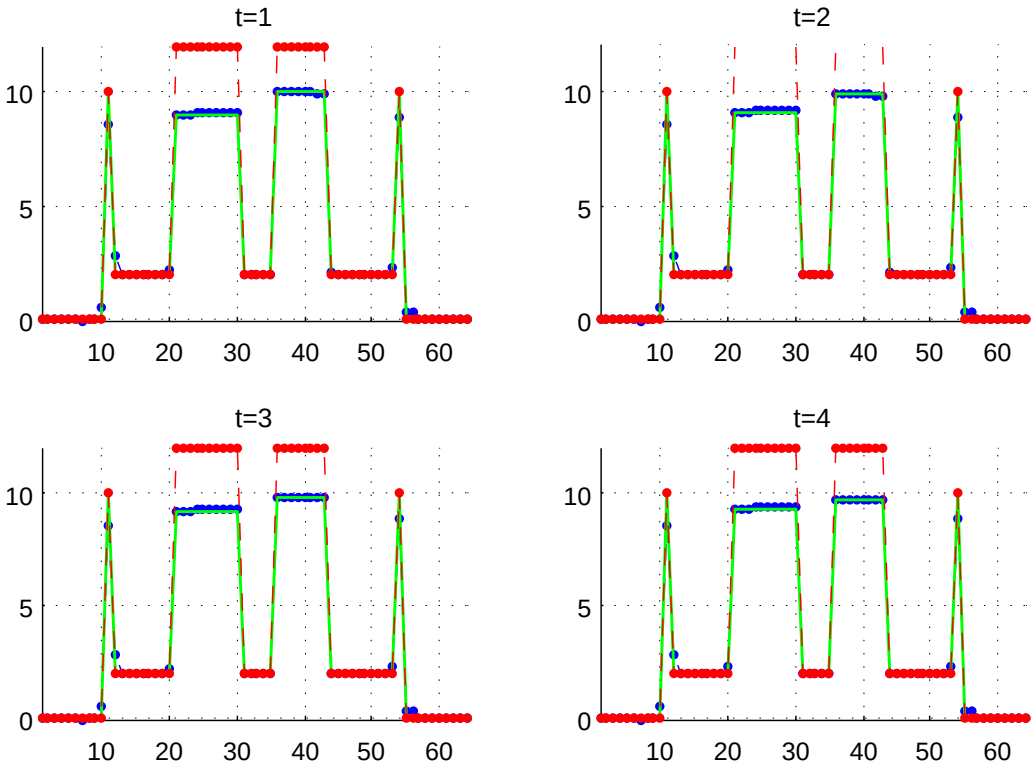

Figure 4.11: Figure 4.9 (left) on separate plots for $t = 1, \ldots, 4$.

t	RelErr$(X_1^*(t))$	RelErr$(X_2^*(t))$	RelErr$(X_3^*(t))$
1	0.72%	31.29%	0.14%
2	0.71%	31.16%	0.14%
3	0.70%	31.25%	0.15%
4	0.69%	31.26%	0.14%

Table 4.3: Comparison of relative errors between the exact time series and the recovered images. The relative error is defined as RelErr$(X) := \|X^{\text{exact}}(t) - X\|_2 / \|X_{\text{exact}}(t)\|$, where $X_{\text{exact}}(t)$ is the exact solution at time t (same as Table 4.2)

Radial Sampling

Finally, we show reconstruction results with m radially sampled Fourier coefficients taken from the sequence of images $\{X(t)\}_{t=1}^4$ shown in Figure 4.2. Figure 4.12 shows the sampling pattern 5 radial lines with $m = 336$, or roughly 8% of the Fourier coefficients of the image.

Figure 4.13 compares the reconstruction with and without the edge data. Figure 4.14 shows a close up of the reconstruction in Figure 4.13 (right), where the ellipse intensities are changing, and Figure 4.15 shows the same reconstruction shown in Figure 4.13 (right),

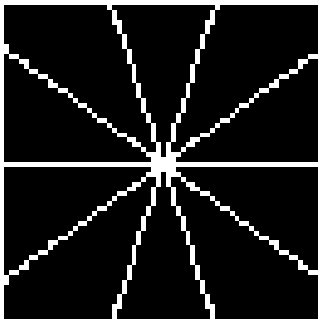

Figure 4.12: Fourier domain sampling pattern with radial sampling. White pixels denote the indices of the Fourier measurement samples.

but on individual plots.

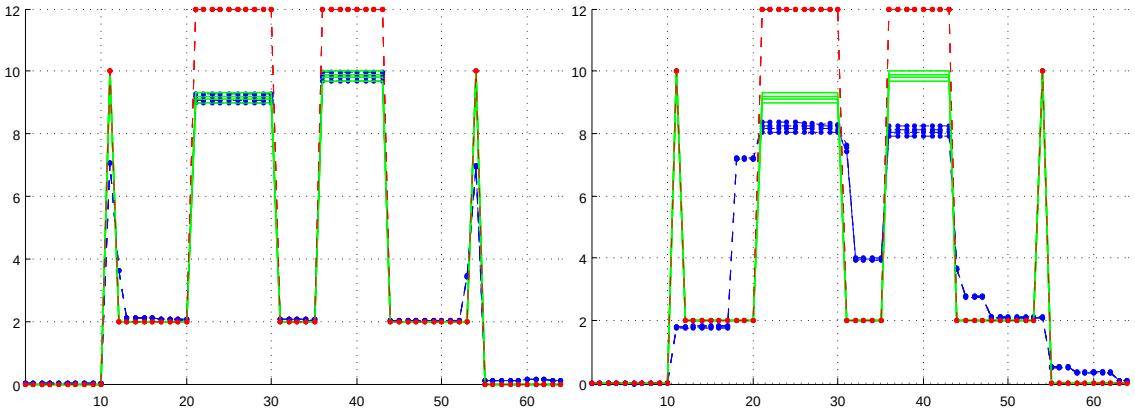

Figure 4.13: (Left) Reconstruction viewed along the center cross-section with 8% radial sampled Fourier coefficients and edge data, i.e. $\lambda = \gamma = .01, \tau = .1$. (Right) Reconstruction along the center cross-section with 8% radial sampled Fourier coefficients and no edge data, i.e. $\gamma = 0$. The blue dotted line represents the reconstruction, the green line represents the true values, $\{X(t)\}_{t=1}^{4}$, and the red line represents the values for the reference image Y.

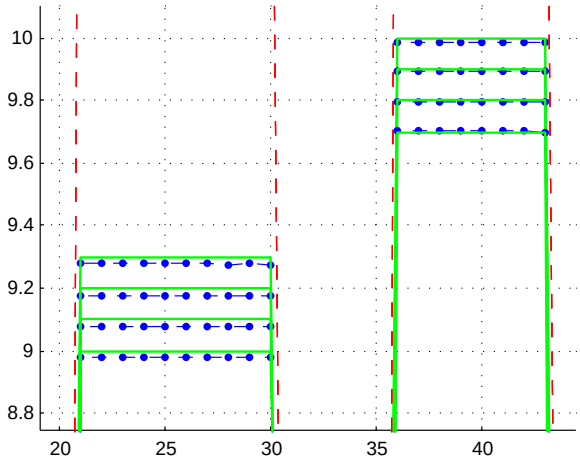

Figure 4.14: Close up of the left illustration in Figure 4.13 (left) at the location of the changing ellipse intensities.

Experiments indicate that one would require roughly 15 radial lines or roughly 23%

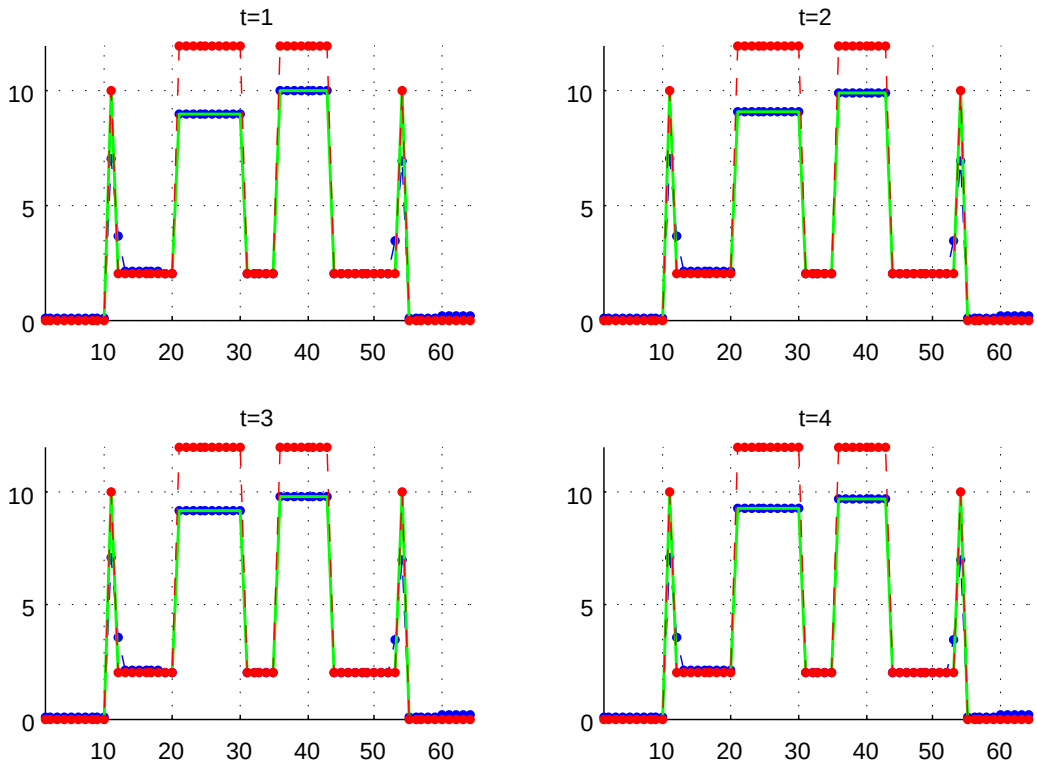

Figure 4.15: Figure 4.13 (left) on separate plots for $t = 1, \ldots, 4$.

Fourier measurements to obtain near exact reconstruction using total-variation techniques alone, without any edge data. This means that an overall reduction by a factor of about 4.3, compared with full Fourier measurement acquisition, can be achieved using total-variation techniques without edge data. In comparison, using our approach, one can achieve an overall reduction by a factor of 12.5, compared with full Fourier measurement acquisition. The additional factor of 3 reduction comes from using the edge information.

Let $X_1^*(t)$ be the solution to (4.1) using the 8% radially sampled Fourier measurements and edge information, $X_2^*(t)$ be the solution to (4.1) using the same 8% Fourier measurements and no edge data, and $X_3^*(t)$ be the solution to (4.1) using 23% radially sampled Fourier measurements and no edge data. Table 4.4 shows the relative errors between the exact solution and the time series reconstructions $X_i^*(t)$, for $i = 1, 2$ and 3.

t	RelErr$(X_1^*(t))$	RelErr$(X_2^*(t))$	RelErr$(X_3^*(t))$
1	0.68%	14.10%	0.19%
2	0.71%	14.08%	0.20%
3	0.75%	14.06%	0.20%
4	0.80%	14.04%	0.20%

Table 4.4: Comparison of relative errors between the exact time series and the recovered images. Relative error is defined as $\text{RelErr}(X) := \|X^{\text{exact}}(t) - X\|_2 / \|X_{\text{exact}}(t)\|$, where $X_{\text{exact}}(t)$ is the exact solution at time t (same as previous tables).

4.3.2 Reconstruction with High Resolution Reference Image

Noiseless Measurements

Consider the reconstruction of the same time series $\{X(t)\}_{t=1}^4$ as the the first set of tests, but with Fourier transform information $\hat{Y} \in \mathbb{C}^{256\times256}$ from a high resolution reference image $\hat{Y} \in \mathbb{C}^{256\times256}$. In order to use the edge information from $\hat{Y}$ for our algorithm, we perform a bilinear interpolation on $\hat{\mathcal{E}}\hat{Y}$. As stated in Section 4.2, this introduces inexact edge information. Figure 4.16 shows the reference image that corresponds to edge information of the interpolant versus the true edge information of the time series images, along a cross-section. The location of the edges for the two center ellipses are not exact anymore. Figure 4.17 shows the Gaussian sampling pattern with $m = 484$, or roughly 12% of the Fourier coefficients of the image. Figure 4.18 shows the reconstruction under the 12% Gaussian sampling pattern with the modified algorithm.

