2008

REGULARIZATION METHODS FOR ILL-POSED POISSON IMAGING

N'Djekornom Dara Laobeul
The University of Montana

REGULARIZATION METHODS FOR ILL-POSED POISSON IMAGING PROBLEMS

by

N'Djekornom Dara Laobeul

B.S. Université De Ouagadougou, BURKINA-FASO, 1999
M.S. Université De Ouagadougou, BURKINA-FASO 2001

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy
in Mathematical Sciences

The University of Montana
Missoula, Montana

May 2008

Approved by:

Dr. David A. Strobel, Dean
Graduate School

Dr. Johnathan Bardsley, Chair
Mathematical Sciences

Dr. Thomas Tonev
Mathematical Sciences

Dr. Leonid Kalachev
Mathematical Sciences

Dr. Brian Steele
Mathematical Sciences

Dr. Jesse Johnson
Computer Sciences

Laobeul, N'djekornom D. Ph.D., May 2008 Mathematics

Regularization Methods For Ill-Posed Poisson Imaging Problems

Committee Chair: Johnathan M. Bardsley, Ph.D.

The noise contained in images collected by a charge coupled device (CCD) camera is predominantly of Poisson type. This motivates the use of the negative logarithm of the Poisson likelihood in place of the ubiquitous least squares fit-to-data. However, if the underlying mathematical model is assumed to have the form $z = Au$, where A is a linear, compact operator, the problem of minimizing the negative log-Poisson likelihood function is ill-posed, and hence some form of regularization is required. In this work, it involves solving a variational problem of the form

$$u_\alpha \stackrel{\text{def}}{=} \arg\min_{u \geq 0} \ell(Au; z) + \alpha J(u),$$

where ℓ is the negative-log of a Poisson likelihood functional, and J is a regularization functional. The main result of this thesis is a theoretical analysis of this variational problem for four different regularization functionals. In addition, this work presents an efficient computational method for its solution, and the demonstration of the effectiveness of this approach in practice by applying the algorithm to simulated astronomical imaging data corrupted by the CCD camera noise model mentioned above.

Acknowledgements

At this point, being near the completion of my Ph.D. I would like to acknowledge and to be grateful to all the persons who have helped me to accomplish this work. Above all, thanks to The Lord Jesus for being my shepherd. He has put some people in my path, I would like to take this time to tell them how much I appreciate all they have done for me. First of all, I thank my advisor Johnathan Bardsley for his guidance and constant support. Thanks to my parents; Laobeul Goreu Dara and Nangtolatam Suzanne Dara for praying for me, teaching me the love of education and their support. I would like to thank all the members of my committee: Dr Leonid Kalachev, Dr Thomas Tonev, Dr Brian Steele, Dr Johnson Jesse. I would also like to thank Dr Emily Stone for the computer she bought me, Dr Jenifer halfpap and Todd Riggs for lending me some of the books I used for my research. I am grateful to Aaron Luttman for reading my dissertation and giving me some feed back. I would also like to acknowledge Dr Anatoly Yagola, whose course on inverse problem influenced my advisor to choose my dissertation project. I cannot forget my brothers and sisters: N'Djekoulmem Dara, Gamalao Dara, Nom-Osso Walbabet, Guedem Dara, Remadji Dara for being there for me whenever I need them. I am so grateful to Cherith, my beloved wife, for her caring love, as well as to her family. I would especially like to thank Laurel Smith, my mother in law, Amy Michel, Gabriel Ansah, Guedem Dara and David Yao for baby sitting Maye, allowing me to put countless hours in the pursuit of my degree. The people from the Graduate School and FSSS (foreign Student and Scholar Service) have allowed me to keep my student status; I would like to say my gratitude to Mary Kamensky, Dr David Strobel and Effie Koehn. All my gratitude to the Cates, they have done so much for me; David and Rosalie, thank you from the bottom of my heart. My gratitude also goes to Ahmed Diallo, Sico Yao and Ouedraego Thierry for hosting me, I extend my gratitude to the whole African community in Missoula. The Granga, the Nadjita and the Nguelkodingar families should be sure that I will never forget what they have done for me, as well as Sazoulang Douh, Telbet Jude, Ngargodje Ngarasta, Ndoudalmbaye Constant, Alex Laba, Ndoudjalbaye Benjamin, Sylva Lombaye, Jean-Paul Koualoa, Djounfoune Dougaye, Olivier Patale, Frederic Telbet, Yobombe Théophile and Mbaïnguimam. I would like to say to "ma tantie preferée" Angele Mbailao, my uncle Mbailao Jonas, my aunt Sarah Goltob and her husband Ngaressem Goltob, my aunt Esther Nokolo and her husband; we did it, thank you so much to all of you. In addition I would like to thank all the members of The UM international Toastmasters Club for helping me with my communication skills. Lastly, but not the least, I would like to thank Dr Jeffrey Adams and his team, and Dr Beatty for taking care of my health. I dedicate this dissertation to May-Lyric Smith, Olivia Smith and Maye N'Djekornom Ohitika Dara.

Contents

List of Figures

Chapter 1

Introduction

The following problem is very common in imaging science: given a blurred, noisy $N \times N$ image array $\mathbf{z}$, obtain an estimate of the underlying $N \times N$ true object array $\mathbf{u}_{\text{exact}}$ by approximately solving a linear system of the form

$$\mathbf{z} = \mathbf{A}\mathbf{u}. \tag{1.1}$$

Here $\mathbf{z}$ has been column stacked so that it is $N^2 \times 1$, and $\mathbf{A}$ is a known $N^2 \times N^2$ ill-conditioned matrix.

The focus of this work is on astronomical imaging, in which case $\mathbf{A}$ is the blurring matrix and $\mathbf{z}$ is an image of an object $\mathbf{u}_{\text{exact}}$ in outer space collected by a charge couple device (CCD) camera. The CCD camera was invented in 1969 by Willard Boyle and George E. Smith at AT & T Bell Labs [1]. It has the ability to receive charge via the photoelectric effect, allowing for the creation of electronic images. A CCD camera consists of a grid of pixels, each of which detects roughly 70% of incident light (photographic film captures only about 2%). At each pixel an electric charge accumulates that is proportional to the amount of light it receives. This charge is converted into voltage, digitized, and stored in memory.

The collected image $\mathbf{z}$ is blurred due to the fact that the light from the object being viewed has travelled through layers in the earth's atmosphere with variable degrees of refraction. Furthermore, diffractive blurring occurs due to the finite aperture of the telescope. This process is represented by the picture in Figure 1.1.

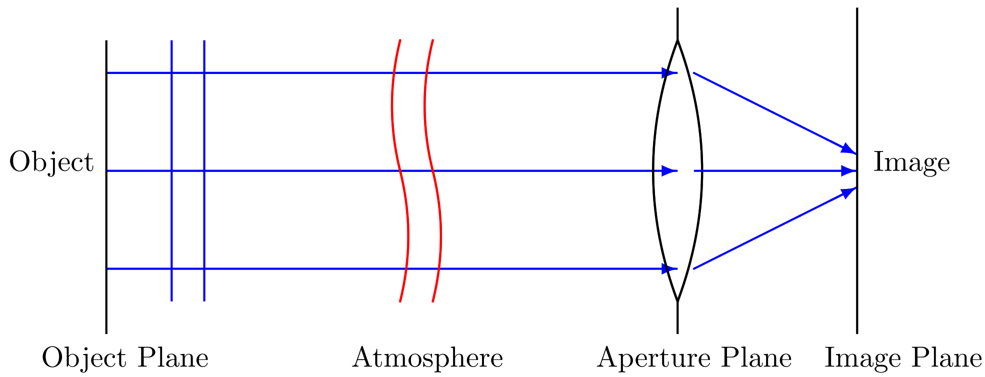

Figure 1.1: Blurring process schematic.

In practice, the noise in the image $\mathbf{z}$ is random. Thus $\mathbf{z}$ is a realization of a random vector $\hat{\mathbf{z}}$. A statistical model for $\hat{\mathbf{z}}$ is given by (c.f. [31])

$$\hat{\mathbf{z}} \sim \text{Poiss}(\mathbf{Au}_{\text{exact}}) + \text{Poiss}(\gamma \cdot \mathbf{1}) + N(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{1.2}$$

Here $\mathbf{1}$ is an $N^2 \times 1$ vector of all ones, and $\mathbf{I}$ is the $N^2 \times N^2$ identity matrix. (1.2) means that each element $\hat{z}_i$ of the vector $\hat{\mathbf{z}}$, has a mixture distribution, it is a random variable with distribution

$$\hat{z}_i = \frac{1}{3}\left(\text{n}_{\text{obj}}(i) + \text{n}_0(i) + g(i)\right), \quad i = 1, \dots, N^2, \tag{1.3}$$

where

- $\text{n}_{\text{obj}}(i)$ is the number of object dependent photoelectrons measured by the ith detector in the CCD array. It is a Poisson random variable with Poisson parameter $[\mathbf{Au}_{\text{exact}}]_i$.
- $\text{n}_0(i)$ is the number of background photoelectrons, which arise from both natural and artificial sources, measured by the ith detector in the CCD array. It is a Poisson random variable with a fixed positive Poisson parameter γ.

- $g(i)$ is the so-called readout noise, which is due to random errors caused by the CCD electronics and errors in the analog-to-digital conversion of measured voltages. It is a Gaussian random variable with mean 0 and fixed variance σ^2.

The random variables $\mathrm{n}_{\mathrm{obj}}(i)$, $\mathrm{n}_0(i)$, and $g(i)$ are assumed to be independent of one another and of $\mathrm{n}_{\mathrm{obj}}(j)$, $\mathrm{n}_0(j)$, and $g(j)$ for $i \neq j$.

The image and object in Figure 1.2 illustrate the light emanating from the star cluster $\mathbf{u}_{\mathrm{exact}}$ on the left hand side, it travels through the earth's atmosphere, and the telescope's CCD camera collects the blurred noisy image $\mathbf{z}$ on the right.

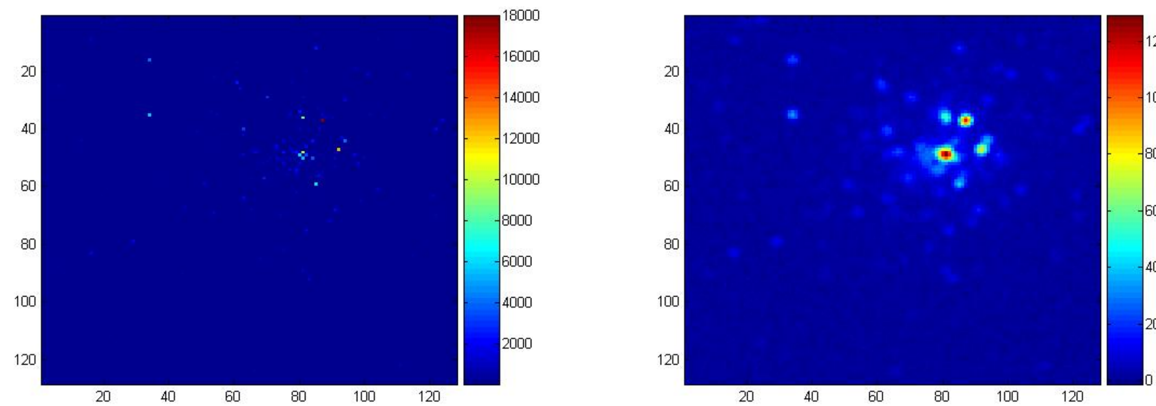

Figure 1.2: Illustration of image formation.

In order to obtain a workable maximum likelihood problem for obtaining estimates of $\mathbf{u}_{\mathrm{exact}}$, approximating (1.2) is necessary. As in [31], following [18, pp. 190 and 245], this is done using the approximation

$$N(\sigma^2, \sigma^2) \approx \mathrm{Poiss}(\sigma^2). \tag{1.4}$$

From this, together with the independence properties of the random variables in (1.3) it follows,

$$\begin{aligned}\hat{\mathbf{z}} + \sigma^2 \cdot \mathbf{1} &\sim \text{Poiss}(\mathbf{A}\mathbf{u}_{\text{exact}}) + \text{Poiss}(\gamma \cdot \mathbf{1}) + N(\sigma^2 \cdot \mathbf{1}, \sigma^2 \mathbf{I}) \\ &\approx \text{Poiss}(\mathbf{A}\mathbf{u}_{\text{exact}}) + \text{Poiss}(\gamma \cdot \mathbf{1}) + \text{Poiss}(\sigma^2 \cdot \mathbf{1}) \\ &\sim \text{Poiss}(\mathbf{A}\mathbf{u}_{\text{exact}} + \gamma \cdot \mathbf{1} + \sigma^2 \cdot \mathbf{1}). \end{aligned} \tag{1.5}$$

Assuming that (1.5) is the true statistical model, the maximum likelihood estimator of $\mathbf{u}_{\text{exact}}$, given a realization $\mathbf{z}$ from $\hat{\mathbf{z}}$ defined in (1.5), is the minimizer with respect to $\mathbf{u}$ of the negative-log Poisson likelihood functional

$$T_0(\mathbf{u}) \stackrel{\text{def}}{=} \sum_{i=1}^{N^2}([\mathbf{A}\mathbf{u}]_i + \gamma + \sigma^2) - \sum_{i=1}^{N^2}(z_i + \sigma^2)\log([\mathbf{A}\mathbf{u}]_i + \gamma + \sigma^2), \tag{1.6}$$

where $[\mathbf{A}\mathbf{u}]_i$ and z_i are the ith components of $\mathbf{A}\mathbf{u}$ and $\mathbf{z}$ respectively.

Before continuing, let address the question of whether or not using (1.5) to approximate (1.2) will have a negative effect on the accuracy of the estimates obtained by minimizing (1.6). For large values of σ^2 (the simulations suggest that $\sigma^2 > 30$ suffices), (1.4) is accurate, in which case (1.5) will well-approximate (1.2). This will also be true if the signal is sufficiently strong, since then the readout noise will have negligible effect. However, there are certainly instances in which (1.5) will not well-approximate (1.2), in particular, in regions of an image with very low light intensity. The likelihood that results, however, from the correct model (1.2) is non-trivial, as can be seen in [30]. In the sequel, the assumption that (1.5) is accurate is made.

Since $\mathbf{A}$ is an ill-conditioned matrix, computing the minimizer of (1.6) is an ill-posed problem. Thus regularization must be used. This involves solving a problem of the form

$$\arg\min_{\mathbf{u}\geq\mathbf{0}} \left\{ T_\alpha(\mathbf{u}) \stackrel{\text{def}}{=} T_0(\mathbf{u}) + \alpha J(\mathbf{u}) \right\}, \tag{1.7}$$

where $J(\mathbf{u})$ and $\alpha > 0$ are known as the regularization function and parameter, respectively, and minimization is subject to the constraint $\mathbf{u} \geq \mathbf{0}$ due to the fact that light intensity is non-negative.

The use of four different regularization functions is the main point of this thesis, also an important subject is that (1.7) can be motivated from a Bayesian perspective. The Bayesian perspective and the computational method used to solve (1.7) are the main topic of Chapter 4. The reader not interested in the theoretical arguments that constitute the remainder of this Chapter, as well as Chapter 2 and 3, should skip to Chapter 4 now.

Practically speaking, the computational problem (1.7) is very important, as it is the solution of (1.7) that will yield a regularized estimate of $\mathbf{u}_{\text{exact}}$. However, of equal importance is the relationship between (1.7) and its analogue in the function space setting, which results from the operator equation analogue of (1.1) given by

$$z(x) = Au(x) \stackrel{\text{def}}{=} \int_\Omega a(x,y)u(y)\,dy. \tag{1.8}$$

Here, $\Omega \subset \mathbb{R}^d$ is the closed bounded computational domain, and a which is known as the point spread function, is nonnegative and bounded, moreover it is reasonable to assumed that it is measurable with respect to the Lebesgue measure, and hence is in $L^2(\Omega \times \Omega)$. Although in the numerical experiments, $d = 2$, in the analysis, d can be any positive integer, unless otherwise specified, so that the results are as general as possible. Given the true image $u_{\text{exact}} \in L^2(\Omega)$, let define $z = Au_{\text{exact}}$. The functional analogue of (1.6) is then given by

$$T_0(Au; z+\gamma) = \int_\Omega \left((Au+\gamma+\sigma^2) - (z+\gamma+\sigma^2)\log(Au+\gamma+\sigma^2)\right)\,dx. \tag{1.9}$$

Note that if mid-point quadrature is used to discretize the integral in (1.8) and a collocation of indices is used to discretize the x-variable, the resulting numerical approximation of (1.9) will have the form (1.6). Thus in addition to stochastic errors, $\mathbf{z}$ also contains numerical dicretization errors. The functional analogue of (1.7) is then given by

$$u_\alpha = \arg\min_{u \geq 0} \left\{ T_\alpha(Au; z+\gamma) \stackrel{\text{def}}{=} T_0(Au; z+\gamma) + \alpha J(u) \right\}, \tag{1.10}$$

where J is now a regularization functional.

The main results of this thesis involve the analysis of problems of the form (1.10). In particular, although minimizing (1.9) is an ill-posed problem, problem (1.10) is well-posed if certain reasonable assumptions hold; that is, solutions of (1.10) exist, are unique, and depend continuously of the data z and the operator A. Noting that this is an important result because in practice there will always be errors in the measurements of z and A. Well-posedness implies that if these errors are small, the estimates will be near to those that would be obtained if the exact data z and operator A were known.

To state this mathematically, let introduce a sequence of operators equations

$$z_n(x) = A_n u(x) + \gamma \stackrel{\text{def}}{=} \int_\Omega a_n(x,y)u(y)\,dy + \gamma, \tag{1.11}$$

where $a_n \in L^2(\Omega \times \Omega)$ and corresponding minimization problems

$$u_{\alpha,n} = \arg\min_{u \geq 0} \left\{ T_\alpha(A_n u; z_n) \stackrel{\text{def}}{=} T_0(A_n u; z_n) + \alpha J(u) \right\}. \tag{1.12}$$

Proving that (1.10) is well posed amounts to showing that if $A_n \to A$ and $z_n \to z + \gamma$ then $u_{\alpha,n} \to u_\alpha$, where u_α is defined in (1.10). Another important result is that as A_n and z_n become arbitrarily close to A and $z+\gamma$, respectively, α_n can be chosen so that $u_{\alpha_n,n}$ becomes arbitrarily close u_{exact}; that is if $A_n \to A$ and $z_n \to z + \gamma$ there exists a positive sequence $\{\alpha_n\}$ such that $\alpha_n \to 0$ and $u_{\alpha_n,n} \to u_{\text{exact}}$. The main results of this thesis are proofs of these important properties for four different regularization functionals.

Chapter 2

Mathematical Preliminaries

In this chapter, the goal is to provide a mathematical background for the theoretical analysis of Chapter 3, which serves as the main result of this thesis.

This begins by proving a number of facts about Fredholm first kind integral operators, of which A, defined in (1.8), is an example.

2.1 Some Results Regarding Fredholm Integral Operators

Most functions will be element of $L^2(\Omega)$, $\|.\|_2$ and $\langle ., . \rangle$ denote the $L^2(\Omega)$ norm and inner product, respectively. The operator A will be assumed to be of Fredholm first kind type, i.e.

$$Au(x) \stackrel{\text{def}}{=} \int_\Omega a(x,y)u(y)dy, \tag{2.1}$$

where $a \in L^2(\Omega \times \Omega)$ and is measurable with respect to the Lebesgue measure. Then $A : L^2(\Omega) \longrightarrow L^2(\Omega)$.

Definition 2.1.1. *A linear operator $A : L^2(\Omega) \longrightarrow L^2(\Omega)$ is said to be compact if it maps the unit ball to a relatively compact set, that is, a set whose closure is a compact subset of $L^2(\Omega)$.*

The proofs in this section follow those in [4, chapter 4].

Theorem 2.1.2. *If $a \in L^2(\Omega\times\Omega)$, then (2.1) is a compact operator and $||A||_2 \leq ||a||_{L^2(\Omega\times\Omega)}$.*

The following lemma is necessary in order to prove Theorem 2.1.2.

Lemma 2.1.3. *If* $\{e_i : i \in I\}$ *is an orthonormal basis for* $L^2(\Omega)$ *and*

$$\phi_{ij}(x,y) = e_i(x)e_j(y) \tag{2.2}$$

for $i,\ j \in I$ *and* $x,\ y \in \Omega$ *then* $\{\phi_{ij} : i,j \in I\}$ *is an orthonormal basis for* $L^2(\Omega \times \Omega)$.

Proof. Since $\int_\Omega \int_\Omega |\phi_{ij}|^2 dxdy = \int_\Omega \int_\Omega |e_i(x)e_j(y)|^2 dxdy = ||e_i||_2^2 ||e_j||_2^2 = 1$, $\phi_{ij} \in L^2(\Omega \times \Omega)$. Moreover, if $(k,l) \neq (i,j)$ then

$$\begin{aligned} \langle \phi_{kl}, \phi_{ij} \rangle &= \int_\Omega \int_\Omega \phi_k(x)\phi_l(y)\phi_i(x)\phi_j(y)dxdy \\ &= \int_\Omega \left(\int_\Omega \phi_k(x)\phi_i(x)dx \right) \phi_l(y)\phi_j(y)dy \\ &= \langle \phi_k, \phi_i \rangle \langle \phi_l, \phi_j \rangle \\ &= 0. \end{aligned}$$

Therefore $\{\phi_{ij}\}$ is an orthonormal set.

To see that $\{\phi_{ij}\}$ is a basis for $L^2(\Omega \times \Omega)$, let choose $\phi \in L^2(\Omega \times \Omega)$ and define $\phi_y(x) = \phi(x,y)$. Then $\phi_y \in L^2(\Omega)$, and hence, $f_i(y) = \langle e_i, \phi_y \rangle = \int_\Omega \phi(x,y)e_i(x)dx$ is well defined. Moreover

$$\begin{aligned} ||f_i||_2^2 &= \sum_j |\langle e_j, f_i \rangle|^2 \\ &= \sum_j \left| \int_\Omega f_i(y)e_j(y)dy \right|^2 \\ &= \sum_j \left| \int_\Omega \int_\Omega \phi(x,y)e_i(x)e_j(y)dxdy \right|^2 \\ &= \sum_j \left| \int_\Omega \int_\Omega \phi(x,y)\phi_{ij}(x,y)dxdy \right|^2 \\ &= \sum_j |\langle \phi, \phi_{ij} \rangle|^2 . \end{aligned}$$

Thus, if ϕ is orthogonal to ϕ_{ij} for all i and j, $f_i = 0$ for all i and hence $\phi_y = 0$ in $L^2(\Omega)$, implying $\phi = 0$. From this it follows immediately that $\{\phi_{ij}\}$ is a basis for $L^2(\Omega \times \Omega)$. □

Now let prove Theorem 2.1.2.

Proof. The first step is to prove that $||A||_2 \leq ||a||_{L^2(\Omega\times\Omega)}$. Let $\{e_i\}$ and $\{\phi_{ij}\}$ be as in the previous lemma. Then

$$\begin{aligned} ||a||^2_{L^2(\Omega\times\Omega)} &= \sum_{i,j} |\langle a, \phi_{ij}\rangle|^2 \\ &= \sum_{i,j} \left| \int_\Omega \int_\Omega a(x,y) e_i(x) e_j(y) dx dy \right|^2 \\ &= \sum_{i,j} \left| \int_\Omega \left[\int_\Omega a(x,y) e_j(y) dy \right] e_i(x) dx \right|^2 \\ &= \sum_{i,j} |\langle Ae_j, e_i\rangle|^2. \end{aligned}$$

Moreover, if $u = \sum_j \alpha_j e_j \ \in \ L^2(\Omega)$, then $||u||_2^2 = \sum_j |\alpha_j|^2 < \infty$, and hence

$$|\langle Au, e_i\rangle|^2 = \left| \sum_j \alpha_j \langle Ae_j, e_i\rangle \right|^2 \leq \left(\sum_j |\alpha_j|^2 \right) \left(\sum_j |\langle Ae_j, e_i\rangle|^2 \right),$$

implies

$$||Au||_2^2 = \sum_i |\langle Au, e_i\rangle|^2 \leq ||a||^2_{L^2(\Omega\times\Omega)} ||u||_2^2,$$

which in turn implies $||A||_2 \leq ||a||_{L^2(\Omega\times\Omega)}$.

It remains to show that A is compact. Let J be the smallest (with respect to the number of its elements) set such that

$$a = \sum_{i,j\in J} \alpha_{ij}\phi_{ij}.$$

Such a representation exists by Lemma 2.1.3. If J is a finite set, then

$$\begin{aligned} Au(x) &= \int_\Omega a(x,y)u(y)dy \\ &= \int_\Omega \left(\sum_{i,j\in J} \alpha_{ij}\phi_{ij}(x,y) \right) u(y)dy \\ &= \sum_{i,j\in J} \alpha_{ij} \int_\Omega e_i(x)e_j(y)u(y)dy \\ &= \sum_{i,j\in J} \alpha_{ij} e_i(x)\langle e_j, u\rangle. \end{aligned}$$

Hence, A has finite rank and is therefore compact since it is bounded [33, chapter 2 page 17]. On the other hand, if J is not finite it is countably infinite, and hence

$$a = \sum_{i,j=1}^{\infty} \alpha_{ij}\phi_{ij}.$$

Let define

$$a_n = \sum_{i,j=1}^{n} \alpha_{ij}\phi_{ij}.$$

Then

$$||a_n - a||_{L^2(\Omega\times\Omega)} \longrightarrow 0 \text{ as } n \longrightarrow \infty.$$

Moreover, if

$$A_n u(x) \stackrel{\text{def}}{=} \int_\Omega a_n(x,y)u(y)dy,$$

then given that $||u||_2$ is finite,

$$|A_n u(x) - Au(x)| = \left| \int_\Omega \left[a_n(x,y) - a(x,y)\right] u(y)dy \right| \leq ||a_n - a||_{L^2(\Omega\times\Omega)} ||u||_2 \longrightarrow 0,$$

and hence, $||A_n - A||_2 \to 0$. Since the A_n's are finite dimensional, they are compact, implying that A is compact [4, chapter 4 page 41]. □

Next, is to show that A, defined in (2.1), has a singular value decomposition, which will be useful when proving that (2.1) is ill-posed.

Definition 2.1.4. *A singular system for a linear operator $A : L^2(\Omega) \longrightarrow L^2(\Omega)$ is a countable set of triples $\{u_j, s_j, v_j\}_j$ with the following properties:*

1. *the right singular vectors v_j form an orthonormal basis for Null(A)$^\perp$, where*

$$\textit{Null}(A) \stackrel{\text{def}}{=} \left\{u \ \in \ L^2(\Omega) \mid Au = 0\right\},$$

   and $\perp$ denotes the orthogonal complement;

2. *the left singular vectors u_j form an orthonormal basis for the closure of Range(A);*

3. *the singular values s_j are positive real numbers and are in nonincreasing order, i.e. $s_1 \geq s_2 \geq \cdots > 0$. Moreover, if Range(A) is infinite-dimensional, one has the additional property $\lim_{j\to\infty} s_j = 0$.*

4. *for all j*

$$Av_j = s_j u_j, \qquad \textit{and } A^* u_j = s_j v_j.$$

   Where A^* is such that

$$\langle Av, u\rangle = \langle v, A^* u\rangle \textit{ for all } u,\, v \ \in \ L^2(\Omega).$$

Next is the statement of an important theorem regarding compact operators.

Theorem 2.1.5. *If A is a compact operator then A has a singular system.*

Proof. The following proof is inspired by Exercise 2.9 of [33].

Let start by proving part 1. If A is compact then A^* is compact as well [4, chapter 4, page 41]. Hence A^*A is compact and self adjoint. By [4, Theorem 5.1] the existence a set of real eigenvalues and a corresponding set of orthonormal eigenfunctions is guaranteed, furthermore they form a basis for $\text{Null}(A^*A)^\perp$. Moreover, these eigenvalues are strictly positive. In addition [14, Theorem 4.9.A] states that $\text{Null}(A)^\perp = \text{Null}(A^*A)^\perp$. This completes the proof of part 1.