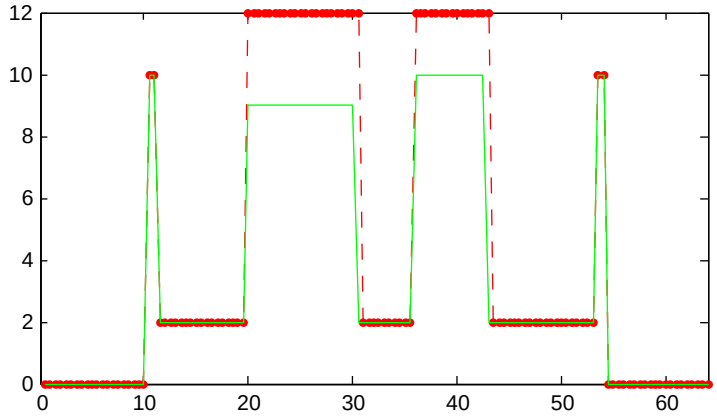

Figure 4.16: Red shows edge information of the reference image, determined by the interpolations on the edge data, i.e. the interpolant $\mathcal{I}_N(\mathcal{E}_H\hat{Y})$, and the green shows the exact edge information for the time series images. The edges are not aligned anymore as in Figure 4.3.

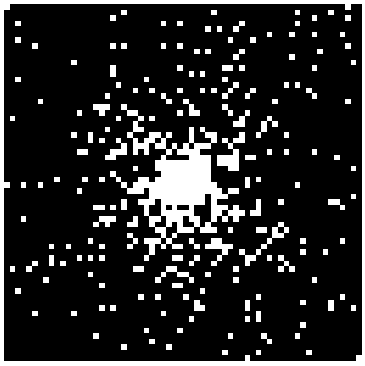

Figure 4.17: Fourier domain sampling pattern with Gaussian sampling. White pixels denote the indices of the Fourier measurement samples.

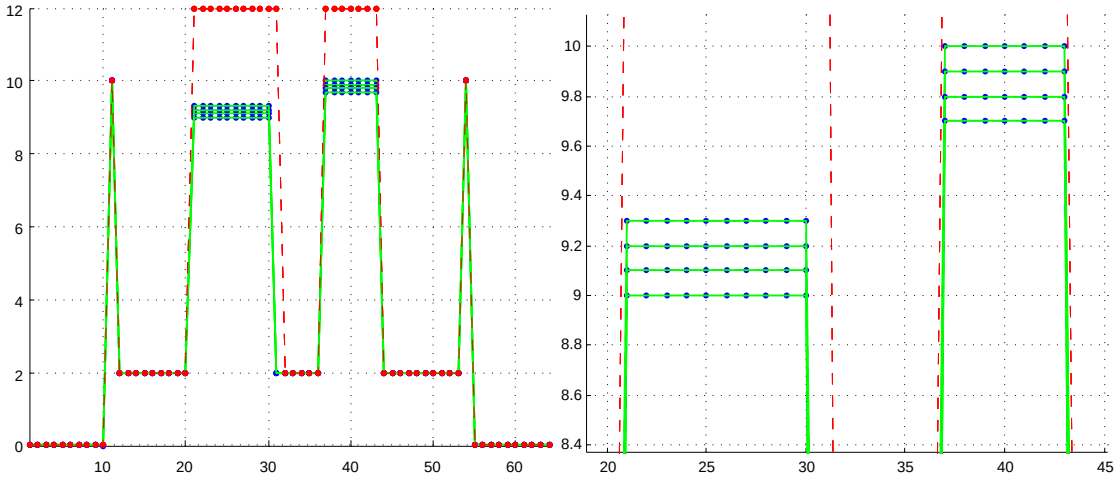

Figure 4.18: (Left) Reconstruction of the time series $X(t)$ viewed along the center cross-section with 12% Gaussian sampled Fourier coefficients and edge data, i.e. $\lambda = \gamma = .01, \tau = .1$, using the modified algorithm. (Right) Close up of the center ellipses.

Let $X_1^*(t)$ be the solution to the modified algorithm using the 12% Gaussian sampled Fourier measurements and edge information. Table 4.5 shows the relative errors between the exact time series solution and the recovered image.

t	RelErr($X^*(t)$)
1	5.21e-5%
2	1.00e-4%
3	7.82e-5%
4	7.82e-4%

Table 4.5: Relative errors between the exact time series and the recovered images using the modified algorithm with 12% Gaussian sampling pattern and edge information. Relative error is defined as $\text{RelErr}(X) := \|X^{\text{exact}}(t) - X\|_2 / \|X_{\text{exact}}(t)\|$, where $X_{\text{exact}}(t)$ is the exact solution at time t.

Again, experiments indicate that one would require roughly 20% Gaussian sampled Fourier measurements to obtain near exact reconstruction using total-variation techniques

alone, without any edge data. This means that an overall reduction by a factor of 5, compared with full Fourier measurement acquisition, can be achieved using total-variation techniques without edge data. In comparison, using our approach, one can achieve an overall reduction by a factor of 8, compared with full Fourier measurement acquisition. The additional factor of 1.6 reduction comes from using the edge information.

Remark 4.1. *Inexact edge measurements may also be attributed to the motion of a test subject in an fMRI study. Thus, these results indicate that one may still be able to use edge information in the presence of motion inaccuracies for exact reconstruction under the modified algorithm.*

Noisy Measurements

In this section we show reconstruction of the same time series as in the previous section, but in the presence of measurement noise. In this example, Gaussian white noise is simply added to the measurements y. The signal-to-noise (SNR) is defined as the ratio between the maximum zero frequency of $X(t)$ divided by the standard deviation of the Gaussian white noise. Figure 4.19 shows reconstruction with the 12% Gaussian Fourier sampling pattern for an SNR of 150 and 100.

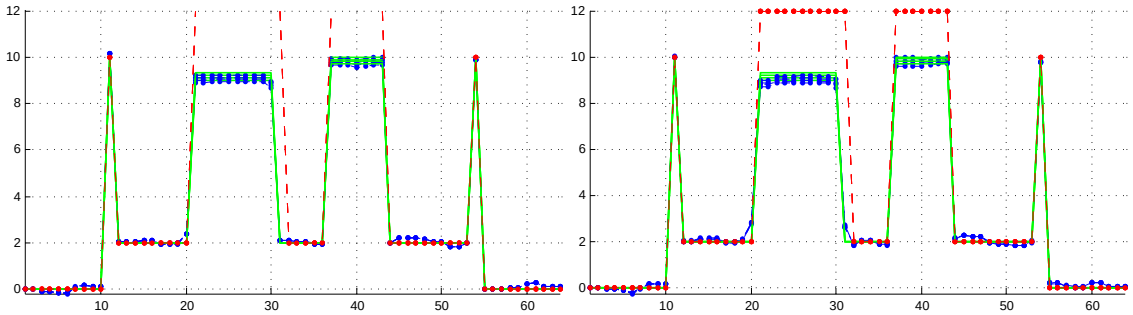

Figure 4.19: Reconstruction with noise. SNR = 150 on the left, and SNR = 100 on the right

Let $X_1^*(t)$ be the solution to the modified algorithm with SNR of 150 and $X_2^*(t)$ be the solution to the modified algorithm with an SNR of 100. The table below displays the relative errors between the exact time series solution and the recovered time series.

t	RelErr 1(snr = 150)	RelErr(snr = 100)
1	3.90%	5.51%
2	3.88%	5.63%
3	3.85%	5.65%
4	3.86%	5.76%

Table 4.6: Relative errors between the exact time series and the recovered images using the modified algorithm with 12% Gaussian sampling pattern and edge information with an SNR of 150 and 100. Relative error is defined as $\text{RelErr}(X) := \|X^{\text{exact}}(t) - X\|_2 / \|X_{\text{exact}}(t)\|$, where $X_{\text{exact}}(t)$ is the exact solution at time t.

4.3.3 2D and 3D Iso-intense Examples

Noiseless Measurements

Let us consider a time series in which a smaller ellipse is iso-intense with its surroundings: (1) the ellipse is the same intensity as its surrounding at some point and (2) the ellipse changes intensity by 1% at another time point. This increase in intensity corresponds to the activation of the region in response to some stimulus in a fMRI time series study. In the noiseless case, the time series will consist of two Shepp-Logan phantom images, where only the intensity of the center most ellipse is changing. Figure 4.20 displays the two images in the time series along with their center cross-sections.

A high resolution reference image also provides us with approximate edge information.[4] The center ellipse intensities for the two images in the time series differ by 1%, while the high resolution reference image intensity remains static. Denote the center ellipse intensity by $e_c(t)$. The ellipse intensities are displayed in the table below.

$X(t)$	$e_c(t)$
$X(1)$	10.0
$X(2)$	10.1
Y	12

Table 4.7: Ellipse intensities for the time series and reference image.

Using the same 12% Gaussian sampling pattern for the Fourier measurements as in Figure 4.17 and inexact edge measurement the reconstruction with the modified algorithm

[4] aproximate and not exact because of interpolation errors.

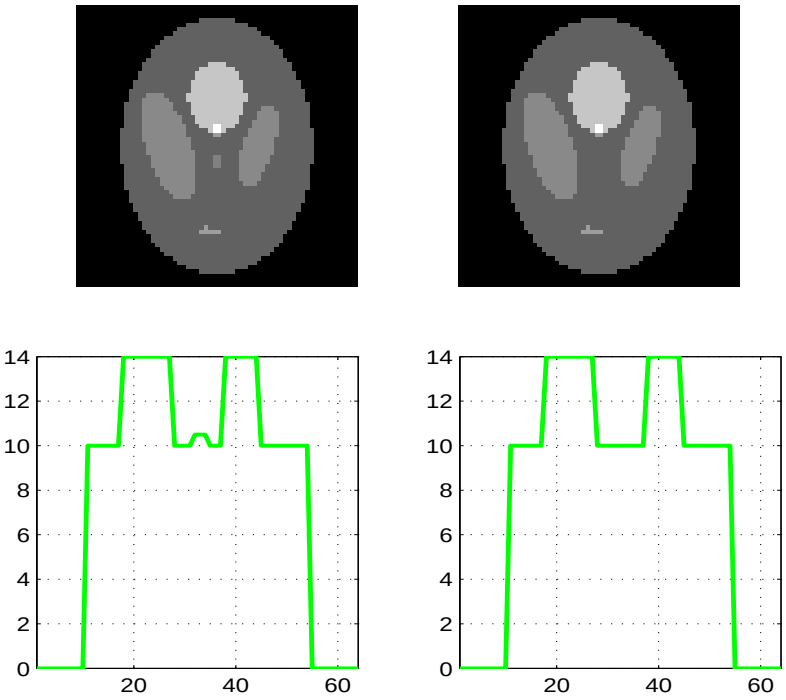

Figure 4.20: Images in the time series (top) and center cross-section of those images (bottom) to illustrate the center ellipse intensities. The ellipse that is iso-intense with its surrounding is displayed in the center.

is shown in the Figure 4.21. We also illustrate the reconstruction with and without edge information with the same 12% Gaussian sampling pattern in Figure 4.22.

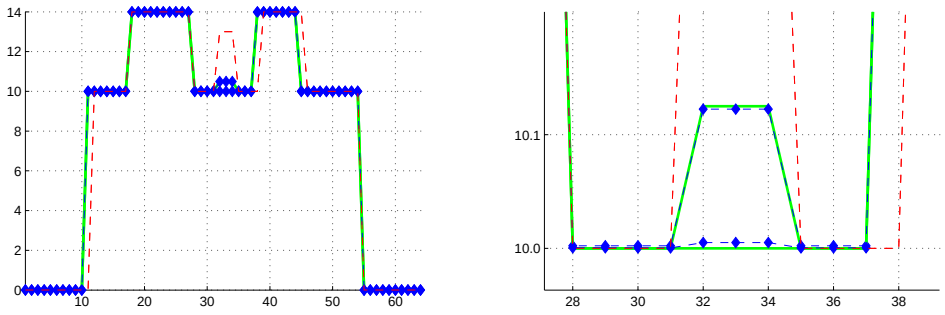

Figure 4.21: (Left) Reconstruction of the time series $X(t)$ viewed along the center cross-section with 12% Gaussian sampled Fourier coefficients and edge data, i.e. $\lambda = \gamma = .01, \tau = .1$, using the modified algorithm. (Right) Close up of the center ellipse. Red indicates the edge information, green indicates the exact solution, and blue indicates the recovered image.