For part 2, 3 and 4 let define $\sigma_p(A^*A)$ to be the set of positive eigenvalues of A^*A. Then if $\lambda_i \in \sigma_p(A^*A)$ corresponds to the eigenfunction v_i and satisfies $\lambda_1 \geq \lambda_2, \cdots > 0$, it is possible to have $s_i = \sqrt{\lambda_i}$ and $u_i = \frac{1}{s_i} A v_i$. This yields

$$\langle u_i, u_j \rangle = \left\langle \frac{1}{s_i} A v_i, \frac{1}{s_j} A v_j \right\rangle = \frac{1}{\sqrt{\lambda_i \lambda_j}} \langle A^* A v_i, v_j \rangle = \frac{\lambda_i}{\sqrt{\lambda_i \lambda_j}} \langle v_i, v_j \rangle = \delta_{ij}.$$

Hence, $\{u_i\}$ is an orthonormal set. Moreover $u_i = \frac{1}{s_i} A v_i$ implies $A^* u_i = s_i v_i \neq 0$, and hence $\{u_i\}$ spans $\mathrm{Null}(A^*)^\perp = \overline{\mathrm{Range}(A)}$ [14]. Finally, noting that if Range(A) is infinite-dimensional, $\sigma_p(A^*A)$ will be infinite-dimensional as well, therefore by [4, corollary 7.8] $\lim_{i\to\infty} \lambda_i = 0 = \lim_{i\to\infty} s_i$. Thus $\{v_i, s_i, u_i\}$ forms a singular system for A. □

Following is the definition of a notion which will be of great importance in the analysis.

Definition 2.1.6. *Let $A : L^2(\Omega) \longrightarrow L^2(\Omega)$ be a nonzero operator. Then the equation*

$$Au = z \tag{2.3}$$

is said to be well-posed provided

1. *for each $z \in L^2(\Omega)$ there exists $u \in L^2(\Omega)$, called a solution, for which (2.3) holds;*
2. *the solution u is unique; and*
3. *the solution is stable with respect to perturbations in z and A. This mean that if $A_m u_m = z_m$ and $Au = z$, then $u_m \to u$ whenever $z_m \to z$ and $A_m \to A$.*

A problem that is not well-posed is said to be ill-posed.

The following result is fundamental to the work in this thesis.

Theorem 2.1.7. *Let $A : L^2(\Omega) \longrightarrow L^2(\Omega)$ be compact, then $Au = z$ is ill-posed.*

Proof. First, suppose dim(Range(A))$= \infty$. The goal is to show that Condition 3 of Definition 2.1.6 fails. Without lost of generality let suppose that $z = 0$; indeed, one can generalize for any nonzero z by using the linearity of A. Let $\{u_m, v_m, s_m\}_m$ be a singular system for A, and define $A_m = A$ and $z_m = s_m^r u_m \in$ Range(A) with $0 < r < 1$. Then $||z_m||_2 = ||s_m^r u_m||_2 = |s_m|^r ||u_m||_2$, implies $z_m \longrightarrow 0$ when $m \longrightarrow \infty$. However $z_m = s_m^r u_m = A(s_m^{r-1} v_m)$, with $||s_m^{r-1} v_m||_2 \longrightarrow \infty$ when $m \to \infty$. Therefore, $A_m \longrightarrow A$ and $z_m \longrightarrow 0$ while $u_m \stackrel{\text{def}}{=} s_m^{r-1} v_m \nrightarrow 0$. Thus Condition 3 of Definition 2.1.6 fails.

If dim(Range$(A)) < \infty$, by the first fundamental theorem on homomorphisms $L^2(\Omega)/\text{Null}(A) \cong \text{Range}(A)$, and hence, Null$(A)$ is not a trivial space. If $z \in L^2(\Omega)\backslash\text{Range}(A)$, then there is no u satisfying $Au = z$, and Condition 1 fails. Otherwise, if u satisfies $Au = z$, and $v \in$ Null(A) is nonzero then $A(u + v) = z$, and Condition 2 fails. □

However, in this thesis, in place of the operator equation $Au = z$, the task is to study the variational problem

$$u = \arg\min_{u \geq 0} T_0(Au; z + \gamma), \tag{2.4}$$

where $A : L^2(\Omega) \to L^2(\Omega)$ is a nonsingular operator satisfying $Au \geq 0$ whenever $u \geq 0$, $z \in L^\infty(\Omega)$ is nonnegative and $\gamma > 0$ is a fixed constant, and T_0 is as defined in (1.9). Accordingly to this variational problem, let now define what it means for (2.4) to be well-posed.

Definition 2.1.8. *The variational problem (2.4) is said to be well-posed provided*

1. *for each nonnegative $z \in L^\infty(\Omega)$, there exists a solution $u \in L^2(\Omega)$ of (2.4);*

2. *this solution is unique;*

3. *the solution is stable with respect to perturbations in $z + \gamma$ and A; that is, if*

$$u_{0,n} \stackrel{\text{def}}{=} \arg\min_{u \geq 0} T_0(A_n u; z_n),$$

where $A_n : L^2(\Omega) \to L^2(\Omega)$ satisfies $A_n u \geq 0$ whenever $u \geq 0$ and $z_n \in L^\infty(\Omega)$ is nonnegative, then $u_n \to u$ whenever $z_n \to (z + \gamma)$ in $L^\infty(\Omega)$ and $a_n \to a$ in $L^\infty(\Omega \times \Omega)$.

With this definition, the statement and proof of the analogue of Theorem 2.1.7 will be,

Theorem 2.1.9. *The variational problem (2.4) is ill-posed.*

The gradient and Hessian of T_0 is useful for proving Theorem 2.1.9. Also, the notion of Fréchet derivative is needed [33, page 22] for a better understanding,

Definition 2.1.10. *An operator $T : L^2(\Omega) \to L^2(\Omega)$ is said to be Fréchet differentiable at $u \in L^2(\Omega)$ if and only if there exists $T'(u)$, element of the set of linear operator from $L^2(\Omega)$ to itself, called the Fréchet derivative of T at u, for which*

$$T(u+h) = T(u) + T'(u)h + o(||h||_2) \tag{2.5}$$

The following identity, stated in Proposition 2.34 of [33], is used to derive the gradient:

$$\frac{d}{d\tau} T_0(u+\tau h)\bigg|_{\tau=0} = \langle \nabla T_0(u), h \rangle. \tag{2.6}$$

As a composite function of the natural logarithm and $Au + \gamma + \sigma^2$, T_0 is twice Fréchet differentiable. Moreover, since $A(u+\tau h) - (z+\gamma+\sigma^2)\log(A(u+\tau h)+\gamma+\sigma^2)$ is twice differentiable with respect to τ, its first and second derivatives are bounded over Ω. Hence the Lebesgue Dominated Convergence Theorem implies

$$\begin{aligned}
\frac{d}{d\tau} T_0(A(u+\tau h); z+\gamma)\bigg|_{\tau=0} &= \frac{d}{d\tau}\int_\Omega ((Au + \tau Ah + \gamma + \sigma^2) \\
&\quad -(z+\gamma+\sigma^2)\log(Au+\tau Ah+\gamma+\sigma^2))\, dx\Big|_{\tau=0} \\
&= \int_\Omega \left(1 - \frac{(z+\gamma+\sigma^2)}{Au+\tau Ah+\gamma+\sigma^2}\right) Ah\, dx\bigg|_{\tau=0} \\
&= \left\langle \left(1 - \frac{(z+\gamma+\sigma^2)}{Au+\gamma+\sigma^2}\right), Ah \right\rangle \\
&= \left\langle A^*\left(\frac{Au-z}{Au+\gamma+\sigma^2}\right), h \right\rangle.
\end{aligned}$$

Thus from (2.6) it is obvious that

$$\nabla T_0(Au; z+\gamma) = A^*\left(\frac{Au-z}{Au+\gamma+\sigma^2}\right). \tag{2.7}$$

In the same fashion, let compute the Hessian using the relation [33, Definition 2.40]

$$\frac{d^2}{d\rho d\tau} T_0(A(u+\tau h+\rho l); z+\gamma)\bigg|_{\rho,\tau=0} = \langle \nabla^2 T_0(u) l, h\rangle. \tag{2.8}$$

Once again using the Lebesque Dominated Convergence Theorem yields

$$\begin{aligned}
\frac{d^2}{d\rho d\tau} T_0(A(u+\tau h+\rho l); z+\gamma)\bigg|_{\tau,\rho=0} &= \int_\Omega \frac{d}{d\rho}\left(Ah - \frac{(z+\gamma+\sigma^2)Ah}{Au+\gamma+\sigma^2}\right) dx\bigg|_{\tau,\rho=0} \\
&= \int_\Omega \left(\frac{Ah(z+\gamma+\sigma^2)Al}{(Au+\tau Ah+\rho Al+\gamma+\sigma^2)^2}\right) dx\bigg|_{\tau,\rho=0} \\
&= \left\langle Al, \left(\mathrm{diag}\left(\frac{(z+\gamma+\sigma^2)}{(Au+\gamma+\sigma^2)^2}\right)\right) Ah\right\rangle \\
&= \left\langle A^*\left(\mathrm{diag}\left(\frac{(z+\gamma+\sigma^2)}{(Au+\gamma+\sigma^2)^2}\right)\right) Al, h\right\rangle,
\end{aligned}$$

where $\mathrm{diag}(v)$ is defined by $\mathrm{diag}(v)w = vw$. Thus, it results, from (2.8), that

$$\nabla^2 T_0(Au; z+\gamma) = A^*\left(\mathrm{diag}\left(\frac{z+\gamma+\sigma^2}{(Au+\gamma+\sigma^2)^2}\right)\right) A. \tag{2.9}$$

Noting, therefore, that

$$\nabla T_0(A_n u; z_n) = A_n^*\left(\frac{A_n u-(z_n-\gamma)}{A_n u+\gamma+\sigma^2}\right),$$

$$\nabla^2 T_0(A_n u; z_n) = A_n^*\left(\mathrm{diag}\left(\frac{z_n+\sigma^2}{(A_n u+\gamma+\sigma^2)^2}\right)\right) A_n.$$

Since z, $z_n \geq 0$, $\nabla^2 T_0(Au, z+\gamma)$ and $\nabla^2 T_0(A_n u, z_n)$ are positive semi-definite operators, implying $T_0(Au; z+\gamma)$ and $T_0(A_n u; z_n)$ are convex [33, Theorem 2.42] and hence have minimizer over $\mathcal{C} = \{u \in L^2(\Omega) \,|\, u \geq 0\}$. Moreover, since A is nonsingular, $T_0(Au; z+\gamma)$ is strictly convex and has unique minimizer u_{exact}, recall that it has been assumed $u_{\text{exact}} \geq 0$ and $Au_{\text{exact}} = z$.

The task is now to prove Theorem 2.1.9.

Proof. Since it has been assumed that A is nonsingular, Condition 1 and 2 hold. Hence, let show that Condition 3 fails. For this let proceed as in the first paragraph of the proof of Theorem 2.1.7, assuming once again without lost of generality, that $z = 0$. Let $\{u_m, v_m, s_m\}_m$ be a singular system for A, and define $A_m = A$ and $z_m = s_m^r u_m + \gamma$ with $0 < r < 1$. Since A is invertible, by our discussion above,

$$u_m = \arg\min_{u \geq 0} T_0(Au; z_m)$$

is unique and satisfies $Au_m = s_m^r v_m$, and hence, $u_m = s_m^{r-1} v_m$. Thus $z_m \to \gamma 1$ while $||u_m||_2 = ||s_m^{r-1} u_m||_2 \to +\infty$, implying that $u_m \nrightarrow 0$, and hence, Condition 3 fails. $\square$

2.2 Regularization Schemes and the Operator $R_\alpha(A, z+\gamma)$

Because (2.4) is ill-conditioned, well-posed methods for its approximate solution are important. Regularization represents the most common such approach. The aim of this thesis is to study a variational problems of the form

$$R_\alpha(A, z+\gamma) \stackrel{\text{def}}{=} \arg\min_{u \in \mathcal{C}} T_\alpha(Au; z+\gamma), \tag{2.10}$$

where $T_\alpha(Au; z+\gamma) = T_0(Au; z+\gamma) + \alpha J(u)$, where J and α are the regularization parameter and functional respectively. The task will be to show that R_α defines a regularization scheme for four different regularization functionals.

In order to define a regularization scheme, let suppose that there is an operator R_* such that for each z in Range(A) there is an unique $R_*(A, z+\gamma) \in L^2(\Omega)$ such that $A(R_*(A, z+\gamma)) = z$. For this work, given the assumption and therefore discussion,

$$R_0(Au; z+\gamma) = \arg\min_{u \in \mathcal{C}} T_0(Au; z+\gamma).$$

From [33] is the following definition

Definition 2.2.1. *$\{R_\alpha(A, z+\gamma)\}_{\alpha\in I}$ is a regularization scheme that converges to $R_*(A, z+\gamma)$ if*

1. *for each $\alpha \in I$, R_α is a continuous operator; and*
2. *given each z in Range(A), for any sequence $\{z_n\} \in L^\infty(\Omega)$ that converges to z and any sequence $\{A_n\}$ of compact operators on $L^2(\Omega)$ that converges to A, one can pick a sequence $\{\alpha_n\} \in I$ such that*

$$\lim_{n\to\infty} R_{\alpha_n}(A_n, z_n) = R_*(A, z+\gamma).$$

Chapter 3

Theoretical Analysis

In this chapter, the task is to prove that $R_\alpha(A, z+\gamma)$, defined by (2.10), is a regularization scheme for four regularization functionals J.

Some standard assumption are made in all of the following arguments. Let Ω be a closed bounded, convex domain with piecewise smooth boundary. Then $|\Omega| = \int_\Omega dx < \infty$. Let $|\cdot|$ denote the Euclidean norm in $\mathbb{R}^d$ and $\|\cdot\|_p$ the Banach space norm on $L^p(\Omega)$ for $1 \le p \le \infty$. Since Ω is bounded, $L^p(\Omega) \subset L^1(\Omega)$ for $p > 1$. It is assumed that the true solution $u_{\text{exact}} \ge 0$ and that $Au \ge 0$ for every $u \ge 0$, with A defined by

$$Au(x) = \int_\Omega a(x,y)u(y)dy,$$

where $a(x,y)$ is nonnegative, measurable with respect to the Lebesgue measure, and bounded on $L^2(\Omega\times\Omega)$. Then, as previously, the exact data satisfies $z = Au_{\text{exact}} \ge 0$. Furthermore, z is assumed to be bounded, hence $z \in L^\infty(\Omega)$. Note that these are reasonable assumptions since in practice the kernel function a is nonnegative and the image z has finite intensity at every point in the computational domain. These assumptions hold for the perturbed operator A_n, defined by

$$A_n u = \int_\Omega a_n(x,y)u(y)dy,$$

and data z_n. Additionally, let $\{a_n\}_{n\ge1} \subset L^\infty(\Omega)$ and $\{z_n\}_{n\ge1} \subset L^\infty(\Omega)$, $a_n \to a \in L^\infty(\Omega\times\Omega)$

and $z_n \to z + \gamma$ in $L^\infty(\Omega)$. Then, it follows that $||A_n - A||_1 \to 0$, $||A_n - A||_2 \to 0$, $||z_n - (z + \gamma)||_\infty \to 0$, and that the a_n's are uniformly bounded in $L^\infty(\Omega \times \Omega)$. All these facts are useful for what follow.