Experiments indicate that one would require roughly 20% Gaussian sampled Fourier measurements to obtain near exact reconstruction using total-variation techniques alone, without any edge data. This means that an overall reduction by a factor of 5 can be

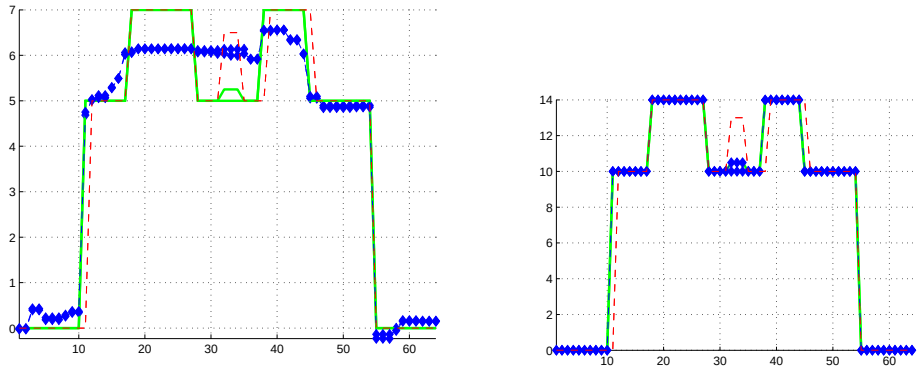

Figure 4.22: Reconstruction with edge data (right) and without edge data (left) using a 12% Gaussian sampling pattern.

achieved using total-variation techniques without edge data. In comparison, using our approach, one can achieve an overall reduction by a factor of 8, compared with full Fourier measurement acquisition. The additional factor of 1.6 reduction comes from using the inexact edge information.

Let $X_1^*(t)$ and $X_2^*(t)$ be the solution to the modified algorithm for the iso-intense time series example with and without edge data. The table below displays the relative errors between $X_i^*(t)$, $i = 1, 2$, and the exact solution.

t	RelErr($X_1^*(t)$)	RelErr($X_2^*(t)$)
1	4.92e-4%	16.10%
2	1.35e-4%	16.29%

Table 4.8: Relative errors for the iso-intense time series example. Relative error is defined as $\text{RelErr}(X) := \|X^{\text{exact}}(t) - X\|_2 / \|X_{\text{exact}}(t)\|$, where $X_{\text{exact}}(t)$ is the exact solution at time t.

Noisy Measurements

Now consider a 32 point iso-intense time series example in which the small center ellipse is only varying by 1% again. Figure 4.23 illustrates the intensities of the center ellipse as a function of time.

Using the modified algorithm with 12% Gaussian sampled Fourier measurements and the high resolution reference image with inexact edge information, Figure 4.24 illustrates

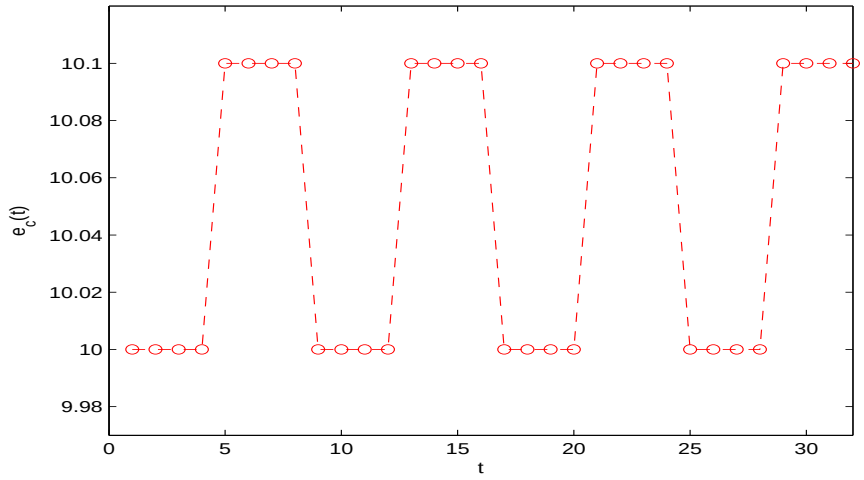

Figure 4.23: Center ellipse intensity as a function of $t = 1, \ldots, 32$.

the ellipse intensities for the reconstruction with SNR values at 150 and 100.[5] The value of the ellipse intensity at each time point is the mean ellipse intensity, i.e., the mean value of points in the center ellipse.

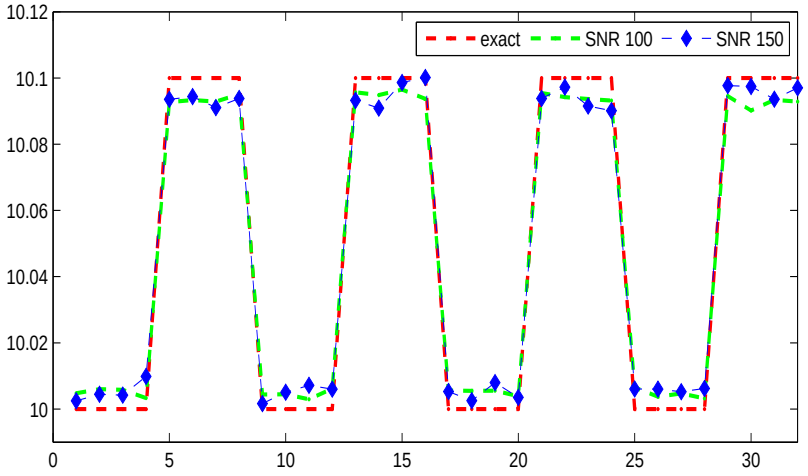

Figure 4.24: Mean ellipse intensities for the reconstruction with SNR 150 (green) and 100 (blue) versus the true ellipse intensities.

The Pearson correlation coefficient for the reconstructed ellipse intensity time series plot in Figure 4.24 is .9998 and .9968 for the SNR 150 and 100, respectively.

[5]The noise is added to each of the images in the 32 point time series.

3D Iso-intense Example

In this example, we consider a three-dimensional $64 \times 64 \times 64$ Shepp-Logan phantom image time series in which a small ellipse is iso-intense with its surroundings. Figure 4.25 shows the three-dimensional Shepp-Logan phantom at a fixed time point. The small maroon-colored ellipse varies in intensity in the time series.

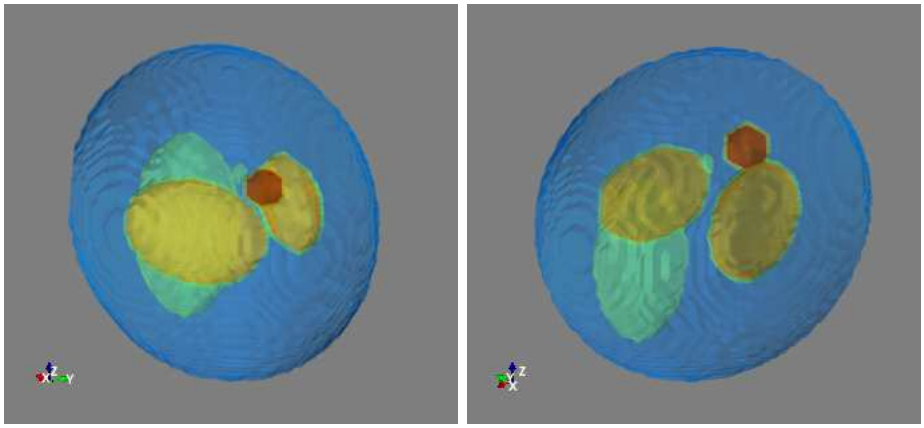

Figure 4.25: Three-dimensional $64 \times 64 \times 64$ Shepp-Logan phantom image from two different angles. The maroon colored ellipse changes ellipse intensities in the time series.

Figure 4.26 shows a two-dimensional slice in the area of the changing ellipse, along with a one-dimensional cross-section illustrating the changing ellipse intensity of the small maroon colored ellipse in Figure 4.25.

A high resolution $256 \times 256 \times 256$ reference image also provides us with approximate edge information.[6] The center ellipse intensities for the two images in the time series differ by 1%, while the high resolution reference image intensity remains static. Denote the top, center ellipse intensity shown in Figure 4.26, by $e_c(t)$. The ellipse intensities are displayed in the table below. We use a 2.2% Gaussian sampling pattern for the Fourier measurements

$X(t)$	$e_c(t)$
$X(1)$	12.0
$X(2)$	12.1
Y	13

Table 4.9: Ellipse intensities for the time series and reference image.

[6] Approximate and not exact because of interpolation errors.

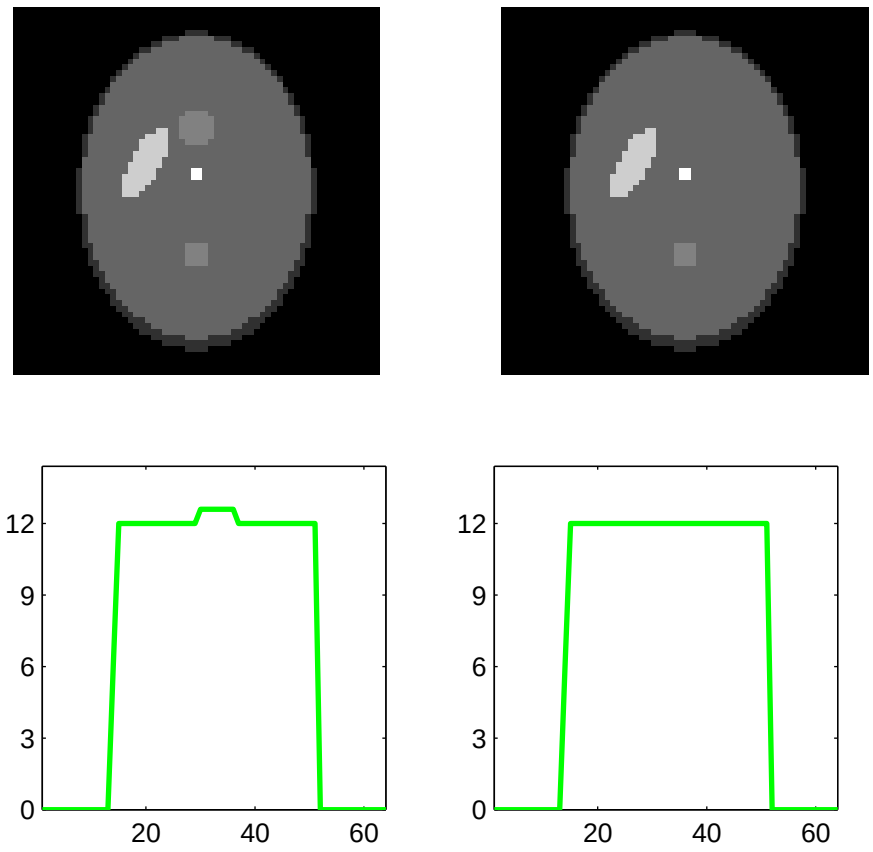

Figure 4.26: Two-dimensional slice (top) and one-dimensional cross-section (bottom) illustrating the changing ellipse intensity. The center, top ellipse in the two-dimensional slice (top) is changing ellipse intensity by 1%.

illustrated in Figure 4.27.[7] Figure 4.28 shows recovery of the three dimensional iso-intense example using the 2.2% Gaussian sampling pattern under the modified algorithm.We also illustrate the three-dimensional reconstruction with and without edge information with the same 2.2% Gaussian sampling pattern.

Experiments indicate that one would require roughly 5.5% Gaussian sampled Fourier measurements to obtain near exact reconstruction using total-variation techniques alone, without any edge data. This means that an overall reduction by a factor of 18 can be achieved using total-variation techniques without edge data. In comparison, using our approach, one can achieve an overall reduction by a factor of 45, compared with full

[7]We can get away with far fewer measurements compared to the two-dimensional cases because the ratio between sparsity and the size of the 64^3 image is significantly lower.

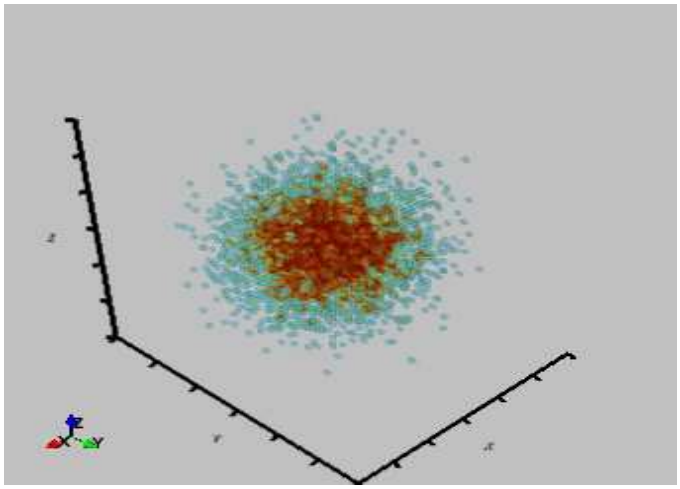

Figure 4.27: Three-dimensional Gaussian sampling pattern. Circles indicate location or indices of Fourier samples. Red contours indicate area of higher sampling.