In order to simplify the notation in the arguments, let $T_\alpha(u)$ denotes $T_\alpha(Au; z+\gamma)$ and $T_{\alpha,n}(u)$ denotes $T_\alpha(A_n u; z_n)$ throughout the remainder of this analysis.

3.1 Tikhonov Regularization

The work in this subsection originally appeared in [8], though the modifications are significant. Let begin the analysis with the most standard regularization functional:

$$J(u) = \frac{1}{2}\|u\|^2. \tag{3.1}$$

Let define

$$\mathcal{C} = \{u \ \in \ L^2(\Omega) \ |u \geq 0\}. \tag{3.2}$$

Then the task is to show that

$$R_\alpha(A, z + \gamma) \stackrel{\text{def}}{=} \arg\min_{u \in \mathcal{C}} T_\alpha(Au; z + \gamma) \tag{3.3}$$

reward a regularization scheme as defined in Definition 2.2.1.

Let begin with definitions and results that will be needed in the later analysis. The following theorem, proven in [35], will be useful.

Theorem 3.1.1. *If $\mathcal{S}$ is a bounded set in $L^2(\Omega)$, every sequence in $\mathcal{S}$ has a weakly convergent subsequence in $L^2(\Omega)$, where u_n converges to u weakly (denoted $u_n \rightharpoonup u$) in $L^2(\Omega)$ if $\langle u_n - u, v\rangle \to 0$ for all $v \in L^2(\Omega)$.*

An important definition is what comes next.

Definition 3.1.2. *A functional $T : L^2(\Omega) \to \mathbb{R}$ is coercive if*

$$T(u) \to \infty \quad \textit{whenever} \quad \|u\|_2 \to \infty. \tag{3.4}$$

Recalling that a functional is said to be weakly lower continuous if

Definition 3.1.3. *T is weakly lower continuous if*

$$T(u_*) \leq \liminf T(u_n) \ \textit{whenever}\ u_n \rightharpoonup u_* \tag{3.5}$$

3.1.1 $R_\alpha(A, z + \gamma)$ is Well-Defined

The following theorem, which is similar to [33, Theorem 2.30] helps to prove the existence and uniqueness of the solutions of (2.10).

Theorem 3.1.4. *If $T : L^2(\Omega) \to \mathbb{R}$ is convex and coercive, then it has a minimizer over $\mathcal{C}$. If T is strictly convex, the minimizer is unique.*

Proof. Let $\{u_n\} \subset \mathcal{C}$ be such that $T(u_n) \to T_* \stackrel{\text{def}}{=} \inf_{u \in \mathcal{C}} T(u)$. Then, by (3.4), the sequence $\{u_n\}$ is bounded in $L^2(\Omega)$. By Theorem 3.1.1, this implies that $\{u_n\}$ has a subsequence $\{u_{n_j}\}$ that converges weakly to some $u_* \in \mathcal{C}$. Now, since T is convex, it is weakly lower semi-continuous [33, page 21], and hence,

$$T(u_*) \leq \liminf T(u_{n_j}) = \lim T(u_n) = T_*.$$

Thus u_* minimizes T on $\mathcal{C}$ and is unique if T is a strictly convex functional since $\mathcal{C}$ is a convex set. □

Corollary 3.1.5. *$R_\alpha(A, z + \gamma)$ is well-defined.*

Proof. The convexity of T_α follows from the convexity of T_0 (see the comments at the end of Chapter 2) and the strict convexity of $J(u)$.

For coercivity, noting that by Jensen's inequality and the properties of the function $x - c\log x$, for $c > 0$,

$$\begin{aligned} T_0(u) &\geq \|Au+\gamma+\sigma^2\|_1 - \|z+\gamma+\sigma^2\|_\infty \log \|Au+\gamma+\sigma^2\|_1, \\ &\geq \|z+\gamma+\sigma^2\|_\infty - \|z+\gamma+\sigma^2\|_\infty \log \|z+\gamma+\sigma^2\|_\infty. \end{aligned} \tag{3.6}$$

Since $z \geq 0$, T_0 is bounded below. The coercivity of $T_\alpha(u) = T_0(u) + \frac{\alpha}{2}\|u\|_2^2$ then follows immediately.

By Theorem 3.1.4, T_α has a unique minimizer in $\mathcal{C}$ and hence $R_\alpha(A, z+\gamma)$ is well-defined. □

3.1.2 $R_\alpha(A, z+\gamma)$ is Continuous

Let u_α be the unique solution of T_α over $\mathcal{C}$ given by Corollary 3.1.5. A similar analysis yields the existence and uniqueness of minimizers $u_{\alpha,n}$ of $T_\alpha(A_n u; z)$ in (1.12) for $\alpha \geq 0$ (recall that A_n and z_n satisfied the same assumptions as A and z).

The following theorem gives conditions that guarantee this result.

Theorem 3.1.6. *Let $u_{\alpha,n}$ be the unique minimizer of $T_{\alpha,n}$ over $\mathcal{C}$, and suppose that*

1. *for any sequence $\{u_n\} \subset L^2(\Omega)$,*

$$\lim_{n\to\infty} T_{\alpha,n}(u_n) = +\infty \quad \textit{whenever} \quad \lim_{n\to\infty} \|u_n\|_2 = +\infty; \tag{3.7}$$

2. *given $B > 0$ and $\epsilon > 0$, there exists N such that*

$$|T_{\alpha,n}(u) - T_\alpha(u)| < \epsilon \quad \textit{whenever} \quad n \geq N,\ \|u\|_2 \leq B. \tag{3.8}$$

Then $u_{\alpha,n}$ converges strongly to u_α in $L^2(\Omega)$.

Proof. Note that $T_{\alpha,n}(u_{\alpha,n}) \leq T_{\alpha,n}(u_\alpha)$. From this and (3.8), yields

$$\liminf T_{\alpha,n}(u_{\alpha,n}) \leq \limsup T_{\alpha,n}(u_{\alpha,n}) \leq T_\alpha(u_\alpha) < \infty. \tag{3.9}$$

Thus by (3.7), the $u_{\alpha,n}$'s are bounded in $L^2(\Omega)$. By Theorem 3.1.1, there exists a subsequence $\{u_{n_j}\}$ that converges weakly to some $\hat{u} \in L^2(\Omega)$. Furthermore, by the weak lower semicontinuity of T_α, (3.8), and (3.9) it follows

$$\begin{aligned} T_\alpha(\hat{u}) &\leq \liminf T_\alpha(u_{n_j}), \\ &= \liminf (T_\alpha(u_{n_j}) - T_{\alpha,n_j}(u_{n_j})) + \liminf T_{\alpha,n_j}(u_{n_j}), \\ &\leq T_\alpha(u_\alpha). \end{aligned}$$

By uniqueness of minimizers, $\hat{u} = u_\alpha$. Thus $\{u_{n_j}\}$ converges weakly to u_α, and hence, $u_{\alpha,n} \rightharpoonup u_\alpha$.

Let now prove strong convergence. This follows if $||u_{\alpha,n}||_2 \to ||u_\alpha||_2$ since $u_{\alpha,n} \rightharpoonup u_\alpha$ [4, Exercise 8, page 128]. By the weak lower semi-continuity of the norm $||u_\alpha||_2 \leq \liminf ||u_{\alpha,n}||_2$. However $T_{\alpha,n}(u_{\alpha,n}) \leq T_{\alpha,n}(u_\alpha)$ implies

$$\begin{aligned} \liminf ||u_{\alpha,n}||_2^2 &\leq ||u_\alpha||_2^2 + \frac{2}{\alpha}(T_0(u_\alpha) - \liminf T_0(u_{\alpha,n})) \\ &\leq ||u_\alpha||_2^2 \end{aligned}$$

Since T_0 is weakly, lower semi continuous. Hence $||u_{\alpha,n}||_2 \to ||u_\alpha||_2$ which in turn implies $u_{\alpha,n} \to u_\alpha$. □

A corollary of Theorem 3.1.6 is the continuity result for (3.3) that concludes this part.

Corollary 3.1.7. *$R_\alpha(A, z+\gamma)$ is continuous.*

Proof. It suffices to show that conditions *(i)* and *(ii)* from Theorem 3.1.6 hold. For condition *(i)*, note that the analogue of inequality (3.6) for $T_{0,n}$ is given by

$$T_{0,n}(u_{\alpha,n}) \geq \|z_n + \sigma^2\|_\infty - \|z_n + \sigma^2\|_\infty \log \|z_n + \sigma^2\|_\infty,$$

which has a lower bound for all n since $\|z_n - (z+\gamma)\|_\infty \to 0$ and $z \in L^\infty(\Omega)$ is nonnegative. Thus $T_{\alpha,n}(u_n) = T_{0,n}(u_n) + \frac{\alpha}{2}\|u_n\|_2^2 \to \infty$ whenever $\|u_n\|_2 \to \infty$, and hence, (3.7) is satisfied.

For condition *(ii)*, note that, using Jensen's inequality and the properties of the logarithm,

$$\begin{aligned}
|T_{\alpha,n}(u) - T_\alpha(u)| &= \left| \int_\Omega \left((A_n - A)u - (z_n + \sigma^2)\log(A_n u + \gamma + \sigma^2)\right)\, dx \right. \\
&\qquad \left. + \int_\Omega \left((z + \gamma + \sigma^2)\log(Au + \gamma + \sigma^2)\right)\, dx \right|, \\
&\leq \|A_n - A\|_1 \|u\|_1 \\
&\quad + \|z_n - (z+\gamma)\|_\infty \log(\|A_n\|_1 \|u\|_1 + (\gamma + \sigma^2)|\Omega|) \qquad (3.10) \\
&\quad + \|z + \gamma + \sigma^2\|_\infty \log \left\|(Au + \gamma + \sigma^2)/(A_n u + \gamma + \sigma^2)\right\|_1 .
\end{aligned}$$

By assumption, $\|A_n - A\|_1, \|z_n - (z+\gamma)\|_\infty \to 0$. Furthermore, by the Banach-Steinhaus Theorem, $\|A_n\|_1$ is uniformly bounded, and since $\|u\|_2$ is bounded by assumption, by Theorem 3.1.1 $\|u\|_1$ is bounded as well. Thus the first two terms on the right-hand side in (3.10) tend to zero as $n \to \infty$. For the third term note that

$$\left\| \frac{Au + \gamma + \sigma^2}{A_n u + \gamma + \sigma^2} - 1 \right\|_1 \leq \left\| \frac{1}{A_n u + \gamma + \sigma^2} \right\|_1 \|A_n - A\|_1 \|u\|_1,$$

which converges to zero since $\|1/(A_n u + \gamma + \sigma^2)\|_1$ is bounded and $\|A_n - A\|_1 \to 0$. Thus $\log(\|(Au + \gamma + \sigma^2)/(Au_n + \gamma + \sigma^2)\|_1) \to \log(1) = 0$, and hence

$$|T_{\alpha,n}(u) - T_\alpha(u)| \to 0. \qquad (3.11)$$

□

3.1.3 $R_\alpha(A, z+\gamma)$ is Convergent

The task in this section is to prove that $R_\alpha(A, z+\gamma)$ is convergent.

Theorem 3.1.8. $R_\alpha(A, z+\gamma)$ *is convergent.*

Proof. Suppose $\alpha_n \to 0$ at a rate such that

$$(T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n}))/\,\alpha_n \to 0. \qquad (3.12)$$

Then since $u_{\alpha_n,n}$ minimizes $T_{\alpha_n,n}$, it is obvious that

$$T_{\alpha_n,n}(u_{\alpha_n,n}) \leq T_{\alpha_n,n}(u_{\text{exact}}). \tag{3.13}$$

Since $\{z_n\}$ and $\{A_n\}$ are uniformly bounded and $A_n \to A$ in the $L^1(\Omega)$ operator norm, $\{T_{\alpha_n,n}(u_{\text{exact}})\}$ is a bounded sequence. Hence $\{T_{\alpha_n,n}(u_{\alpha_n,n})\}$ is bounded by (3.13).

Subtracting $T_{0,n}(u_{0,n})$ from both sides of (3.13) and dividing by α_n yields

$$\begin{aligned}(T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n}))/\alpha_n + \frac{1}{2}\|u_{\alpha_n,n}\|_2^2 \quad \leq \quad & (T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n}))/\,\alpha_n \\ & + \frac{1}{2}\|u_{\text{exact}}\|_2^2. \end{aligned} \tag{3.14}$$

By (3.12), the right-hand side is bounded, implying the left hand side is bounded. Since $T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n})$ is nonnegative, this implies that $\{u_{\alpha_n,n}\}$ is bounded in $L^2(\Omega)$.

Let now show that $u_{\alpha_n,n} \to u_{\text{exact}}$ in $L^2(\Omega)$ by showing that every subsequence of $\{u_{\alpha_n,n}\}$ contains a subsequence that converges to u_{exact}. Since $\{u_{\alpha_n,n}\}$ is bounded in $L^2(\Omega)$, by Theorem 3.1.1, each of its subsequences in turn has a subsequence that converges weakly in $L^2(\Omega)$. Let $\{u_{\alpha_{n_j},n_j}\}$ be such a sequence and $\hat{u}$ its weak limit. Then

$$\begin{aligned} T_0(\hat{u}) \quad = \quad & \int_\Omega (A(\hat{u} - u_{\alpha_{n_j},n_j}) + (A - A_{n_j})u_{\alpha_{n_j},n_j}\, dx \\ & + \int_\Omega (z_{n_j} - (z+\gamma)) \log(A\hat{u} + \gamma + \sigma^2))\, dx \\ & - \int_\Omega (z_{n_j} + \sigma^2) \log((A_{n_j} u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2))\, dx \\ & + T_{0,n_j}(u_{\alpha_{n_j},n_j}), \end{aligned}$$

which, as in previous arguments, yields

$$\begin{aligned} |T_{0,n_j}(u_{\alpha_{n_j},n_j}) \quad - \quad & T_0(\hat{u})| \\ \leq \quad & \left| \int_\Omega A\left(\hat{u} - u_{\alpha_{n_j},n_j}\right)\, dx \right| \end{aligned}$$

$$+\|z_{n_j}-(z+\gamma)\|_\infty \log(\|A\|_1\|\hat{u}\|_1+\gamma|\Omega|)$$
$$+\|z_{n_j}+\sigma^2\|_\infty \log\|(A_{n_j}u_{\alpha_{n_j},n_j}+\gamma+\sigma^2)/(A\hat{u}+\gamma+\sigma^2)\|_1$$
$$+\|A-A_{n_j}\|_1\|u_{\alpha_{n_j},n_j}\|_1.$$

Then

$$\|z_{n_j}-(z+\gamma)\|_\infty \log(\|A\|_1\|\hat{u}\|_1+(\gamma+\sigma^2)|\Omega|)\to 0,$$

since $\|z_{n_j}-(z+\gamma)\|_\infty\to 0$ and $\log(\|A\|_1\|\hat{u}\|_1+(\gamma+\sigma^2)|\Omega|)$ is constant, and

$$\|A-A_{n_j}\|_1\|u_{\alpha_{n_j},n_j}\|_1\to 0$$

since $\|A-A_{n_j}\|_1\to 0$ and $\|u_{\alpha_{n_j},n_j}\|_1$ is bounded.

Since A is a bounded linear operator on $L^2(\Omega)$ with Ω a set of finite measure, $F(u)=\int_\Omega Au\,dx$ is a bounded linear functional on $L^2(\Omega)$. The weak convergence of $\{u_{\alpha_{n_j},n_j}\}$ then implies $\int_\Omega Au_{\alpha_{n_j},n_j}\,dx\to\int_\Omega A\hat{u}\,dx$, which yields $\int_\Omega A\left(\hat{u}-u_{\alpha_{n_j},n_j}\right)\,dx\to 0$.

Since A is compact, $u_{\alpha_{n_j},n_j}$ converges weakly to $\hat{u}$, hence $\|Au_{\alpha_{n_j},n_j}-A\hat{u}\|_1\to 0$ (cf. [4, Prop. 3.3]). Thus, since $\left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1$ is bounded, and

$$\begin{aligned}\left\|\frac{A_{n_j}u_{\alpha_{n_j},n_j}+\gamma+\sigma^2}{A\hat{u}+\gamma+\sigma^2}-1\right\|_1 &\le \left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1\|A_{n_j}u_{\alpha_{n_j},n_j}-A\hat{u}\|_1,\\ &\le \left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1\times\\ &\quad\left(\left\|A_{n_j}-A\right\|_1\|u_{\alpha_{n_j},n_j}\|_1+\|Au_{\alpha_{n_j},n_j}-A\hat{u}\|_1\right),\end{aligned}$$

it results that $\|z_{n_j}+\sigma^2\|_\infty\log\|(A_{n_j}u_{\alpha_{n_j},n_j}+\gamma+\sigma^2)/(A\hat{u}+\gamma+\sigma^2)\|_1\to 0$. Therefore

$$T_0(\hat{u})=\lim_{n_j\to\infty}T_{0,n_j}(u_{\alpha_{n_j},n_j}).$$

Invoking (3.14), (3.12), and (3.11), respectively, yields

$$\lim_{n_j\to\infty} T_{0,n_j}(u_{\alpha_{n_j},n_j}) = \lim_{n_j\to\infty} T_{0,n_j}(u_{\alpha_{n_j},n_j}) = \lim_{n_j\to\infty} T_{0,n_j}(u_{\text{exact}}) = T_0(u_{\text{exact}}).$$

Thus $T_0(\hat{u}) = T_0(u_{\text{exact}})$. Since u_{exact} is the unique minimizer of T_0, it follows $\hat{u} = u_{\text{exact}}$. Therefore $\{u_{\alpha_{n_j},n_j}\}$ converges weakly to u_{exact} in $L^2(\Omega)$.

With weak convergence in hand, the next task is to prove strong convergence, which is true provided $||u_{\alpha_n,n}||_2 \to ||u_{\text{exact}}||_2$. In order to prove this is the case, noting that by the weak lower semi-continuity of the norm, $||u_{\text{exact}}||_2^2 \leq \liminf ||u_{\alpha_n,n}||_2^2$, and since $T_{0,n}(u_{0,n}) \leq T_{0,n}(u_{\alpha_n,n})$, then

$$\begin{aligned}
T_{\alpha,n}(u_{\alpha_n,n}) &\leq T_{\alpha_n,n}(u_{\text{exact}}) \\
\Rightarrow \frac{\alpha_n, n}{2}||u_{\alpha_n,n}||_2^2 + T_{0,n}(u_{\alpha_n,n}) &\leq \frac{\alpha_n, n}{2}||u_{\text{exact}}||_2^2 + T_{0,n}(u_{\text{exact}}) \\
\Rightarrow \frac{\alpha_n, n}{2}||u_{\alpha_n,n}||_2^2 + T_{0,n}(u_{0,n}) &\leq \frac{\alpha_n, n}{2}||u_{\text{exact}}||_2^2 + T_{0,n}(u_{\text{exact}}) \\
\Rightarrow ||u_{\alpha_n,n}||_2^2 &\leq ||u_{\text{exact}}||_2^2 - 2\left(\frac{T_{0,n}(u_{0,n}) - T_{0,n}(u_{\text{exact}})}{(\alpha_n, n)}\right),
\end{aligned}$$

therefore with (3.12),

$$\begin{aligned}
\liminf ||u_{\alpha_n,n}||_2^2 &\leq ||u_{\text{exact}}||_2^2 - 2\liminf\left(\frac{T_{0,n}(u_{0,n}) - T_{0,n}(u_{\text{exact}})}{(\alpha_n, n)}\right) \\
&\leq ||u_{\text{exact}}||_2^2.
\end{aligned}$$

Hence it is clear that $||u_{\alpha_n,n}||_2^2 \to ||u_{\text{exact}}||_2^2$ consequently $u_{\alpha_n,n}$ converges to u_{exact} strongly in $L^2(\Omega)$. □

A summary of the results is the next Theorem.

Theorem 3.1.9. *$R_\alpha(A, z+\gamma)$, defined in (3.3) defines a regularization scheme.*

Proof. By Corollaries 3.1.5 and 3.1.7 R_α is well-defined and continuous and therefore satisfies Conditions 1 and 2 of Definition 2.2.1. Theorem 3.1.8 then gives convergence. □

3.2 Differential Regularization Theory

The work in this section appears in [7], though it has been modified the in presentation significantly. In this section, the focus is to study the use of first-order differential operators for regularization. Let define

$$\langle u, v\rangle_{H^1(\Omega)} \stackrel{\text{def}}{=} \langle u, v\rangle_2 + \langle \nabla u, \nabla v\rangle_2, \tag{3.15}$$

where "∇" denotes the gradient. The set of all functions $u \in C^1(\Omega)$ such that

$$\|u\|_{H^1(\Omega)} = \sqrt{\langle u, u\rangle_{H^1(\Omega)}}$$

is finite is a normed linear space whose closure in $L^2(\Omega)$ is the Sobolev space $H^1(\Omega)$ [34]. Noting, moreover, that with the inner-product defined in (3.37), $H^1(\Omega)$ is a Hilbert space. Now, let introduce the $d \times d$ matrix valued function

$$[G(x)]_{ij} = g_{ij}(x), \quad 1 \le i, j \le d,$$

where $G(x)$ is non-singular for all $x \in \bar{\Omega}$. Moreover, let assume that the g_{ij}'s are continuously differentiable for all i and j. Then a regularization functional can be defined, as in [24], by

$$J(u) = \frac{1}{2}\|G(x)\nabla u\|_2^2. \tag{3.16}$$

$\mathcal{C}$ is defined by

$$\mathcal{C} = \{u \ \in \ H^1(\Omega) \mid u \ge 0\} \tag{3.17}$$

and

$$R_\alpha(A, z+\gamma) = \arg\min_{u\in\mathcal{C}} \left\{ T_0(Au; z+\gamma) + \frac{\alpha}{2}\|G(x)u\|^2 \right\}. \tag{3.18}$$

The task in this subsection is to show that (3.18) defines a regularization scheme as defined in Definition 2.2.1.

3.2.1 $R_\alpha(A, z+\gamma)$ is Well-Defined

With the $\|\cdot\|_{H^1(\Omega)}$ norm a function T is *coercive* if

$$T(u) \to +\infty \quad \text{whenever} \quad \|u\|_{H^1(\Omega)} \to +\infty. \tag{3.19}$$

In order to prove the existence and uniqueness of solutions of (3.18), let introduce the following theorem, which is similar to [33, Theorem 2.30].

Theorem 3.2.1. *If $T : H^1(\Omega) \to \mathbb{R}$ is convex and coercive, then it has a minimizer on $\mathcal{C}$. If T is strictly convex it has a unique minimizer over $\mathcal{C}$.*

Proof. Let $\{u_n\} \subset \mathcal{C}$ be such that $T(u_n) \to T_* \stackrel{\text{def}}{=} \inf_{u\in\mathcal{C}} T(u)$. Then, by (3.19), the sequence $\{u_n\}$ is bounded in $H^1(\Omega)$. By the Rellich-Kondrachov Compactness Theorem [16, Theorem 1, Section 5.7], $\{u_n\}$ has a subsequence $\{u_{n_j}\}$ that converges to some $u_* \in \mathcal{C}$. Now, since T is convex, it is weakly lower semi-continuous [35], and hence,

$$T(u_*) \leq \liminf T(u_{n_j}) = \lim T(u_n) = T_*.$$

Thus u_* minimizes T on $\mathcal{C}$. Uniqueness follows immediately if T strictly convex. □

Corollary 3.2.2. *$R_\alpha(A; z+\gamma)$ is well-defined.*

Proof. First, recall that from the above arguments, T_0 is strictly convex since A is invertible. Next, note that

$$\|G(x)\nabla u\|^2 = \sum_{\ell=1}^{d} \int_\Omega \left(\sum_{j=1}^{d} g_{lj}(x)\partial_j u \right)^2 dx \tag{3.20}$$

$$= \sum_{\ell=1}^{d} \int_\Omega \bar{g}_{jk}(x)(\partial_j u)(\partial_k u)\, dx \tag{3.21}$$

where $\bar{g}_{jk}(x) \stackrel{\text{def}}{=} \sum_{\ell=1}^{d} g_{lj}(x) g_{lk}(x)$. It is readily seen that if $[\bar{G}(x)]_{jk} \stackrel{\text{def}}{=} \bar{g}_{jk}(x)$, then $\bar{G}(x) = G^T(x)G(x)$, and hence $\bar{G}(x)$ is symmetric and positive definite for all x. Since $\bar{\Omega}$ is a closed and bounded set, the eigenvalues of $\bar{G}(x)$ will be uniformly bounded away from 0 on Ω. Thus there exists a constant $\theta > 0$ such that

$$\sum_{j,k=1}^{n} \bar{g}_{jk}(x)\xi_j\xi_k \geq \theta|\boldsymbol{\xi}|^2 \tag{3.22}$$

for all $x \in \Omega$ and $\boldsymbol{\xi} = (\xi_1, \cdots, \xi_n) \in \mathbb{R}^n$. Equation (3.22) implies that for $u \in H^1(\Omega)$,

$$J(u) \geq \theta \int_{\Omega} \nabla u \cdot \nabla u \, dx. \tag{3.23}$$

Moreover, the Poincaré inequality [16, Theorem 3, Chapter 5], implies that there exists a constant C such that

$$\|u\|_2^2 \leq C\|\nabla u\|_2^2.$$

Therefore,

$$J(u) + C\theta\|\nabla u\|_2^2 \geq \theta\|u\|_{H^1(\Omega)}^2.$$

But by (3.23), $\|\nabla u\|_2^2 \leq \theta^{-1}J(u)$, and so this yields to, finally,

$$J(u) \geq \frac{\theta}{1+C}\|u\|_{H^1(\Omega)}^2. \tag{3.24}$$

Thus J is coercive, and is convex by results shown in [13].

The strict convexity of T_α over $\mathcal{C}$ follows immediately.

To show that T_α is coercive, noting that by Jensen's inequality and the properties of the function $x - c\log x$ for $c > 0$,

$$\begin{aligned} T_0(u) &\geq \|Au + \gamma + \sigma^2\|_1 - \|z + \gamma + \sigma^2\|_\infty \log \|Au + \gamma + \sigma^2\|_1, \\ &\geq \|z + \gamma + \sigma^2\|_\infty - \|z + \gamma + \sigma^2\|_\infty \log \|z + \gamma + \sigma^2\|_\infty. \end{aligned} \tag{3.25}$$

Since $z \geq 0$, T_0 is bounded below. Thus the coercivity of J implies the coercivity of T_α.

The desired result then follows from Theorem 3.2.1 □

Remark: The choice of boundary conditions will also determine the space in which solutions can be expected to be found. For example, if the boundary condition $u = 0$ on $\partial\Omega$ is used, then J will be strictly convex on the set $H_0^1(\Omega) \stackrel{\text{def}}{=} \{u \in H^1(\Omega) \mid u = 0 \text{ on } \partial\Omega\}$ [16], and hence, by the arguments above T_α will have a unique minimizer on $\mathcal{C}_0 = \{u \in H_0^1(\Omega) \mid u \geq 0\}$.

3.2.2 $R_\alpha(A, z+\gamma)$ is Continuous

Recalling that A_n and z_n in (1.11) satisfy the same assumptions as A and z, the above arguments give that $u_{\alpha,n}$ (defined in (1.12)) exists, and is unique if A_n is invertible.