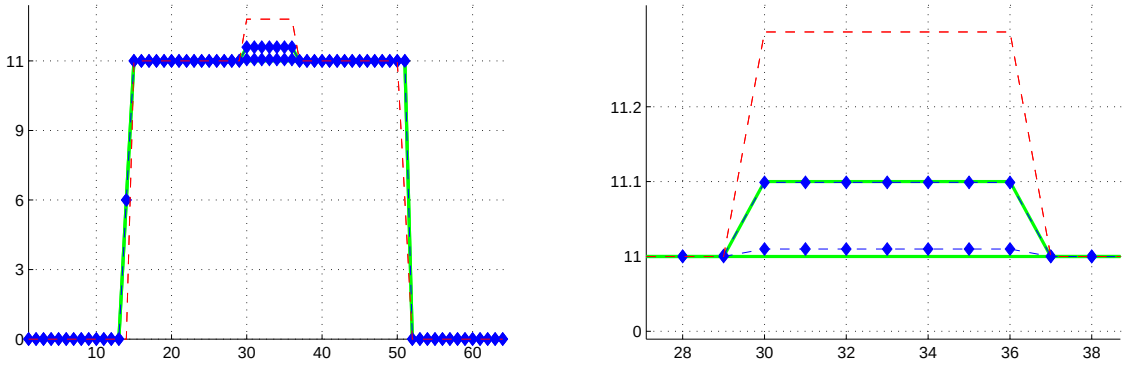

Figure 4.28: (Left) Reconstruction of the three-dimensional time series $X(t)$ viewed along the center cross-section with 2.2% Gaussian sampled Fourier coefficients and edge data, i.e. $\lambda = \gamma = .001, \tau = .01$, using the modified algorithm. (Right) Close up of the center ellipse. Red indicates the edge information, green indicates the exact solution, and blue indicates the recovered image.

Fourier measurement acquisition. The additional factor of 2.5 reduction comes from using the inexact edge information.

Let $X_1^*(t)$ and $X_2^*(t)$ be the solution to the modified algorithm for the iso-intense time series example with and without edge data. Table 4.10 displays the relative errors between $X_i^*(t)$, $i = 1, 2$, and the exact solution.

We also consider a 32 point iso-intense time series example in which the center ellipse in Figure 4.26 (top) is varying by 1% every four time points. Figure 4.30 illustrates the

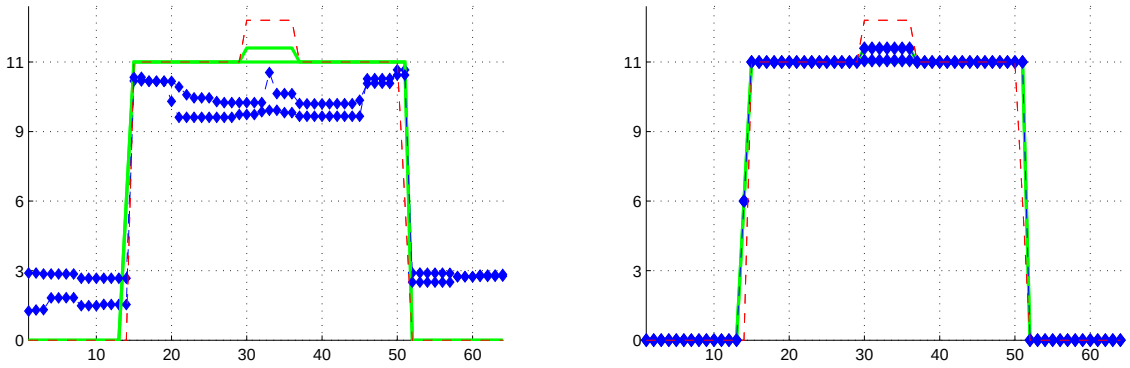

Figure 4.29: Reconstruction with edge data (right) and without edge data (left) using a 2.2% Gaussian sampling pattern.

t	RelErr($X_1^*(t)$)	RelErr($X_2^*(t)$)
1	0.07%	8.95%
2	0.02%	8.31%

Table 4.10: Relative errors for the iso-intense time series example. Relative error is defined as $\text{RelErr}(X) := \|X^{\text{exact}}(t) - X\|_2 / \|X_{\text{exact}}(t)\|$, where $X_{\text{exact}}(t)$ is the exact solution at time t.

intensities of the center ellipse as a function of time in this 32 point time series.

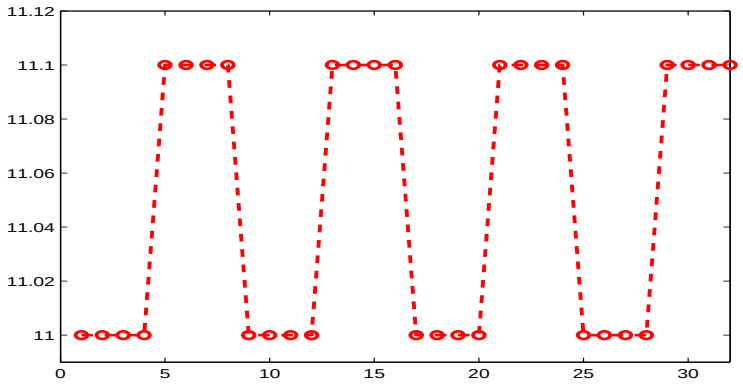

Figure 4.30: Center ellipse intensity as a function of $t = 1, \ldots, 32$.

Using the modified algorithm with 2.2% Gaussian sampled Fourier measurements and the high resolution reference image with inexact edge information, Figure 4.31 illustrates the ellipse intensities for the reconstruction with SNR values at 150 and 100.[8] The value of the ellipse intensity at each time point is the mean ellipse intensity, i.e., the mean value of points in the top, center ellipse.

[8]The noise is added to each of the images in the 32 point time series.

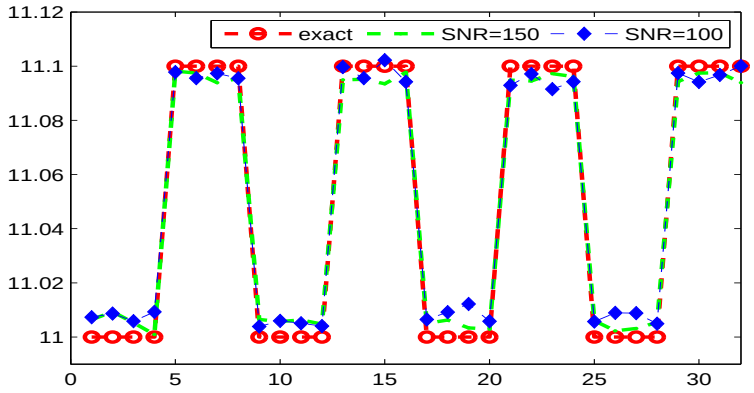

Figure 4.31: Mean ellipse intensities for the reconstruction with SNR 150 (green) and 100 (blue) versus the true ellipse intensities.

The Pearson correlation coefficient for the reconstructed ellipse intensity time series plot in Figure 4.31 is .9998 and .9955 for the SNR 150 and 100, respectively.

Chapter Five

Discussion and Concluding Remarks

In this work we introduce a new algorithm to reconstruct sparse gradient time series images from limited Fourier measurements and prior edge information from a reference image. Results show that the number of Fourier measurements required to obtain accurate reconstruction can be reduced by a factor of 1.6 - 3, beyond CS total-variation techniques alone, with the addition of Fourier edge information. In particular, numerical experiments indicate that Fourier sampling patterns which place more points near the zero frequencies, e.g., Gaussian or radial patterns, show the most significant reduction in the requisite Fourier measurements for exact and stable recovery. Moreover, the proposed algorithm is robust in the sense that the regularization parameters do not need to be extensively re-tuned for every image in the time series, the edge information can be inexact, and the jump intensities of the reference image can be within 20-30% of the true intensities in the time series images.

Another feature of our algorithm is that it can be easily adapted to existing compressed sensing methods by simply adding a second order conic constraint term which incorporates the prior edge measurements from the reference image. Thus, there is potential to combine the ideas in this paper with state-of-the-art algorithms (see, for example, [30] or [39]). In particular, there has been recent work in combining compressed sensing with parallel imaging techniques for improved accuracy and robustness [21, 22, 31, 17]. While our algorithm could be used in lieu of this approach, it is also easily conceivable to combine our technique with parallel imaging methods such as SENSE [31] or GRAPPA [17] to provide a further reduction in the amount of Fourier space data acquired for an fMRI time series study.

Appendix A

Relative Entropy and gPC Distributions

A.1 Relative Entropy

A.1.1 Properties

In this appendix we review some properties of relative entropy [7, Section 1.4].

Proposition A.1. *Let $\mathcal{X}$ be a Polish metric space and $\mathcal{P}(\mathcal{X})$ be the collection of probability measures on $\mathcal{X}$. Consider $\nu, \mu \in \mathcal{P}(\mathcal{X})$. Then, the following properties hold.*

a) $R(\nu\|\mu) \geq 0$ *and* $R(\nu\|\mu) = 0$ *if and only* $\nu = \mu$.

b) $R(\nu\|\mu)$ *is a convex, lower semicontinuous function of* $(\nu, \mu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X})$. *In particular,* $R(\nu\|\mu)$ *is a convex, lower semicontinuous function of each variable* ν *or* μ *separately.*

c) For each $\mu \in \mathcal{P}(\mathcal{X})$, $R(\cdot\|\mu)$ *has compact level sets, i.e., for each* $M < \infty$, *the set* $\{\nu \in \mathcal{P}(\mathcal{X}) : R(\nu\|\mu) \leq M\}$ *is a compact subset of* $\mathcal{P}(\mathcal{X})$.

Proof. For part (a) it suffices to consider the case when $R(\nu\|\mu) < \infty$ only. Note that $s \log s \geq s - 1$ with equality if and only if $s = 1$. Therefore,

$$\begin{aligned} R(\nu\|\mu) &\triangleq \int_{\mathcal{X}} \log\left(\frac{d\nu}{d\mu}(x)\right)\nu(dx) \\ &= \int_{\mathcal{X}} \frac{d\nu}{d\mu}(x)\log\left(\frac{d\nu}{d\mu}(x)\right)\mu(dx) \quad \geq \quad \int_{\mathcal{X}}\left(\frac{d\nu}{d\mu}(x) - 1\right)\mu(dx) \quad = \quad 0, \end{aligned}$$

and equality holds when $\nu = \mu$ almost everywhere with respect to μ. For the proofs of (b) and (c) see [7, Lemma 1.4.3]. □

Another property of relative entropy is the chain rule (see [7, Theorem C.3.1] and [7, Corollary C.3.3]).

Proposition A.2. ***(Chain Rule)*** *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish metric spaces and $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ be the collection of probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively. Consider α and*

β to be joint probability measures on $\mathcal{X} \times \mathcal{Y}$. We denote $\alpha_1 \in \mathcal{P}(\mathcal{X})$ and $\beta_1 \in \mathcal{P}(\mathcal{X})$ to be the first marginals of α and β, and by $\alpha(dy|x) \in \mathcal{P}(\mathcal{Y})$ and $\beta(dy|x) \in \mathcal{P}(\mathcal{Y})$ the stochastic kernels on $\mathcal{Y}$ given $\mathcal{X}$ (this is essentially a conditional distribution on $\mathcal{Y}$ given $X = x$) for which we have the decompositions (for a precise definition see [7, page 35])

$$\begin{aligned} \alpha(dx \times dy) &= \alpha_1(dx) \otimes \alpha(dy|x), \\ \beta(dx \times dy) &= \beta_1(dx) \otimes \beta(dy|x). \end{aligned}$$

Then the function mapping $x \in \mathcal{X} \mapsto R(\alpha(\cdot|x)\|\beta(\cdot|x))$ is measurable and

$$R(\alpha|\beta) = R(\alpha_1\|\beta_1) + \int_{\mathcal{X}} R(\alpha(\cdot|x)\|\beta(\cdot|x))\alpha_1(dx).$$

Corollary A.1. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Polish metric spaces and $\mathcal{P}(\mathcal{X})$ and $\mathcal{P}(\mathcal{Y})$ be the collection of probability measures on $\mathcal{X}$ and $\mathcal{Y}$, respectively. If $\gamma, \theta \in \mathcal{P}(\mathcal{X})$ and $\lambda, \mu \in \mathcal{P}(\mathcal{Y})$, then*

$$R(\gamma \times \lambda\|\theta \times \mu) = R(\gamma\|\theta) + R(\lambda\|\mu).$$

A.1.2 Formulas

Here we evaluate relative entropies for some of the most common types of distributions listed in the last section. Since most of the distributions listed in this table fall within the exponential family, we start with the general form

$$f_X(x;\boldsymbol{\theta}) = h(x) \exp\left(\sum_{i=1}^{s} \eta_i(\boldsymbol{\theta}) T_i(x) - A(\eta_1(\boldsymbol{\theta}), ..., \eta_s(\boldsymbol{\theta}))\right), \tag{A.1}$$

where $T_i(x), h(x), \eta_i(\boldsymbol{\theta})$, and $A(\boldsymbol{\theta})$ are known functions, $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^T$ is the parameter of the family.