Theorem 3.2.3. *Let u_α be the unique minimizer of T_α over $\mathcal{C}$, and for each $n \in \mathbb{N}$ let $u_{\alpha,n}$ a minimizer of $T_{\alpha,n}$ over $\mathcal{C}$. Suppose, furthermore, that*

1. *for any sequence $\{u_n\} \subset H^1(\Omega)$,*

$$\lim_{n\to\infty} T_{\alpha,n}(u_n) = +\infty \quad \textit{whenever} \quad \lim_{n\to\infty} \|u_n\|_{H^1(\Omega)} = +\infty; \tag{3.26}$$

2. *given $B > 0$ and $\epsilon > 0$, there exists N such that*

$$|T_{\alpha,n}(u) - T_\alpha(u)| < \epsilon \quad \textit{whenever} \quad n \geq N,\ \|u\|_{H^1(\Omega)} \leq B. \tag{3.27}$$

Then

$$u_{\alpha,n} \to u_\alpha \textit{ in } L^2(\Omega). \tag{3.28}$$

Proof. Note that $T_{\alpha,n}(u_{\alpha,n}) \leq T_{\alpha,n}(u_\alpha)$. From this and (3.27), it is obvious that

$$\liminf T_{\alpha,n}(u_{\alpha,n}) \leq \limsup T_{\alpha,n}(u_{\alpha,n}) \leq T_\alpha(u_\alpha) < \infty. \tag{3.29}$$

Thus by (3.26), the $u_{\alpha,n}$'s are bounded in $H^1(\Omega)$. The Rellich-Kondrachov Compactness Theorem [16, Theorem 1, Section 5.7] then tells that there exists a subsequence $\{u_{n_j}\}$ that converges strongly to some $\hat{u} \in L^2(\Omega)$. Furthermore, the weak lower semicontinuity of T_α, (3.27), and (3.29) imply

$$\begin{aligned} T_\alpha(\hat{u}) &\leq \liminf T_\alpha(u_{n_j}), \\ &= \liminf (T_\alpha(u_{n_j}) - T_{\alpha,n_j}(u_{n_j})) + \liminf T_{\alpha,n_j}(u_{n_j}), \\ &\leq T_\alpha(u_\alpha). \end{aligned}$$

By uniqueness of minimizers, $\hat{u} = u_\alpha$. Thus every convergent subsequence of $\{u_{\alpha,n}\}$ converges strongly to u_α, and hence, it results (3.28). □

The following corollary of Theorem 3.2.3 is the stability result for (3.18) that is desired.

Corollary 3.2.4. *$R_\alpha(A, z+\gamma)$ is continuous.*

Proof. It suffices to show that conditions *(i)* and *(ii)* from Theorem 3.2.3 hold. For condition *(i)*, note that the analogue of inequality (3.25) for $T_{0,n}$ is given by

$$T_{0,n}(u_{\alpha,n}) \geq \|z_n + \sigma^2\|_\infty - \|z_n + \sigma^2\|_\infty \log \|z_n + \sigma^2\|_\infty,$$

which has a lower bound for all n since $\|z_n - (z+\gamma)\|_\infty \to 0$ and $z \in L^\infty(\Omega)$ is nonnegative. Thus by (3.24) $T_{\alpha,n}(u_n) = T_{0,n}(u_n) + \alpha J(u_n) \to +\infty$ whenever $\|u_n\|_{H^1(\Omega)} \to \infty$, and hence (3.26) is satisfied.

For condition *(ii)*, note that using Jensen's inequality and the properties of the logarithm

$$\begin{aligned} |T_{\alpha,n}(u) - T_\alpha(u)| &= \Big| \int_\Omega \big((A_n - A)u - (z_n + \sigma^2)\log(A_n u + \gamma + \sigma^2)\big)\, dx \\ &\qquad + \int_\Omega \big((z+\gamma+\sigma^2)\log(Au + \gamma + \sigma^2)\big)\, dx \Big|, \\ &\leq \|A_n - A\|_1 \|u\|_1 \\ &\quad + \|z_n - (z+\gamma)\|_\infty \log(\|A_n\|_1 \|u\|_1 + (\gamma + \sigma^2)|\Omega|) \qquad (3.30) \\ &\quad + \|z + \gamma + \sigma^2\|_\infty \log \big\|(Au + \gamma + \sigma^2)/(A_n u + \gamma + \sigma^2)\big\|_1 . \end{aligned}$$

By assumption, $\|A_n - A\|_1, \|z_n - (z+\gamma)\|_\infty \to 0$. Furthermore, by the Banach-Steinhaus

Theorem, $\|A_n\|_1$ is uniformly bounded. Since it is assumed that $\|u\|_{H^1(\Omega)}$ is bounded, and $H^1(\Omega) \subset L^2(\Omega)$, $\|u\|_2$ will be bounded as well. Moreover, since Ω is a bounded set, this implies that $\|u\|_1$ is bounded. Thus the first two terms on the right-hand side in (3.30) tend to zero as $n \to \infty$. For the third term note that

$$\left\| \frac{Au + \gamma + \sigma^2}{A_n u + \gamma + \sigma^2} - 1 \right\|_1 \leq \left\| \frac{1}{A_n u + \gamma + \sigma^2} \right\|_1 \|A_n - A\|_1 \, \|u\|_1,$$

which converges to zero since $\|1/(A_n u + \gamma + \sigma^2)\|_1$ is bounded and $\|A_n - A\|_1 \to 0$. Thus $\log(\|(Au + \gamma + \sigma^2)/(Au_n + \gamma + \sigma^2)\|_1) \to \log(1) = 0$, and hence

$$|T_{\alpha,n}(u) - T_\alpha(u)| \to 0. \tag{3.31}$$

The desired result now follows from Theorem 3.2.3. □

3.2.3 $R_\alpha(A, z + \gamma)$ is Convergent

The task now, is to prove that $R_\alpha(A, z + \gamma)$ is convergent, that is, Condition 2 of Definition 2.2.1 holds. For this, let assume that A is invertible so that u_{exact} is the unique solution of $Au = z$.

Theorem 3.2.5. *$R_\alpha(A, z + \gamma)$ is convergent.*

Proof. Suppose $\alpha_n \to 0$ at a rate such that

$$(T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n}))/\,\alpha_n \tag{3.32}$$

is bounded. Then since $u_{\alpha_n,n}$ minimizes $T_{\alpha_n,n}$, this yields

$$T_{\alpha_n,n}(u_{\alpha_n,n}) \leq T_{\alpha_n,n}(u_{\text{exact}}). \tag{3.33}$$

Since $\{z_n\}$ and $\{A_n\}$ are uniformly bounded and $A_n \to A$ in the $L^1(\Omega)$ operator norm, $\{T_{\alpha_n,n}(u_{\text{exact}})\}$ is a bounded sequence. Hence $\{T_{\alpha_n,n}(u_{\alpha_n,n})\}$ is bounded by (3.33).

Subtracting $T_{0,n}(u_{0,n})$ from both sides of (3.33) and dividing by α_n yields

$$\begin{aligned}(T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n}))/\alpha_n + \alpha J(u_{\alpha_n,n}) \quad &\leq \quad (T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n}))/\,\alpha_n \\ &\qquad + \alpha J(u_{\text{exact}}). \end{aligned} \tag{3.34}$$

By (3.32), the right-hand side is bounded, implying that the left hand side is bounded. Since $T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n})$ is nonnegative, this implies that $\{J(u_{\alpha_n,n})\}$ is bounded. Equation (3.24) then tells that $\{u_{\alpha_n,n}\}$ is bounded in $H^1(\Omega)$.

Next is to show that $u_{\alpha_n,n} \to u_{\text{exact}}$ in $L^2(\Omega)$ by showing that every subsequence of $\{u_{\alpha_n,n}\}$ contains a subsequence that converges to u_{exact}. Since $\{u_{\alpha_n,n}\}$ is bounded in $H^1(\Omega)$, it is relatively compact in $L^2(\Omega)$ by the Rellich-Kondrachov Compactness Theorem [16, Theorem 1, Section 5.7]. Thus each of its subsequences in turn has a subsequence that converges strongly in $L^2(\Omega)$. Let $\{u_{\alpha_{n_j},n_j}\}$ be such a sequence and $\hat{u}$ its limit. Then

$$\begin{aligned} T_0(\hat{u}) \quad = \quad & \int_\Omega (A(\hat{u} - u_{\alpha_{n_j},n_j}) + (A - A_{n_j})u_{\alpha_{n_j},n_j}\, dx \\ & + \int_\Omega (z_{n_j} - (z+\gamma)) \log(A\hat{u} + \gamma + \sigma^2))\, dx \\ & - \int_\Omega (z_{n_j} + \sigma^2) \log((A_{n_j} u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2))\, dx \\ & + T_{0,n_j}(u_{\alpha_{n_j},n_j}), \end{aligned}$$

which, as in previous arguments, yields

$$\begin{aligned} |T_{0,n_j}(u_{\alpha_{n_j},n_j}) \quad - \quad & T_0(\hat{u})| \\ \leq \quad & \int_\Omega A\left(\hat{u} - u_{\alpha_{n_j},n_j}\right)\, dx \\ & + \|z_{n_j} - (z+\gamma)\|_\infty \log(\|A\|_1 \|\hat{u}\|_1 + \gamma|\Omega|) \\ & + \|z_{n_j} + \sigma^2\|_\infty \log \|(A_{n_j} u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2)\|_1 \\ & + \|A - A_{n_j}\|_1 \|u_{\alpha_{n_j},n_j}\|_1. \end{aligned}$$

Then

$$\|z_{n_j} - (z+\gamma)\|_\infty \log(\|A\|_1\|\hat{u}\|_1 + (\gamma+\sigma^2)|\Omega|) \to 0,$$

since $\|z_{n_j} - (z+\gamma)\|_\infty \to 0$ and $\log(\|A\|_1\|\hat{u}\|_1 + (\gamma+\sigma^2)|\Omega|)$ is constant, and

$$\|A - A_{n_j}\|_1 \|u_{\alpha_{n_j},n_j}\|_1 \to 0$$

since $\|A - A_{n_j}\|_1 \to 0$, and $\|u_{\alpha_{n_j},n_j}\|_1$ is bounded since $\|u_{\alpha_{n_j},n_j}\|_{H^1(\Omega)}$ is bounded and $H^1(\Omega)$ is compactly embedded in $L^2(\Omega) \subset L^1(\Omega)$.

Knowing that A is a bounded linear operator and Ω is a set of finite measure, gives $F(u) = \int_\Omega Au\, dx$ is a bounded linear functional on $L^p(\Omega)$. The convergence of $\{u_{\alpha_{n_j},n_j}\}$ then implies $\int_\Omega Au_{\alpha_{n_j},n_j}\, dx \to \int_\Omega A\hat{u}\, dx$, which yields $\int_\Omega A\left(\hat{u} - u_{\alpha_{n_j},n_j}\right)\, dx \to 0$.

Since A is compact, it is completely continuous, i.e. $u_{\alpha_{n_j},n_j} \to \hat{u}$ implies that $\|Au_{\alpha_{n_j},n_j} - A\hat{u}\|_1 \to 0$ (cf. [4, Prop. 3.3]). Thus, since $\left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1$ is bounded and

$$\begin{aligned}\left\|\frac{A_{n_j}u_{\alpha_{n_j},n_j}+\gamma+\sigma^2}{A\hat{u}+\gamma+\sigma^2} - 1\right\|_1 &\le \left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1 \|A_{n_j}u_{\alpha_{n_j},n_j} - A\hat{u}\|_1, \\ &\le \left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1 \times \\ &\quad \left(\left\|A_{n_j} - A\right\|_1 \|u_{\alpha_{n_j},n_j}\|_1 + \|Au_{\alpha_{n_j},n_j} - A\hat{u}\|_1\right),\end{aligned}$$

hence $\|z_{n_j} + \sigma^2\|_\infty \log\|(A_{n_j}u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2)\|_1 \to 0$. Therefore

$$T_0(\hat{u}) = \lim_{n_j\to\infty} T_{0,n_j}(u_{\alpha_{n_j},n_j}).$$

Invoking (3.34), (3.32), and (3.31), respectively, yields

$$\lim_{n_j\to\infty} T_{0,n_j}(u_{\alpha_{n_j},n_j}) = \lim_{n_j\to\infty} T_{0,n_j}(u_{\text{exact}}) = T_0(u_{\text{exact}}).$$

Thus $T_0(\hat{u}) = T_0(u_{\text{exact}})$. Since u_{exact} is the unique minimizer of T_0, it results $\hat{u} = u_{\text{exact}}$. Therefore $\{u_{\alpha_{n_j},n_j}\}$ converges strongly to u_{exact} in $L^2(\Omega)$. □

The result obtained in this subsection are summarized in the following Theorem.

Theorem 3.2.6. *$R_\alpha(A, z+\gamma)$ defined in (3.18) defines a regularization scheme.*

Proof. By Corollaries 3.2.2 and 3.2.4, $R_\alpha(A, z+\gamma)$ is well-defined and continuous, and therefore satisfies Condition 1 of Definition 2.2.1. Theorem 3.2.5 then gives Condition 2 and hence R_α defines a regularization scheme. □

3.3 Higher Order Differential Regularization Theory

In this section, the focus is the use of m-order differential operators for regularization purpose. Let define

$$\tau_i = \sum_{j=1}^{m+1} q_{ij}(x)\partial_i^{j-1}, \tag{3.35}$$

and

$$\tau \stackrel{\text{def}}{=} \sum_{i=1}^{d} \tau_i e_i, \tag{3.36}$$

$\{e_i\}_{1\le i\le d}$ being the canonical basis of $\mathbb{R}^d$ and $\partial_i^{j-1} = (\partial/\partial x_i)^{j-1}$. The norm in this section is

$$||u||_{H_m^0(\Omega)} = \left[\sum_{|\alpha|\le m} \int_\Omega |\partial^\alpha u|^2\, dx\right]^{\frac{1}{2}} \tag{3.37}$$

$H_m^0(\Omega)$ is the completion of $C_c^\infty(\Omega)$ with respect to $||\cdot||_{H_m^0(\Omega)}$.

In addition Q is a $d\times(m+1)$ real valued matrix of functions

$$[Q(x)]_{ij} = q_{ij}(x), \quad 1\le i\le d, \quad 1\le j\le m+1$$

where the $q_{ij}(x)$'s are nonnegative and $C^\infty(\Omega)$ for all $x\in\bar{\Omega}$. Furthermore by assumption the $q_{i,n+1}(x)$'s are strictly positive. The regularizing functional is given by

$$J(u) = \|\tau u\|_2^2 = \sum_{i=1}^{d} \int_\Omega \left(\sum_{k,l=1}^{m+1} q_{ik}(x)q_{il}(x)\partial_i^{k-1}u\partial_i^{l-1}u\right)\, dx. \tag{3.38}$$

Lastly, $Q(x)$ is such that there is $\theta > 0$,

$$J(u) \geq \theta \sum_{i=1}^{d} \int_{\Omega} (q_{i,m+1}(x))^2 (\partial_i^m u)^2 \, dx. \tag{3.39}$$

The definition of $\mathcal{C}$ is then

$$\mathcal{C} = \{ \, u \, \in \, H_m^0(\Omega) \, | \, u \geq 0 \, \}. \tag{3.40}$$

The regularization scheme is

$$R_\alpha(A, z + \gamma) = \arg\min_{u \in \mathcal{C}} \left\{ T_0(Au; z + \gamma) + \frac{\alpha}{2} \|\tau u\|^2 \right\}. \tag{3.41}$$

The task in this subsection is to show that (3.41) defines a regularization scheme as defined in Definition 2.2.1.

3.3.1 $R(A; z + \gamma)$ is Well-Defined

Accordingly to the $H_m^0(\Omega)$-norm a function T is *coercive* if

$$T(u) \to +\infty \quad \text{whenever} \quad \|u\|_{H_m^0(\Omega)} \to +\infty. \tag{3.42}$$

In order to prove the existence and uniqueness of solutions of (3.41), the use of the following theorem, which is similar to [33, Theorem 2.30] helps.

Theorem 3.3.1. *If $T : H_m^0(\Omega) \to \mathbb{R}$ is convex and coercive, then it has a minimizer on $\mathcal{C}$. If T is strictly convex it has a unique minimizer over $\mathcal{C}$.*

Proof. Let $\{u_n\} \subset \mathcal{C}$ be such that $T(u_n) \to T_* \stackrel{\text{def}}{=} \inf_{u \in \mathcal{C}} T(u)$. Then, by (3.42), the sequence $\{u_n\}$ is bounded in $H_m^0(\Omega)$. By the Rellich's Theorem [5, Theorem 6.14], this implies that $\{u_n\}$ has a subsequence $\{u_{n_j}\}$ that converges to some $u_* \in \mathcal{C}$. Now, since T is convex, it is weakly lower semi-continuous [35], and hence,

$$T(u_*) \leq \liminf T(u_{n_j}) = \lim T(u_n) = T_*.$$

Thus u_* minimizes T on $\mathcal{C}$. Uniqueness follows immediately if T is strictly convex. □

Corollary 3.3.2. *$R_\alpha(A, z+\gamma)$ is well-defined.*

Proof. Since the $q_{i,m+1}$'s are strictly positive , there is $\{c_{i,m+1}\}_{1\leq i\leq d}$ such that

$$q_{i,m+1}(x) \geq c_{i,m+1} > 0 \text{ for all } i \text{ and } x \in \Omega.$$

Hence

$$\int_\Omega \sum_{i=1}^{d} (q_{i,n+1})^2 (\partial_i^n u)^2 \, dx \geq \int_\Omega \sum_{i=1}^{d} (c_{i,n+1})^2 (\partial_i^n u)^2 \, dx, \tag{3.43}$$

by Theorem [5, 7.14], since the right hand side of (3.43) is such that $\sum_{i=1}^{d}(c_{i,n+1})^2\xi_i^{2n} \neq 0$ for all nonzero $\xi \in \mathbb{R}^d$ [5, page 210], it is elliptic with real coefficients, hence coercive on $H_m^0(\Omega)$. Thus, by (3.39) J is coercive, and is convex as a composite functional of linear functions and the norm squared.

The convexity of T_α over $\mathcal{C}$ follows immediately, with strict convexity if A is invertible.

To show that T_α is coercive, noting that by Jensen's inequality and the properties of the function $x - c\log x$ for $c > 0$,

$$\begin{aligned} T_0(u) &\geq \|Au + \gamma + \sigma^2\|_1 - \|z + \gamma + \sigma^2\|_\infty \log \|Au + \gamma + \sigma^2\|_1, \\ &\geq \|z + \gamma + \sigma^2\|_\infty - \|z + \gamma + \sigma^2\|_\infty \log \|z + \gamma + \sigma^2\|_\infty. \end{aligned} \tag{3.44}$$

Since $z \geq 0$, T_0 is bounded below. Thus the coercivity of J implies the coercivity of T_α.

The desired result then follows from Theorem 3.3.1 □

Remark: Even though, if $u \in H_m^0(\Omega)$ then $u = 0$ on the boundary [5, Corollary 6.48], this analysis can generalized to reflexive and periodic boundary conditions since in both cases no boundary condition is assumed on the boundary of the extended function, which is the same as assuming zero boundary condition on the extended function [22, page 31].

3.3.2 $R_\alpha(A, z+\gamma)$ is Continuous

Recalling that A_n and z_n in (1.11) satisfy the same assumptions as A and z, from the above arguments $u_{\alpha,n}$ (defined in (1.12)) exists, and is unique if A_n is invertible.

To prove stability, let assume that T_α has a unique minimizer u_α over $\mathcal{C}$ and then show that $A_n \to A$ and $z_n \to z+\gamma$ implies $u_{\alpha,n} \to u_\alpha$. The following theorem gives conditions that guarantee this result. For completeness, let present the proof.

Theorem 3.3.3. *Let u_α be the unique minimizer of T_α over $\mathcal{C}$, and for each $n \in \mathbb{N}$ let $u_{\alpha,n}$ a minimizer of $T_{\alpha,n}$ over $\mathcal{C}$. Suppose, furthermore, that*

1. *for any sequence $\{u_n\} \subset H_m^0(\Omega)$,*

$$\lim_{n\to\infty} T_{\alpha,n}(u_n) = +\infty \quad \textit{whenever} \quad \lim_{n\to\infty} \|u_n\|_{H_m^0(\Omega)} = +\infty; \tag{3.45}$$

2. *given $B > 0$ and $\epsilon > 0$, there exists N such that*

$$|T_{\alpha,n}(u) - T_\alpha(u)| < \epsilon \quad \textit{whenever} \quad n \geq N,\ \|u\|_{H_m^0(\Omega)} \leq B. \tag{3.46}$$

Then

$$\lim_{n\to\infty} \|u_{\alpha,n} - u_\alpha\|_{H_k^0(\Omega)} = 0 \textit{ for all } k < m. \tag{3.47}$$

Proof. Note that $T_{\alpha,n}(u_{\alpha,n}) \leq T_{\alpha,n}(u_\alpha)$. From this and (3.46), it follows

$$\liminf T_{\alpha,n}(u_{\alpha,n}) \leq \limsup T_{\alpha,n}(u_{\alpha,n}) \leq T_\alpha(u_\alpha) < \infty. \tag{3.48}$$

Thus by (3.45), the $u_{\alpha,n}$'s are bounded in $H_m^0(\Omega)$. Rellich's Theorem [5, Theorem 6.14] then tells us that there exists a subsequence $\{u_{n_j}\}$ that converges to some $\hat{u} \in H_k^0(\Omega)$ for $k < m$.

Furthermore, by the weak lower semicontinuity of T_α, (3.46), and (3.48)

$$\begin{aligned} T_\alpha(\hat{u}) &\leq \liminf T_\alpha(u_{n_j}), \\ &= \liminf (T_\alpha(u_{n_j}) - T_{\alpha,n_j}(u_{n_j})) + \liminf T_{\alpha,n_j}(u_{n_j}), \\ &\leq T_\alpha(u_\alpha). \end{aligned}$$

By uniqueness of minimizers, $\hat{u} = u_\alpha$. Thus every convergent subsequence of $\{u_{\alpha,n}\}$ converges to u_α, which implies, (3.47). □

The following corollary of Theorem 3.3.3 is the stability result for (3.41) that is sought.

Corollary 3.3.4. *$R_\alpha(A, z+\gamma)$ is continuous in $H_k^0(\Omega)$ for $k < m$.*

Proof. It suffices to show that conditions *(i)* and *(ii)* from Theorem 3.3.3 hold. For condition *(i)*, note that the analogue of inequality (3.44) for $T_{0,n}$ is given by

$$T_{0,n}(u_{\alpha,n}) \geq \|z_n + \sigma^2\|_\infty - \|z_n + \sigma^2\|_\infty \log \|z_n + \sigma^2\|_\infty,$$

which has a lower bound for all n since $\|z_n - (z+\gamma)\|_\infty \to 0$ and $z \in L^\infty(\Omega)$ is nonnegative. Thus by the coercivity of J, $T_{\alpha,n}(u_n) = T_{0,n}(u_n) + \alpha J(u_n) \to +\infty$ whenever $\|u_n\|_{H_m^0(\Omega)} \to \infty$, and hence (3.45) is satisfied.

For condition *(ii)*, note that using Jensen's inequality and the properties of the logarithm

$$\begin{aligned} |T_{\alpha,n}(u) - T_\alpha(u)| &= \left| \int_\Omega \left((A_n - A)u - (z_n + \sigma^2)\log(A_n u + \gamma + \sigma^2)\right) dx \right. \\ &\qquad \left. + \int_\Omega \left((z + \gamma + \sigma^2)\log(Au + \gamma + \sigma^2)\right) dx \right|, \\ &\leq \|A_n - A\|_1 \|u\|_1 \\ &\quad + \|z_n - (z+\gamma)\|_\infty \log(\|A_n\|_1 \|u\|_1 + (\gamma + \sigma^2)|\Omega|) \\ &\quad + \|z + \gamma + \sigma^2\|_\infty \log \left\|(Au + \gamma + \sigma^2)/(A_n u + \gamma + \sigma^2)\right\|_1 . \end{aligned} \tag{3.49}$$

By assumption, $\|A_n - A\|_1, \|z_n - (z+\gamma)\|_\infty \to 0$. Furthermore, by the Banach-Steinhaus

Theorem, $\|A_n\|_1$ is uniformly bounded. Since it is assumed that $\|u\|_{H_m^0(\Omega)}$ is bounded, by Rellich's Theorem [5, Theorem 6.14] $\|u\|_{H_k^0(\Omega)}$ will be bounded as well for $k < m$. Moreover, since Ω is a bounded set, this implies that $\|u\|_1$ is bounded. Thus the first two terms on the right-hand side in (3.49) tend to zero as $n \to \infty$. For the third term note that

$$\left\| \frac{Au + \gamma + \sigma^2}{A_n u + \gamma + \sigma^2} - 1 \right\|_1 \leq \left\| \frac{1}{A_n u + \gamma + \sigma^2} \right\|_1 \|A_n - A\|_1 \|u\|_1,$$

which converges to zero since $\|1/(A_n u + \gamma + \sigma^2)\|_1$ is bounded and $\|A_n - A\|_1 \to 0$. Thus $\log(\|(Au + \gamma + \sigma^2)/(Au_n + \gamma + \sigma^2)\|_1) \to \log(1) = 0$, and hence

$$|T_{\alpha,n}(u) - T_\alpha(u)| \to 0. \tag{3.50}$$

The desired result now follows from Theorem 3.3.3. □

3.3.3 $R(A, z + \gamma)$ is Convergent

It remains to show that a sequence of positive regularization parameters $\{\alpha_n\}$ can be chosen so that $u_{\alpha_n,n} \to u_{\text{exact}}$ as $\alpha_n \to 0$.

Theorem 3.3.5. $R_\alpha(A, z + \gamma)$ *is convergent.*

Proof. Suppose $\alpha_n \to 0$ at a rate such that

$$(T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n}))/\,\alpha_n \tag{3.51}$$

is bounded. Since $u_{\alpha_n,n}$ minimizes $T_{\alpha_n,n}$,

$$T_{\alpha_n,n}(u_{\alpha_n,n}) \leq T_{\alpha_n,n}(u_{\text{exact}}). \tag{3.52}$$

Since $\{z_n\}$ and $\{A_n\}$ are uniformly bounded and $A_n \to A$ in the $L^1(\Omega)$ operator norm, $\{T_{\alpha_n,n}(u_{\text{exact}})\}$ is a bounded sequence. Hence $\{T_{\alpha_n,n}(u_{\alpha_n,n})\}$ is bounded by (3.80).

Subtracting $T_{0,n}(u_{0,n})$ from both sides of (3.52) and dividing by α_n yields

$$(T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n}))/\alpha_n + \alpha J(u_{\alpha_n,n}) \leq (T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n}))/\,\alpha_n + \alpha J(u_{\text{exact}}). \tag{3.53}$$

By (3.51), the right-hand side is bounded, implying the left hand side is bounded. Since $T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n})$ is nonnegative, this implies that $\{J(u_{\alpha_n,n})\}$ is bounded. The coercivity of J then tells us that $\{u_{\alpha_n,n}\}$ is bounded in $H_m^0(\Omega)$.

Coming next is proving that $u_{\alpha_n,n} \to u_{\text{exact}}$ in $H_k^0(\Omega)$ by showing that every subsequence of $\{u_{\alpha_n,n}\}$ contains a subsequence that converges to u_{exact}. Since $\{u_{\alpha_n,n}\}$ is bounded in $H_m^0(\Omega)$, each of its subsequences in turn has a subsequence that converges strongly in $H_k^0(\Omega)$ for $k < m$ [5, Theorem 6.14]. Let $\{u_{\alpha_{n_j},n_j}\}$ be such a sequence and $\hat{u}$ its limit. Then

$$\begin{aligned} T_0(\hat{u}) \;=\; & \int_\Omega (A(\hat{u} - u_{\alpha_{n_j},n_j}) + (A - A_{n_j})u_{\alpha_{n_j},n_j}\, dx\, dy \\ & + \int_\Omega (z_{n_j} - (z+\gamma)) \log(A\hat{u} + \gamma + \sigma^2))\, dx\, dy \\ & - \int_\Omega (z_{n_j} + \sigma^2) \log((A_{n_j} u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2))\, dx\, dy \\ & + T_{0,n_j}(u_{\alpha_{n_j},n_j}), \end{aligned}$$

which, as in previous arguments, yields

$$\begin{aligned} |T_{0,n_j}(u_{\alpha_{n_j},n_j}) \;-\; & T_0(\hat{u})| \\ \leq\; & \int_\Omega A\left(\hat{u} - u_{\alpha_{n_j},n_j}\right)\, dx\, dy \\ & + \|z_{n_j} - (z+\gamma)\|_\infty \log(\|A\|_1 \|\hat{u}\|_1 + \gamma|\Omega|) \\ & + \|z_{n_j} + \sigma^2\|_\infty \log \|(A_{n_j} u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2)\|_1 \\ & + \|A - A_{n_j}\|_1 \|u_{\alpha_{n_j},n_j}\|_1. \end{aligned}$$

Then

$$\|z_{n_j} - (z+\gamma)\|_\infty \log(\|A\|_1 \|\hat{u}\|_1 + (\gamma + \sigma^2)|\Omega|) \to 0,$$

since $\|z_{n_j} - (z+\gamma)\|_\infty \to 0$ and $\log(\|A\|_1\|\hat{u}\|_1 + (\gamma+\sigma^2)|\Omega|)$ is constant, and

$$\|A - A_{n_j}\|_1 \|u_{\alpha_{n_j},n_j}\|_1 \to 0$$

since $\|A - A_{n_j}\|_1 \to 0$, and $\|u_{\alpha_{n_j},n_j}\|_1$ is bounded since $\|u_{\alpha_{n_j},n_j}\|_{H_m^0(\Omega)}$ is bounded and $H_m^0(\Omega)$ is compactly embedded in $L^2(\Omega) \subset L^1(\Omega)$.