Example A.1. *For Gaussian distributions $\theta = (\mu, \sigma)^T$, where μ is the mean and σ is the variance. This is put in the general form by setting*

$$\boldsymbol{\theta} = (\mu, \sigma)^T, \quad \eta = \left(\frac{\mu}{\sigma^2}, \frac{-1}{2\sigma^2}\right)^T, \quad h(x) = \frac{1}{\sqrt{2\pi}}, \quad T(x) = (x, x^2)^T,$$

$$A(\eta(\boldsymbol{\theta})) = \frac{\mu^2}{2\sigma^2} + \log|\sigma| = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}\log\left|\frac{1}{2\eta_2}\right|,$$

producing

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}.$$

We now state the relative entropy formula for distributions which have a probability density function given by the form of exponential family in (A.1). The relative entropy between a probability measure P with density $p(x)$ and another probability measure Q with density $q(x)$ is given by

$$R(P\,\|Q\,) = \int_\Gamma p(x) \log \frac{p(x)}{q(x)} dx \tag{A.2}$$

whenever P is absolutely continuous with respect to Q, where Γ is the support of p and q. In all other cases it is defined to be ∞.

A simplified version of the relative entropy formula for distributions from the same exponential family type is available. Suppose that $q(x)$ and $p(x)$ have the same h, η_i, T_i, A, but with different parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ (note that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are vectors). If we associate $q(x)$ with $\boldsymbol{\theta}_1$ and $p(x)$ with $\boldsymbol{\theta}_2$, then it can be shown that the relative entropy formula reduces to

$$R(P\,\|Q\,) = \sum_{i=1}^{s} (\eta_i(\boldsymbol{\theta}_2) - \eta_i(\boldsymbol{\theta}_1)) \int_\Gamma T_i(x) p(x) dx - A(\boldsymbol{\eta}(\boldsymbol{\theta_2})) + A(\boldsymbol{\eta}(\boldsymbol{\theta}_1)). \tag{A.3}$$

Gaussian. If $Q = N(\mu_1, \sigma_1)$ and $P = N(\mu_2, \sigma_2)$, then using the fact that $\int T_1(x) p(x) dx =$

μ_2 and $\int T_2(x)p(x)dx = \mu_2^2 + \sigma_2^2$, we get

$$R(P\,\|Q\,) = \frac{1}{2\sigma_1^2}\left((\mu_1^2 - \mu_2^2) + (\sigma_2 - \sigma_1)\right) + \log\frac{\sigma_1}{\sigma_2}. \tag{A.4}$$

Beta. For the beta distribution, the parameters in the exponential family are given by

$$\eta = (\alpha - 1, \beta - 1)^T, T = (\log(x), \log(1-x))^T, h(x) = 1, A = \log\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

where

$$f_X(x;\alpha,\beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}x^{\alpha-1}(1-x)^{\beta-1}, \qquad 0 < x < 1, \alpha, \beta > 1.$$

Another form of the beta distribution, which is more closely related to the Jacobi polynomials, is given by

$$\tilde{f}_X(k;\tilde{\alpha},\tilde{\beta}) = \frac{(1-x)^{\tilde{\alpha}}(1+x)^{\tilde{\beta}}}{2^{\tilde{\alpha}+\tilde{\beta}+1}B(\tilde{\alpha}+1,\tilde{\beta}+1)}, \qquad -1 < x < 1, \tilde{\alpha}, \tilde{\beta} > -1,$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ is the beta function. Note that the relation between the two is simply a rescaling and a variable substitution given by $\tilde{\alpha} = \beta - 1, \tilde{\beta} = \alpha - 1$. We will use the primary form to determine the relative entropy formula. Letting $P = \text{beta}(\alpha_2, \beta_2)$ and $Q = \text{beta}(\alpha_1, \beta_1)$,

$$\int_0^1 T_1(x)p(x)dx = \psi(\alpha_2) - \psi(\alpha_2+\beta_2), \int_\Gamma T_2(x)p(x)dx = \psi(\beta_2) - \psi(\alpha_2+\beta_2),$$

where $\psi(x)$ is known as the digamma function, or the zeroth order of the polygamma function. Then,

$$\begin{aligned} R(P\,\|Q\,) &= (\alpha_2 - \alpha_1)(\psi(\alpha_2) - \psi(\alpha_2+\beta_2)) \\ &\quad +(\beta_2 - \beta_1)(\psi(\beta_2) - \psi(\alpha_2+\beta_2)) + \ln\frac{B(\alpha_1,\beta_1)}{B(a_2,\beta_2)}. \end{aligned}$$

Gamma. Here we have

$$\eta = (-\beta, \alpha - 1)^T, T = (x, \ln(x))^T, h(x) = 1, A = \log\Gamma(\alpha) - \alpha \log \beta$$

If $P = \text{gamma}(\alpha_2, \beta_2), Q = \text{gamma}(\alpha_1, \beta_1)$, where α_i, β_i are the so-called shape parameters, then

$$R(P \,\|Q\,) = \frac{\alpha_1}{\beta_1}(\beta_2 - \beta_1) + (\alpha_1 - \alpha_2)(\psi(\alpha_1)) - \log(\beta_1)) + \log \frac{\Gamma(\alpha_2)\beta_1^{\alpha_1}}{\Gamma(\alpha_1)\beta_2^{a_2}}.$$

Binomial. Here

$$\eta = (-\beta, \alpha - 1)^T, T = (x, \log(x))^T, h(x) = 1, A = \log\Gamma(\alpha) - \alpha \log \beta.$$

If $P = \text{binomial}(n, p_2), Q = \text{binomial}(n, p_1)$, then

$$R(P \,\|Q\,) = \log \frac{p_1^{\mu_1}(1 - p_2)^{\mu_1 - n}}{p_2^{\mu_1}(1 - p_1)^{\mu_1 - n}}, \qquad \mu_1 = np_1.$$

Poisson. In this case

$$\eta = \log \lambda, T = x, h(x) = \frac{1}{x!}, A = \lambda$$

If $P = \text{Poisson}(\lambda_2), Q = \text{Poisson}(\lambda_1)$, then the relative entropy distance is

$$R(P \,\|Q\,) = \lambda_1 - \lambda_2 + \lambda_2 \log \frac{\lambda_2}{\lambda_1}.$$

Note that relative entropy formula is invariant under shifting and scaling of both distributions simultaneously. In fact, suppose that ψ and its inverse are both well defined and measurable, X and Y have distributions P and Q, and that $\psi(X)$ and $\psi(Y)$ have

distributions $\bar{P}$ and $\bar{Q}$. Then [7, Lemma E.2.1] $R(\bar{P}\,\|\bar{Q}) = R(P\,\|Q)$. A list of relative entropy formulas follows.

Distribution	$\mathbf{R}((\cdot)_2 \,\|\, (\cdot)_1)$
Gaussian	$\frac{1}{2\sigma_1^2}\left[(\mu_1^2 - \mu_2^2) + (\sigma_2 - \sigma_1)\right] + \log\frac{\sigma_1}{\sigma_2}$
Beta/ Uniform	$(\alpha_2 - \alpha_1)\left[\psi(\alpha_2) - \psi(\alpha_2 + \beta_2)\right]$ $+(\beta_2 - \beta_1)\left[\psi(\beta_2) - \psi(\alpha_2 + \beta_2)\right] + \log\frac{B(\alpha_1, \beta_1)}{B(\alpha_2, \beta_2)}$
Gamma	$\frac{\alpha_1}{\beta_1}(\beta_2 - \beta_1) + (\alpha_1 - \alpha_2)\left[\psi(\alpha_1) - \log\beta_1\right]$ $+\log\left[\Gamma(\alpha_2)\beta_1^{\alpha_1}/\Gamma(\alpha_1)\beta_2^{\alpha_2}\right]$
Binomial	$\log\left[p_1^{\mu_1}(1-p_2)^{\mu_1 - n}/p_2^{\mu_1}(1-p_1)^{\mu_1 - n}\right]$
Poisson	$\lambda_1 - \lambda_2 + \lambda_2 \log\frac{\lambda_2}{\lambda_1}$

Table A.1: Relative entropy formulas for common exponential families

A.2 Distributions for Polynomial Basis Functions

We list some of the most common distributions, both continuous and discrete, and their corresponding gPC basis representations (see [41]).

Distribution	**gPC Polynomial Basis**
Gaussian	Hermite
Gamma	Laguerre
Beta	Jacobi
Uniform	Legendre
Poisson	Charlier
Binomial	Krawtchouk
Negative Binomial	Meixner
Hypergeometric	Hahn

Table A.2: Probability distributions and their corresponding gPC polynomial basis functions

Appendix B

Tools for Compressed Sensing and Edge Detection

B.1 Matrix and Vector Norm Properties

Consider a complex vector $x \in \mathbb{C}^n$ and its associated ℓ_p-norm defined as

$$\begin{aligned} \|x\|_p &= \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}, \quad 0 < p < \infty, \\ \|x\|_\infty &:= \max_{1 \le j \le n} |x_j|. \end{aligned}$$

For $0 < p < 1$, the ℓ_p-norm is only a quasi norm, e.g. the triangle inequality is not satisfied. Consider a complex matrix $A \in \mathbb{C}^{m \times n}$ with $m \le n$ and typically $m \ll n$. Define the operator norm of a matrix as

$$\|A\|_p := \max_{x \in \mathbb{C}^n, \|x\|_p = 1} \|Ax\|_p$$

where $Ax \in \mathbb{C}^m$. For $p = 1, 2, \infty$ the matrix norm takes on the following explicit forms:

$$\begin{aligned} \|A\|_1 &= \max_{j=1,\ldots,N} \sum_{i=1}^{m} |a_{ij}|, \quad \text{(max absolute column sum)} \\ \|A\|_2 &= \sqrt{\lambda_{\max}(A^* A)}, \quad \text{(max singular value of } A\text{)} \\ \|A\|_\infty &= \max_{i=1,\ldots,N} \sum_{j=1}^{m} |a_{ij}|. \quad \text{(max absolute row sum)} \end{aligned}$$

The following norm equivalences also hold between vector norms,

$$\begin{aligned} \|x\|_2 &\le \|x\|_1 \le \sqrt{n}\|x\|_2, \\ \|x\|_\infty &\le \|x\|_2 \le \sqrt{n}\|x\|_\infty. \end{aligned}$$

and matrix norms

$$\begin{aligned} \frac{1}{\sqrt{n}}\|A\|_2 &\le \|A\|_1 \le \sqrt{m}\|A\|_2, \\ \frac{1}{\sqrt{m}}\|A\|_1 &\le \|A\|_2 \le \sqrt{n}\|A\|_1. \end{aligned}$$

The above inequalities can be proven from the definition, matrix norm properties, and Jensen's inequality. Finally, for a Hermitian matrix $A \in \mathbb{C}^{n \times n}$, the eigenvalues are neces-

sary real-valued and lie in the set

$$\{\langle Ax, x\rangle : x \in \mathbb{C}^n, \|x\|_2 = 1\} \subset \mathbb{R}.$$

Moreover,

$$\|A\|_2 \quad = \quad \sup_{\|x\|_2=1} |\langle Ax, x\rangle| .$$

B.2 Proofs of the Main CS Theorems

We prove the main results for exact and stable recovery of s-sparse signals via ℓ_1-minimization [34, 12] .

B.2.1 Proof of Theorem 2.3

Let us show that A satisfies the NSP. This means that for $v \in \mathbb{C}^N, v \in \ker(A)$, it must be true that

$$\|v_{S^c}\|_1 - \|v_S\|_1 \quad > \quad 0, \quad \forall S \subset [N], |S| = s,$$

where $[N] \stackrel{\circ}{=} \{1, \ldots, N\}$. It is enough to show the result for the index set S_0 of the s largest entries (largest in modulus) of the v. Why is this true? Let S be an arbitrary index set such that $|S| = |S_0| = s$. Then, $\|v_{S_0}\|_1 \geq \|v_S\|_1$ and, consequently, $\|v_{S_0^c}\|_1 \leqslant \|v_{S^c}\|_1$. Therefore,

$$\|v_{S^c}\|_1 - \|v_S\|_1 \quad \geq \quad \|v_{S_0^c}\|_1 - \|v_{S_0}\|_1.$$

Therefore, if $\|v_{S_0^c}\|_1 - \|v_{S_0}\|_1 > 0$, clearly, it holds true for any other index set S with $|S| = s$.

Now, divide $S_0^c = S_0 \cup S_1 \cup \cdots$ into disjoint s index sets S_i, where S_1 is the index set of the next s largest entries of v in $[N]|\backslash S_0$, with $|S_1| = s$, S_2 the s largest entries in

$[N]\backslash S_0 \cup S_1$, and so on. Since v is in the null space of A, we have

$$A\left(v_{S_0} + \sum_i v_{S_i}\right) \quad = \quad 0$$

and therefore $Av_{S_0} = -A(v_{S_1} + v_{S_2} + \cdots)$. The properties of the RIP constant δ_{2s} give

$$\begin{aligned} \|v_{S_0}\|_2^2 \quad \leq \quad \frac{1}{1-\delta_{2s}}\|Av_{S_0}\|_2^2 \quad &= \quad \frac{1}{1-\delta_{2s}}\left\langle Av_{S_0}, Av_{S_0}\right\rangle \\ &= \quad \frac{1}{1-\delta_{2s}}\left\langle Av_{S_0}, -A\left(v_{S_1}+\cdots\right)\right\rangle \\ &= \quad \frac{1}{1-\delta_{2s}}\sum_{i\geq 1}\left\langle Av_{S_0}, Av_{S_i}\right\rangle. \end{aligned}$$

Using part (c) of Proposition 2.1, since $S_0 \cap S_i = \emptyset$, and $\mathrm{supp}(S_0) + \mathrm{supp}(S_i) = 2s$, we have

$$\left|\left\langle Av_{S_0}, Av_{S_i}\right\rangle\right| \quad \leq \quad \delta_{2s}\|v_{S_0}\|_2\|v_{S_i}\|_2.$$

Thus,

$$\|v_{S_0}\|_2 \quad \leq \quad \frac{\delta_{2s}}{1-\delta_{2s}}\sum_{i\geq 1}\|v_{S_i}\|_2.$$

Also, since the s non-zero entries of v_{S_i} do not exceed the s non-zero entries of $v_{S_{i-1}}$ for $i \geq 1$, i.e.,

$$\min_j |(v_{S_{i-1}})_j| \quad \geq \quad \max_j |(v_{S_i})_j|, \quad \forall i,$$

then,

$$\sum_{l\in S_{i-1}} |v_l| \quad \geq \quad s|v_j|, \qquad \forall j \in S_i.$$

Therefore, $|v_i| \leq \frac{1}{s}\|v_{S_{i-1}}\|_1$. Now,

$$\|v_{S_i}\|_2^2 \quad = \quad \sum_{l\in S_i}|v_l|^2 \quad \leq \quad \frac{1}{s}\|v_{S_{i-1}}\|_1^2$$

By the equivalence of norms (see Section B.1), $\|x\|_1 \leq \sqrt{n}\|x\|_2$. Thus,

$$\begin{aligned}
\|v_{S_0}\|_1 \quad \leq \quad \sqrt{s}\|v_{S_0}\|_2 \quad &= \quad \sqrt{s}\frac{\delta_{2s}}{1-\delta_{2s}}\sum_{i\geq 1}\|v_{S_i}\|_2 \\
&\leq \quad \sqrt{s}\frac{\delta_{2s}}{1-\delta_{2s}}\sum_{i\geq 1}\frac{1}{\sqrt{s}}\|v_{S_{i-1}}\|_1 \\
&= \quad \frac{\delta_{2s}}{1-\delta_{2s}}\sum_{i\geq 1}\|v_{S_{i-1}}\|_1 \\
&\leq \quad \frac{\delta_{2s}}{1-\delta_{2s}}\left(\|v_{S_0}\|_1 + \|v_{S_0^c}\|_1\right)
\end{aligned}$$

By assumption, $\delta_{2s} < \frac{1}{3}$, which means $1 - \delta_{2s} > \frac{2}{3}$ and that $\frac{\delta_{2s}}{1-\delta_{2s}} < \frac{1}{2}$. Finally,

$$\|v_{S_0}\|_1 \quad < \quad \frac{1}{2}\left(\|v_{S_0}\|_1 + \|v_{S_0^c}\|_1\right)$$

and so $\|v_{S_0^c}\|_1 > \|v_{S_0}\|_1$ and the NSP is proven! □

B.2.2 Proof of Theorem 2.4

Let $v \in \mathbb{C}^N$ be an arbitrary vector, not necessarily in the null space of the measurement matrix $A \in \mathbb{C}^{m\times N}$. Again, let us partition the index set into decreasing, disjoint sets of cardinality s, denoted by $S_0, S_1, \ldots$. By the RIP, observe that

$$\begin{aligned}
\|v_{S_0}\|_2^2 + \|v_{S_1}\|_2^2 \quad &\leq \quad \frac{1}{1-\delta_{2s}}\|A(v_{S_0} + v_{S_1})\|_2^2 \\
&= \quad \frac{1}{1-\delta_{2s}}\left\langle Av, A\left(v_{S_0} + v_{S_1}\right)\right\rangle + \frac{1}{1-\delta_{2s}}\sum_{i\geq 2}\left\langle A(-v_{S_i}), Av_{S_0}\right\rangle \\
&\quad + \left\langle A(-v_{S_i}), Av_{S_1}\right\rangle .
\end{aligned}$$

By part (c) of Proposition 2.1,

$$\left|\left\langle A(-v_{S_i}), Av_{S_0}\right\rangle\right| \quad \leq \quad \delta_{2s}\|v_{S_i}\|_2\|v_{S_0}\|_2$$

for an abitrary i, so that

$$\langle A(-v_{S_i}), Av_{S_0}\rangle + \langle A(-v_{S_i}), Av_{S_1}\rangle \;\leq\; \delta_{2s}\|v_{S_i}\|_2\left(\|v_{S_0}\|_2 + \|v_{S_1}\|_2\right).$$

Then,

$$\|v_{S_0}\|_2^2 + \|v_{S_1}\|_2^2 \;\leq\; (c + d\Sigma)\left(\|v_{S_0}\|_2 + \|v_{S_1}\|_2\right),$$

where $c \triangleq \dfrac{1+\delta_{2s}}{1-\delta_{2s}}\|Av\|_2$, $d \triangleq \delta_{2s}/(1-\delta_{2s})$, and $\Sigma \triangleq \displaystyle\sum_{i\geq 2}\|v_{S_i}\|_2$. Furthermore, since $\|v_{S_0}\|_1 \leq \sqrt{s}\,\|v_{S_0}\|_2$ (see B.1),

$$\|v_{S_0}\|_1 \;\leq\; \sqrt{s}\cdot\frac{1+\sqrt{2}}{2}\cdot(c + d\Sigma).$$

Combining this with the inequality $\Sigma \leq \dfrac{1}{\sqrt{s}}\|v_{S_0^c}\|_1$, we get

$$\|v_{S_0}\|_1 \;\leq\; \sqrt{s}\frac{\lambda}{2}\|Av\|_2 + \mu\left\|v_{S_0^c}\right\|_1, \tag{B.1}$$

where $\lambda \triangleq (1+\sqrt{2})\cdot\dfrac{1+\delta_{2s}}{1-\delta_{2s}}$ and $\mu := \dfrac{1+\sqrt{2}}{2}\cdot\dfrac{\delta_{2s}}{1-\delta_{2s}}$.

Now, let S_0 be the set of indices of the s largest absolute value components of the exact solution x, and define $v \triangleq x - x^*$ where x^* is the minimizer of the ℓ_1 minimization with noise. Since x^* is the minimizer, we have

$$\|x\|_1 \;\geq\; \|x^*\|_1 \;.$$

The triangle inequality gives

$$\|x_{S_0}\|_1 + \left\|x_{S_0^c}\right\|_1 \;\geq\; \|x_{S_0}\|_1 - \|v_{S_0}\|_1 + \|v_{S_0^c}\|_1 - \|x_{S_0^c}\|_1.$$

and upon rearangement of the inequality

$$\left\|v_{S_0^c}\right\|_1 \;\leq\; 2\left\|x_{S_0^c}\right\|_1 + \|v_{S_0}\|_1 \;=\; 2\sigma_s(x)_1 + \|v_{S_0}\|_1.$$

Incorporating the noise ε and using B.1 gives

$$\left\|v_{S_0^c}\right\|_1 \;\leq\; 2\sigma_s(x)_1 + \sqrt{s}\lambda e + \mu\left\|v_{S_0^c}\right\|_1$$

By assumption, $\delta_{2s} < 2(3-\sqrt{2})/7$, so that

$$\frac{\delta_{2s}}{1-\delta_{2s}} \;<\; 2\left(\sqrt{2}-1\right)$$

which means $\mu < 1$. Therefore,

$$\left\|v_{S_0^c}\right\|_1 \;\leq\; \frac{2}{1-\mu}\sigma_s(x)_1 + \sqrt{s}\frac{\lambda}{1-\mu}e.$$

and

$$\|v\|_1 \;\leq\; \|v_{S_0}\|_1 + \left\|v_{S_0^c}\right\|_1 \;\leq\; \frac{2(1+\mu)}{1-\mu}\sigma_s(x)_1 + \sqrt{s}\frac{2\lambda}{1-\mu}e. \tag{B.2}$$

Finally, we get

$$\|v\|_1 \;\leq\; C_1\sigma_s(x)_1 + \sqrt{s}D_1 e, \tag{B.3}$$

where

$$C_1 \;\triangleq\; \frac{2(1+\mu)}{1-\mu}, \quad D_1 \;\triangleq\; \frac{2\lambda}{1-\mu} \quad .$$

For the ℓ_2 estimate, note that

$$\|v\|_2 \;\leq\; \sum_{k\geq 0}\|v_{S_k}\|_2 \;\leq\; \left(1+\sqrt{2}\right)c + \left(\left(1+\sqrt{2}\right)d + 1\right)\Sigma$$

and

$$\left(1+\sqrt{2}\right)c \;=\; \lambda\|Av\|_2 \;\leq\; 2\lambda e,$$

since $\|Av\|_2 \leq 2\varepsilon$. Also,

$$\left(1+\sqrt{2}\right)d + 1 \;=\; \frac{1}{2}\left(\lambda + 1 - \sqrt{2}\right).$$

If we define $\nu \triangleq \frac{1}{2}\left(\lambda + 1 - \sqrt{2}\right)$, then

$$\|v\|_2 \quad \leq \quad 2\lambda e + \nu\Sigma \ .$$

Finally, since $\Sigma \leq \frac{1}{\sqrt{s}}\|v_{S_0^c}\|_1$ and by B.2, we get

$$\|v\|_2 \quad \leq \quad C_2\frac{\sigma_s(x)_1}{\sqrt{s}} + D_2 e \tag{B.4}$$

where

$$C_2 \quad := \quad \frac{\lambda + 1 - \sqrt{2}}{1-\mu}, \quad D_2 \quad := \quad \frac{\lambda\left(\lambda + 1 - \sqrt{2}\right)}{2(1-\mu)} + 2\lambda.$$

B.3 and B.4 show the inequalities we want. □

B.3 Proof of the Discrete Uncertainty Principle

For convenience, we use the notation $\mathbb{Z}/N\mathbb{Z} = \{0, \ldots, N-1\}$. First, Plancherel's identity gives

$$\sum_{x\in\mathbb{Z}/N\mathbb{Z}} |f(x)|^2 \quad = \quad \sum_{\xi\in\mathbb{Z}/N\mathbb{Z}} |\hat{f}(\xi)|^2 \ .$$

By definition of the discrete Fourier transform,

$$f(x) \quad = \quad \frac{1}{\sqrt{N}} \sum_{\xi\in\mathbb{Z}/N\mathbb{Z}} \hat{f}(\xi) e^{-i2\pi\xi x},$$

we get

$$\sup_{x\in\mathbb{Z}/N\mathbb{Z}} |f(x)| \quad \leq \quad \frac{1}{\sqrt{N}} \sum_{\xi\in\mathbb{Z}/N\mathbb{Z}} |\hat{f}(\xi)| \tag{B.5}$$

since $|e^{-i2\pi x\xi}| = 1$. Also, it is easy to see that

$$\sum_{x\in\mathbb{Z}/N\mathbb{Z}} |f(x)|^2 \quad \leq \quad |\operatorname{supp}(f)| \left(\sup_{x\in\mathbb{Z}/N\mathbb{Z}} |f(x)|\right)^2 \tag{B.6}$$

where $\text{supp}(f)$ is the number of elements in the vector f such that. $f \neq 0$. Combining (B.5) and (B.6) gives us

$$\sum_{x \in \mathbb{Z}/N\mathbb{Z}} |f(x)|^2 \quad \leq \quad |\operatorname{supp}(f)| \left(\frac{1}{\sqrt{N}} \sum_{\xi \in \mathbb{Z}/N\mathbb{Z}} |\hat{f}(\xi)| \right)^2 . \tag{B.7}$$

Jensen's inequality implies

$$\left(\sum_{\xi \in \mathbb{Z}/N\mathbb{Z}} |\hat{f}(\xi)| \right)^2 \quad \leq \quad |\operatorname{supp}(\hat{f})| \left(\sum_{\xi \in \mathbb{Z}/N\mathbb{Z}} |\hat{f}(\xi)|^2 \right) \tag{B.8}$$

Combining (B.7) and (B.8) gives

$$\sum_{x \in \mathbb{Z}/N\mathbb{Z}} |f(x)|^2 \quad \leq \quad \frac{1}{N} |\operatorname{supp}(f)| |\operatorname{supp}(\hat{f})| \left(\sum_{\xi \in \mathbb{Z}/N\mathbb{Z}} |\hat{f}(\xi)|^2 \right).$$

Finally, applying Plancherel's identity again, we get

$$\operatorname{supp}(f) \times \operatorname{supp}(\hat{f}) \quad \geq \quad N.$$

□

B.4 Cross-Validation

Consider the ℓ_1-regularized least squares functional

$$\Lambda_\lambda(x) \quad \triangleq \quad \|Ax - y\|_2^2 + \lambda \|x\|_1. \tag{B.9}$$

In the cross-validation approach, we divide the measurements into two disjoint reconstruction and validation sets. For measurements $y \in \mathbb{C}^m$, define $\Omega_r \subset \{1, \ldots, m\}$ and $\Omega_v \subset \{1, \ldots, m\}$ to be the reconstruction and validation sets, respectively, such that $\Omega_r \cap \Omega_v = \emptyset$, $m_r = |\Omega|$, $m_v = |\Omega|$, and $m_r + m_v = m$. Let $y_r \in \mathbb{C}^{m_r}$ be the restriction of

y to the set Ω_r and similarly for $y_v \in \mathbb{C}^{m_v}$. Also, define the reconstruction measurement matrix $A_{\Omega_r} \in \mathbb{C}^{m_r \times n}$ to be the m_r rows of A according to index set Ω_r, and similarly for $A_{\Omega_v} \in \mathbb{C}^{m_v \times n}$.

Now, (2.20) is solved using only y_r for different values of $\lambda > 0$. Let us define

$$x^*_{r,\lambda} \stackrel{\circ}{=} \arg\min_{x \in \mathbb{C}^n} \|A_r x - y_r\|_2^2 + \lambda \|x\|_1 \tag{B.10}$$

to be the minimizer computed using only the reconstruction measurements. Then the preferred choice of λ is the one that minimizes the validation error

$$\|A_v x^*_{r,\lambda} - y_r\|_2.$$

One may choose the initial set of parameters $\lambda > 0$ in order to explore the Pareto front (see [2, Section 4.7.3]) by selecting finitely many values $\lambda \in (0, L)$ where $L > 0$. Once the optimal λ is determined, (2.20) can be solved with the entire set of measurements y and the optimal λ. The cross-validation approach is summarized below

1. Divide the m measurements, y, into m_r reconstruction and m_v validation samples, y_r and y_v, respectively. This will correspond to a new set of indices Ω_r and Ω_v for reconstruction and validation, respectively.

2. Choose multiple values of λ to solve the ℓ_1-regularization problem using only y_r. Let $x^*_{r,\lambda}$ denote the respective minimizers (B.10) for the different choices of λ.

3. For each of the minimizers in Step 2, compute the validation error $\|A_v x^*_{r,\lambda} - y_v\|_2$ using the validation samples y_v.

4. Choose λ such that the validation error in Step 3 is the smallest. Denote this choice by λ^*.

5. Solve the ℓ_1-regularization problem (2.20) with the entire set of measurements y to obtain the solution $x^*_{\lambda^*}$.

B.5 Proof of ℓ_1-regularization Convergence

The results in this section are based on the chapter on homotopy methods discussed in [32].

B.5.1 Preliminaries

Consider the ℓ_1 minimization problem

$$\min_{x\in\mathbb{R}^n} \|x\|_1 \quad \text{subject to} \quad Ax = y \tag{B.11}$$

and its unique solution x^*. Consider the ℓ_1-regularized least squares functional

$$F_\lambda(x) \quad = \quad \frac{1}{2}\|Ax - y\|_2^2 + \lambda\|x\|_1, \quad x \in \mathbb{R}^n.$$

Let x_λ be the minimizer of $F_\lambda(x)$. We will show that $\lim_{\lambda\to 0} x_\lambda = x^*$.[1] In order to prove this result, we will need to introduce the concept of a coercive function, followed by an important lemma about sequences of minimizers.

Definition B.1. *A function $F : K \to \mathbb{R}, K \subset \mathbb{R}^n$ is called coercive if the level sets $K_t \doteq \{x \in \mathbb{R}^n, F(x) \leq t\}$ are compact.*

A coercive function is a function that grows rapidly at the extremes. It can be shown that a function F is coercive if and only if

$$\lim_{\|x\|_2\to\infty} |F(x)| \quad = \quad +\infty.$$

For example, $F(x) \equiv 1$ is not coercive but $F(x) = x^T x$ is. With the definition of coercivity at hand, we now state the following lemma about sequences of minimizers.

[1] For the complex-valued case, we can rewrite the ℓ_1-minimization problem in terms of the real and imaginary parts and the proof will be analogous.

Lemma B.1. *Let $F_k : K \to \mathbb{R}, K \subset \mathbb{R}^n$ be a decreasing sequence of continuous function converging pointwise to $F : K \to \mathbb{R}$. That is,*

$$F_{k+1}(x) \quad \leq \quad F_k(x)$$

and

$$\lim_{k\to\infty} F_k(x) \quad = \quad F(x).$$

Assume that F is continuous and coercive and suppose that x_k minimizes F_k over K. Then the limit points of the sequence $\{x_k\}_{k\in\mathbb{N}}$ are the minimizers of F. Moreover, if the minimizer of F is unique, then x_k converges to it.

Proof. Let $\{z_k\}$ be a sequence converging to some $z \in K$. By continuity, pointwise convergence, and monotonicity, we have that

$$\lim_{k\to\infty} F_k(x_k) \quad \leq \quad \lim_{k\to\infty} F_k(z_k) \quad = \quad F(z) \tag{B.12}$$

since x_k minimizes F_k. Since z was abritrary, we have that

$$\inf_{z\in K} F(z) \quad \geq \quad \lim_{k\to\infty} F_k(x_k). \tag{B.13}$$

Since $F_k \geq F$ and F_k is a decreasing sequence, we have

$$\{x : F_k(x) \leq t\} \quad \subset \quad \{x : F(x) \leq t\}$$

for all t. The latter set is compact due to the coercivity of F. Since x_k minimizes $F_k(x)$, the sequence x_k is contained in a compact set. By the Bolzano-Weierstrass Theorem, there exists a converging subsequence $\{x_{k_j}\}$ which converges to one of the limit points x' of x_k.

Then, (B.12) and (B.13) gives

$$\begin{aligned} \inf_{z\in K} F(z) \;\leq\; F(x') &= \lim_{j\to\infty} F_{k_j}(x_{k_j}) \\ &= \lim_{k_j\to\infty} \left(\min_{z\in K} F_{k_j}(z)\right) \;\leq\; \inf_{z\in K} F(z). \end{aligned}$$

This means that $F(x') = \inf_{z\in K} F(x)$ so that x' is a minimizer of F on K. This must be true for all limit points, and in particular, if the limit is unique, all subsequences of x_k must converge to the unique minimizer. □

B.5.2 Main Proof

We are now ready to prove Proposition 2.3, which states that $\lim_{\lambda\to 0} x_\lambda = x^*$. Pick an arbitrary sequence $(\lambda_n)_{n\in\mathbb{N}} \subset (0,\infty)$ that converges monotonically to 0. It suffices to show that $\lim_{n\to\infty} x_{\lambda_n} = x^*$. For simplicity, we write $x^n = x_{\lambda_n}$. Since x^n is the minimizer of F_{λ_n}, we get

$$\|x^n\|_1 \;\leq\; \frac{1}{\lambda_n} F_{\lambda_n}(x^n) \;\leq\; \frac{1}{\lambda_n} F_{\lambda_n}(x^*) \;=\; \|x^*\|_1,$$

where the last equality follows since $Ax^* = y$. Since this is true for all $n \in \mathbb{N}$, we may restrict ourselves to the compact set

$$K \;\doteq\; \{x \in \mathbb{R}^n, \|x\|_1 \leq \|x^*\|_1\},$$

which is the ℓ_1-ball of radius $\|x^*\|_1$. The boundedness follows since $x^n \in K$. Let us introduce the function $F : K \to \mathbb{R}, F(x) = \|Ax - y\|_2^2$, which attains the minimum at $Ax = y$. It is easy to see that $F(x)$ is coercive (in fact, since K is compact, this follows by definition).

Let F_n denote the function F_{λ_n} restriced to the set K. Since λ_n's are decreasing and F_n is a continuous function of λ, F_n's are monotonically decreasing in n and converge to F. Thus, by Lemma B.1, any limit point of the sequence x^n is a minimizer x' of F, which means that $Ax' = y$. This means that x' is a feasible solution to (B.11) where x^* is the

minimizer. Thus, $\|x'\|_1 \geq \|x^*\|$. On the other hand, by the compactness of K and the continuity of the norm, we have $\|x'\|_1 = \lim_{n\to\infty} \|x^n\|_1 \leq \|x^*\|_1$. Hence, we must have

$$\|x'\|_1 \quad = \quad \|x^*\|_1,$$

and so every limit point of x^n is a minimizer of (B.11). If the minimum x^* is unique, then this argument shows that any sequence x^n will converge to x^*. □

B.6 Edge Interpolation

We show how one can utilize high resolution edge information for lower resolution images. First, consider the high resolution edge operator $\mathcal{E}_H : \mathbb{C}^{N_H \times N_H} \to \mathbb{C}^{2N_H \times N_H}$, and the low resolution edge operator $\mathcal{E} : \mathbb{C}^{N\times N} \to \mathbb{C}^{2N\times N}$ where $N_H > N$. To determine the approximate edge measurements for the lower resolution image, we will essentially perform a bilinear interpolation on $\mathcal{E}_H\hat{Y}$. Define the high resolution one-dimensional N_H point grid

$$x_j \quad = \quad \frac{j}{N_H}, \quad j = 1, \ldots, N_H.$$

Similarly, define the lower resolution one-dimensional N point grid as

$$\tilde{x}_j \quad = \quad \frac{j}{N}, \quad j = 1, \ldots, N.$$

Denote the i^{th} row of $\mathcal{E}_H\hat{Y}$ by $\{f_{i,j}\}_{j=1}^{N}$. Then, the pairs $\{(x_j, f_{i,j}\}_{j=1}^{N_H}$ can be interpreted as discrete samples of a one-dimensional function $f_i(x)$ at the high resolution grid points x_j, and we can perform a linear interpolation of $f_i(x)$ at the courser grid points $\{\tilde{x}_j\}_{j=1}^{N}$. Let us perform this linear interpolation along each row. Afterwards, for each fixed column, we perform the analogous linear interpolation on the first N_H rows, and then on the last N_H rows separately (remember that each row of $\mathcal{E}_H\hat{Y}$ is of length $2N_H$). Once the bilinear interpolation is complete, we will have edge measurements, denoted by $\mathcal{I}_N(\mathcal{E}_H\hat{Y}) \in \mathbb{C}^{2N\times N}$, on the lower resolution grid.

B.7 Nonlinear Conjugate Gradient Method

B.7.1 Approximation to the Gradients

Consider the general unconstrained ℓ_1-regularized least-square functional

$$F_\lambda(x) \quad \triangleq \quad \|Ax - y\|_2^2 + \lambda \|Sx\|_1 \tag{B.14}$$

where $A \in \mathbb{C}^{m \times n}, S \in \mathbb{C}^{n \times n}, y \in \mathbb{C}^m$ and let x_λ^* be the minimizer. The penalty parameter $\lambda > 0$ can be determined via cross-validation or by simply choosing a λ such that $\|Ax_\lambda^* - y\|_2 < \varepsilon$, for some given tolerancce $\varepsilon > 0$. We propose solving this problem via the Fletcher-Reeves non-linear conjugate gradient method with a backtracking line search and gradient restarts.

This method requires the calculation of the gradient $\nabla F_\lambda(x)$. First, because $x \in \mathbb{C}^n$ is complex valued, we use the approach discussed in [20, Chapter 3]. Second, because the ℓ_1-norm is non-differentiable at zero, we use the approximation

$$|z| \quad \approx \quad \sqrt{\overline{z}z + \mu}$$

where $z \in \mathbb{C}$ and $\overline{z}$ is the complex conjugate of z, so that

$$\frac{d|z|}{dz} \quad = \quad \frac{z}{\sqrt{\overline{z}z + \mu}}.$$

Then, the partial derivatives of $\|Sx\|_1$ in (B.14) are

$$\begin{aligned} \frac{\partial}{\partial \overline{x}_k} \|Sx\|_1 &= \sum_{i=1}^{n} \frac{\partial}{\partial \overline{x}_k} \sqrt{|\Sigma_j S_{ij} x_j|^2 + \mu} \\ &= \sum_{i=1}^{n} \frac{S_{ik}^*}{2\sqrt{(\Sigma_j \overline{S}_{ij} \overline{x}_j)(\Sigma_l S_{il} x_l) + \mu}} (\Sigma_l S_{il} x_l) \end{aligned}$$

where S^* in the conjugate transpose of S. In matrix-vector form, we have

$$\nabla\|Sx\|_1 \quad = \quad S^*W^{-1}Sx \tag{B.15}$$

where $W \in \mathbb{C}^{n\times n}$ is a diagonal matrix with entries

$$W_{i,i} \quad = \quad 2\sqrt{(\Sigma_j \overline{S}_{ij}\overline{x}_j)(\Sigma_l S_{il}x_l) + \mu)}.$$

To show this, we have

$$\begin{aligned}(S^*W^{-1}Sx)_i &= \sum_{k,m,j=1}^{n} S^*_{ki}W^{-1}_{km}S_{mj}x_j \\ &= \sum_{k=1,j}^{n} S^*_{ki}w_k S_{kj}x_j \\ &= \sum_{k,j=}^{n} S^*_{ki}w_k S_{kj}x_j.\end{aligned}$$

Therefore, the gradient of $F_\lambda(x)$ can be approximated by

$$\nabla F_\lambda(x) \quad \approx \quad 2A^*(A^*x - y) + \lambda S^*W^{-1}Sx. \tag{B.16}$$

When the ℓ_1 regularization term $\|Sx\|_1$ is replaced by $\|X\|_{\mathrm{TV_{iso}}}$ or $\|X\|_{\mathrm{TV_{aniso}}}$, we have

$$\nabla\|X\|_{\mathrm{TV_{iso}}} \quad = \quad D_h^*W_{\mathrm{iso}}^{-1}D_hX + D_v^*W_{\mathrm{iso}}^{-1}D_vX \tag{B.17}$$

and

$$\nabla\|X\|_{\mathrm{TV_{aniso}}} \quad = \quad D_h^*W_{1,\mathrm{aniso}}^{-1}D_hX + D_v^*W_{2,\mathrm{aniso}}^{-1}D_vX, \tag{B.18}$$

where $W_{1,\mathrm{aniso}}, W_{2,\mathrm{aniso}}, W_{\mathrm{iso}} \in \mathbb{C}^{N\times N}$ are diagonal matrices with entries

$$\begin{aligned}(W_{1,\mathrm{aniso}})_{i,i} &= \sqrt{|\Sigma_j(D_h)_{ij}x_j|^2 + \mu}, \\ (W_{2,\mathrm{aniso}})_{i,i} &= \sqrt{|(\Sigma_j D_v)_{ij}x_j|^2 + \mu}, \\ (W_{\mathrm{iso}})_{i,i} &= \sqrt{|\Sigma_j(D_h)_{ij}x_j|^2 + |\Sigma_j(D_v)_{ij}x_j|^2 + \mu}.\end{aligned}$$

B.7.2 Non-linear Conjugate Gradient Algorithm

With the gradients properly defined, the non-linear conjugate gradient algorithm with gradient restarts is as follows.

INPUTS

$y \in \mathbb{C}^m$ measurements
$A \in \mathbb{C}^{m \times n}$ measurement matrix
$S \in \mathbb{C}^{n \times n}$ sparsifying transform
$\lambda > 0$ regularization parameter

PARAMETERS

TolGrad - stopping criteria for gradient
MaxInnerIter - stopping criteria for inner iterations
MaxOuterIter - stopping criteria for outer iterations
α, β - line search parameters

OUTPUTS

% Initialize
$k = 0; x =$ random;
% Iterations
for $i = 1, \ldots,$ MaxOuterIter
  $g_k = \nabla f(x)$
  $d_k = -g_k$
  while ($\|g_0\|_2^2 \geq$ TolGrad and $k <$ MaxInnerIter)
    %Backtracking line search to determine t
    while $(f(x_k + td_k) > f(x_k) + \alpha t \cdot \text{Real}(g_k^* d_k))\, \{t = \beta t\}$
    $x_{k+1} = x_k + td_k$
    $g_{k+1} = \nabla f(x_{k+1})$
    $\gamma = \dfrac{\|g_{k+1}\|_2^2}{\|g_k\|_2^2}$
    $d_{k+1} = -g_{k+1} + \gamma d_k$
    $k = k + 1$

Table B.1: Non-linear conjugate gradient algorithm

Bibliography

[1] R. K. Boel, M. R. James, and I. R. Petersen. Robustness and risk sensitive filtering. *IEEE Trans. on Auto. Control*, 3:451–461, 2002.

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2009.

[3] E. Candes and J. Romberg. *l1-Magic: Recovery of Sparse Signals via Convex Programming*.

[4] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52:489–509, 2006.

[5] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

[6] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 43(1):129–159, 2001.

[7] P. Dupuis and R. S. Ellis. *A Weak Convergence Approach to the Theory of Large Deviations*. John Wiley & Sons, New York, 1997.

[8] P. Dupuis, M. R. James, and I. R. Petersen. Robust properties of risk–sensitive control. *Math. Control Signals Systems*, 13:318–332, 2000.

[9] M. Eldred and L. Swiler. Efficient algorithms for mixed aleatory-epistemic uncertainty quantification with applications to radiation-hardened electronics. Technical report, Sandia National Laboratories, 2009.

[10] M. Fornasier. *Theoretic Foundations and Numerical Methods for Sparse Recovery*. Radon Series on Computational and Applied Mathematics. De Gruyter, 2010.

[11] M. Fornasier and H. Rauhut. Compressive Sensing. In O. Scherzer, editor, *Handbook of Mathematical Methods in Imaging*, pages 187–228. Springer, 2011.

[12] S. Foucart. A note on guaranteed sparse recovery via ℓ_1-minimization. *Applied and Computational Harmonic Analysis*, 29(1):97–103, 2010.

[13] A. Gelb and E. Tadmor. Detection of edges in spectral data II. nonlinear enhancement. *SIAM Journal on Numerical Analysis*, 38:1389–1408.

[14] A. Gelb and E. Tadmor. Detection of edges in spectral data. *Applied and Computational Harmonic Analysis*, 7:101–135, 1999.

[15] A. Gelb and E. Tadmor. Spectral reconstruction of piecewise smooth functions from their discrete data. *Mathematical Modelling and Numerical Analysis*, 36(2):155–175, 2002.

[16] A. Gelb and E. Tadmor. Detection of edges in spectral data iii - refinement of the concentration method. *Journal of Scientific Computing*, 36:1–43, 2008.

[17] M. A. Griswold, R. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, B. Kiefer, and A. Haase. Generalized autocalibrating partially parallel acquisitions (GRAPP). *Magnetic Resonance in Medicine*, 47(6):1202–1210, June 2002.

[18] W. Guo and W. Yin. Edgecs: Edge guided compressive sensing reconstruction. In *Visual Communications and Image Processing*, 2010.

[19] J. S. Hesthaven, D. Gottlieb, and S. Gottlieb. *Spectral Methods for Time-Dependent Problems*. Cambridge University Press, 2000.

[20] D. Johnson. *Statistical signal processing*. Rice University, 2012.

[21] K. F. King. Combining compressed sensing and parallel imaging. *Proceedings of the International Society on Magnetic Resonance in Medicine*, 16, 2008.

[22] R. Liu, F. M. Sebert, Y. Zou, and L. Ying. Sparsesense: Randomly-sampling parallel imaging using compressed sensing.

[23] M. Lustig, D. Donoho, and J.M. Pauly. Sparse MRI: the application of compressed sensing for rapid MR imaging. *Mangetic Resonance in Medicine*, 58:1182–1195, 2007.

[24] O. P. Le Maitre and O. M. Knio. *Spectral Methods for Uncertainty Quantification*. Springer, New York, 2010.

[25] D. Needell and R. Ward. Stable image reconstruction using total variation minimization. *arXiv*, April 2012.

[26] D. Nishimura. *Principles of Magnetic Resonance Imaging*. 2010.

[27] J. Nocedal and S. Wright. *Numerical Optimization*. Springer, 1st edition edition, 1999.

[28] S. Ogawa, D. W. Tank, R. Menon, J. M. Ellerman, S. G. Kim, H. Merkle, and K. Ugurbil. Intrinsic signal changes accompanying sensory stimulation: functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences*, 89:5951–5955, 1992.

[29] V. Patel, G. R. Easley, D. M. Healy, and R. Chellappa. Compressed synethic aperture radar. *IEEE Journal of Selected Topics in Signal Processing*, 4(2):244–254, April 2010.

[30] V. Patel, R. Maleh, A. Gilbert, and R. Chellappa. Gradient-based image recovery methods from incomplete fourier measurements. *IEEE Transactions on Image Processing,* accepted for publication, 2011.

[31] K. P. Preussmann, M. Weiger, M. Scheidegger, and P. Boesiger. SENSE: Sensitivity encoding for fast mri. *Magnetic Resonance in Medicine*, 42:952–962, 1999.

[32] H. Rauhut. Monograph on compressed sensing. Unpublished manuscript, Chapter on Homotopy Methods.

[33] H. Rauhut. Random sampling of sparse trigonometric polynomials. *Applied and Computational Harmonic Analysis*, 22:16–42, 2009.

[34] H. Rauhut. *Compressive Sensing and structured random matrices.* Radon Series on Computational and Applied Mathematics, 2010.

[35] L. A. Shepp and B. F. Logan. The fourier reconstruction of a head section. *IEEE Transactions on Nuclear Science*, 21(3):21–43, 1974.

[36] T. Strohmer and R. W. Heath. Grassmannian frames with applications to coding and communication. *Applied and Computational Harmonic Analysis*, 14(3):257–275, 2003.

[37] J. Tukey and J. Cooley. An algorithm for the machine calculation of complex fourier series. *Mathematics of Computation*, 19(90):297–301, 1965.

[38] S. R. S. Varadhan. *Large Deviations and Applications.* CBMS-NSF Regional Conference Series in Mathematics. SIAM, Philadelphia, 1984.

[39] Y. Wang, J. Yang, W. Yin, and Y. Zhang. A new alternating minimization algorithm for total variation image reconstruction. *SIAM Journal on Imaging Science*, 1(3):348–272, 2008.

[40] D. Xiu. Efficient collocational approach for parametric uncertianty analysis. *Journal of Computational Physics*, 2:293–309, 2007.

[41] D. Xiu. Fast numerical methods for stochastic computations. *Comm. in Computational Physics*, 5:242–272, 2009.

[42] D. Xiu and J. S. Hesthaven. High-order collocation methods for differential equations with random inputs. *Soc. for Industrial and App. Mathematics*, 27:1118–1139, 2005.

[43] D. Xiu, J. Jakeman, and M. Eldred. Numerical approach for quantification of epistemic uncertainty. *Comm. in Computational Physics*, 229:4648–4663, 2010.

[44] D. Xiu and G. Karniadakis. The Weiner-Askey polynomial chaos for stochastic differential equations. *SIAM Journal of Sci. Comp.*, 24:619–644, 2002.