Since A is a bounded linear operator and Ω is a set of finite measure, therefore $F(u) = \int_\Omega Au\, dx$ is a bounded linear functional on $L^p(\Omega)$. The convergence of $\{u_{\alpha_{n_j},n_j}\}$ then implies $\int_\Omega Au_{\alpha_{n_j},n_j}\, dx \to \int_\Omega A\hat{u}\, dx$, which yields $\int_\Omega A\left(\hat{u} - u_{\alpha_{n_j},n_j}\right)\, dx \to 0$.

Since A is compact, it is completely continuous, i.e. $u_{\alpha_{n_j},n_j} \to \hat{u}$ implies that $\|Au_{\alpha_{n_j},n_j} - A\hat{u}\|_1 \to 0$ (cf. [4, Prop. 3.3]). Thus, since $\left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1$ is bounded and

$$\begin{aligned}
\left\|\frac{A_{n_j}u_{\alpha_{n_j},n_j} + \gamma + \sigma^2}{A\hat{u}+\gamma+\sigma^2} - 1\right\|_1 \quad &\leq \quad \left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1 \|A_{n_j}u_{\alpha_{n_j},n_j} - A\hat{u}\|_1, \\
&\leq \quad \left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1 \times \\
&\quad \left(\left\|A_{n_j} - A\right\|_1 \|u_{\alpha_{n_j},n_j}\|_1 + \|Au_{\alpha_{n_j},n_j} - A\hat{u}\|_1\right),
\end{aligned}$$

on the other hand, $\|z_{n_j} + \sigma^2\|_\infty \log \|(A_{n_j}u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2)\|_1 \to 0$. Therefore

$$T_0(\hat{u}) = \lim_{n_j\to\infty} T_{0,n_j}(u_{\alpha_{n_j},n_j}).$$

Invoking (3.53), (3.51), and (3.50), respectively, yields

$$\lim_{n_j\to\infty} T_{0,n_j}(u_{\alpha_{n_j},n_j}) = \lim_{n_j\to\infty} T_{0,n_j}(u_{\text{exact}}) = T_0(u_{\text{exact}}).$$

Thus $T_0(\hat{u}) = T_0(u_{\text{exact}})$. Since u_{exact} is the unique minimizer of T_0, $\hat{u} = u_{\text{exact}}$. Therefore $\{u_{\alpha_{n_j},n_j}\}$ converges strongly to u_{exact} in $H_k^0(\Omega)$ for $k < m$. □

In this second part let suppose that $u_{\text{exact}} \notin \mathcal{C}$.

Definition 3.3.6. *Let $u_{\text{exact}} \in L^2(\Omega)$. u_{bapprox} is called a best approximation to u_{exact} if*

$$u_{\text{bapprox}} = \arg\min_{u \in \mathcal{C}} \|u - u_{\text{exact}}\|. \tag{3.54}$$

The existence of u_{bapprox} is guaranteed since $\mathcal{C}$ is closed [11, Theorem V.2].

Now the goal is to show that $\{\alpha_n\}$ can be chosen so that $u_{\alpha_n,n} \to u_{\text{bapprox}}$ as $\alpha_n \to 0$

Corollary 3.3.7. *There is a sequence $\{\alpha_n\}$ such that $u_{\alpha_n,n} \in \mathcal{C}$ and*

$$u_{\alpha_n,n} \to u_{\text{bapprox}}. \tag{3.55}$$

Proof. Since $\mathcal{C}$ is a closed subspace of $L^2(\Omega)$, it follows

$$L^2(\Omega) = \mathcal{C} \oplus \mathcal{C}^\perp$$

hence

$$\begin{aligned} u_{\text{exact}} &= u_{\text{bapprox}} + (I - P_{\mathcal{C}})u_{\text{exact}} \\ \Rightarrow \quad z &= Au_{\text{bapprox}} + A(I - P_{\mathcal{C}})u_{\text{exact}} \end{aligned}$$

Let have

$$Au_{\text{bapprox}} = z_{\text{bapprox}}$$

$$\begin{aligned} \nabla T_0(Au; z + \gamma) &= A^* \left(\frac{Au - z_{\text{bapprox}}}{Au + \gamma + \sigma^2} \right) \\ &+ A^* \left(\frac{A(P_{\mathcal{C}} - I)u_{\text{exact}}}{Au + \gamma + \sigma^2} \right). \end{aligned} \tag{3.56}$$

From the above decomposition of $\nabla T_0(Au; z+\gamma)$ it is readily seen that u_{bapprox} is the minimizer of $T_0(Au; z_{\text{bapprox}} + \gamma)$. By theorem 3.3.5, since $u_{\text{bapprox}} \in \mathcal{C}$ there is $\{\alpha_n\}$ such that $u_{\alpha_n,n} \to$

u_{bapprox}. Moreover, when applied to $T_0(Au; z+\gamma)$ this sequence will lead to u_{bapprox} since the first term of the right hand side of (3.56) will go to zero. □

Putting together the well-posedness of $R_\alpha(A, z+\gamma)$, its continuity and convergence yield to the final Theorem of this section.

Theorem 3.3.8. *$R_\alpha(A, z+\gamma)$ defined in (3.41) is a regularization scheme.*

3.4 Total Variation Regularization

The work in this section appears in [9]. In this section, the subject is to study total variation regularization [2, 12, 26]. In theory, this is accomplished by taking

$$J(u) = J_\beta(u) \stackrel{\text{def}}{=} \sup_{v\in\mathcal{V}} \int_\Omega \left(-u\nabla\cdot v + \sqrt{\beta(1-|v(x)|^2)}\right) dx, \tag{3.57}$$

where $\beta \geq 0$ and

$$\mathcal{V} = \{v \in C_0^1(\Omega; \mathbb{R}^d) \ : \ |v(x)| \leq 1 \quad x \in \Omega\}.$$

Noting that if u is continuously differentiable on Ω, (3.57) takes the recognizable form [2, Theorem 2.1]

$$J_\beta(u) \stackrel{\text{def}}{=} \int_\Omega \sqrt{|\nabla u|^2 + \beta}\, dx. \tag{3.58}$$

$J_0(u)$ is known as the *total variation* of u.

When (3.57) is used, minimizers of T_α will lie (as it will be shown later) in the space of functions of bounded variation on Ω, which is defined by

$$BV(\Omega) = \{u \in L^1(\Omega) : J_0(u) < +\infty\}. \tag{3.59}$$

This motivates the following definition of $\mathcal{C}$

$$\mathcal{C} = \{u \in BV(\Omega) \mid u \geq 0 \text{ almost everywhere}\}. \tag{3.60}$$

Let now give a number of results and definitions regarding $BV(\Omega)$; further background and details on BV spaces can be found in [17, 19]. First of all, $BV(\Omega)$ is a Banach space with respect to the norm

$$\|u\|_{BV} = \|u\|_1 + J_0(u).$$

$\sqrt{x} \le \sqrt{x+\beta} \le \sqrt{x} + \sqrt{\beta}$ for $\beta, x \ge 0$, yields

$$J_0(u) \le J_\beta(u) \le J_0(u) + \sqrt{\beta}|\Omega|. \tag{3.61}$$

Inequality (3.61) will allow to assume, without loss of generality, that $\beta = 0$ in several of the arguments.

With this norm, two useful definitions for this section are

Definition 3.4.1. *A set $\mathcal{S} \subset BV(\Omega)$ is said to be BV-*bounded *if there exists $M > 0$ such that $\|u\|_{BV} \le M$ for all $u \in \mathcal{S}$.*

Definition 3.4.2. *A functional $T : L^p(\Omega) \to \mathbb{R}$ is said to be BV-*coercive *if*

$$T(u) \to +\infty \quad \textit{whenever} \quad \|u\|_{BV} \to +\infty. \tag{3.62}$$

Note that $BV(\Omega) \subset L^1(\Omega)$, by definition. Also, as a consequence of the following theorem, whose proof is found in [2], $BV(\Omega) \subset L^p(\Omega)$ for $1 \le p \le d/(d-1)$, where $d/(d-1) \overset{\text{def}}{=} +\infty$ for $d = 1$.

Theorem 3.4.3. *Let $\mathcal{S}$ be a BV-bounded set of functions. Then $\mathcal{S}$ is relatively compact, i.e. its closure is compact, in $L^p(\Omega)$ for $1 \le p < d/(d-1)$. $\mathcal{S}$ is bounded and thus relatively weakly compact for dimensions $d \ge 2$ in $L^p(\Omega)$ for $p = d/(d-1)$.*

Let define

$$R_\alpha(A; z+\gamma) = \arg\min_{u \in \mathcal{C}} \left\{ T_0(Au; z+\gamma) + \alpha J_\beta(u) \right\}. \tag{3.63}$$

The task is to show that $R_\alpha(A, z+\gamma)$ defines a regularization scheme.

3.4.1 $R_\alpha(A, z+\gamma)$ is Well-Defined

In order to prove the existence and uniqueness of solutions of (3.63), (3.60), the following theorem, which is similar to [2, Theorem 3.1] will be useful.

Theorem 3.4.4. *If $T : L^p(\Omega) \to \mathbb{R}$ is convex and BV-coercive, and $1 \le p \le d/(d-1)$, then T has a minimizer over $\mathcal{C}$. If T is strictly convex there is a unique minimizer.*

Proof. Let $\{u_n\} \subset \mathcal{C}$ be such that $T(u_n) \to \inf_{u \in \mathcal{C}} T(u)$. Then $T(u_n)$ is bounded, and hence, by (3.62), $\{u_n\}$ is BV-bounded. Theorem 3.4.3 implies that there exists a subsequence $\{u_{n_j}\}$ that converges to some $\hat{u} \in L^p(\Omega)$. Convergence is weak if $p = d/(d-1)$. Since T is convex, it is weakly lower semi-continuous [35], and hence,

$$T(\hat{u}) \le \liminf T(u_{n_j}) = \lim T(u_n) = T_*,$$

where T_* is the infimum of T on $\mathcal{C}$. Thus $\hat{u}$ minimizes T on $\mathcal{C}$. Uniqueness follows immediately if T is a strictly convex functional and $\mathcal{C}$ is a convex set. □

Using Theorem 3.4.4, requires showing that T_α is both strictly convex and BV-coercive.

Lemma 3.4.5. *T_α is strictly convex on $\mathcal{C}$.*

Proof. It was shown above that T_0 is strictly convex since A is nonsingular. The strict convexity of T_α then follows immediately from the fact that J_β is convex, which is proved in [2]. □

Lemma 3.4.6. *T_α is BV-coercive on $\mathcal{C}$.*

Proof. By (3.62), if $\|u\|_{BV} \to +\infty$, then $T_\alpha(u) \to +\infty$. A straightforward computation yields the following decomposition of a function $u \in BV(\Omega)$:

$$u = v + w, \tag{3.64}$$

where

$$w = \left(\int_\Omega u\, dx \Big/ |\Omega| \right) \chi_\Omega, \quad \text{and} \quad \int_\Omega v\, dx = 0. \tag{3.65}$$

Here χ_Ω is the indicator function on Ω. It is shown in [2] that there exists $C_1 \in \mathbb{R}^+$ such that, for any $u = v + w$ in $BV(\Omega)$,

$$\|v\|_p \leq C_1 J_0(v), \tag{3.66}$$

for $1 \leq p \leq d/(d-1)$. Equation (3.66), the triangle inequality, and the fact that $J_0(w) = 0$ yield

$$\|u\|_{BV} \leq \|w\|_1 + (C_1 + 1) J_0(v). \tag{3.67}$$

Let $\{u_n\} \subset BV(\Omega)$ be a sequence with $u_n = v_n + w_n$ as above, and suppose that $\liminf T_\alpha(u_n) = K < +\infty$. Let $\{u_{n_k}\} \subset \{u_n\}$ be a subsequence such that $T_\alpha(u_{n_k}) \to \liminf T_\alpha(u_n) = K$. Then, since $T_\alpha(u_{n_k})$ is uniformly bounded, $\alpha J_0(v_{n_k}) \leq T_\alpha(u_{n_k}) - T_0(u_{\text{exact}})$ implies that $J_0(v_{n_k})$ also is uniformly bounded. Noting that

$$\|Aw_{n_k}\|_1 = \left(\|A\chi_\Omega\|_1 / |\Omega| \right) \|w_{n_k}\|_1 = C_2 \|w_{n_k}\|_1 \tag{3.68}$$

and $Au \geq 0$ for all $u \in \mathcal{C}$, Jensen's inequality together with (3.66) and (3.68) yields

$$\begin{aligned} T_\alpha(u_{n_k}) \;&\geq\; \|Au_{n_k} + \gamma + \sigma^2\|_1 - \|z + \gamma + \sigma^2\|_\infty \log \|Au_{n_k} + \gamma + \sigma^2\|_1, \qquad (3.69)\\ &\geq\; \|Aw_{n_k}\|_1 - \|Av_{n_k}\|_1 - (\gamma + \sigma^2)|\Omega| \\ &\quad - \|z + \gamma + \sigma^2\|_\infty \log \left(\|Aw_{n_k}\|_1 + \|Av_{n_k}\|_1 + (\gamma + \sigma^2)|\Omega| \right), \\ &\geq\; C_2 \|w_{n_k}\|_1 - \|A\|_1 C_1 J_0(v_{n_k}) - (\gamma + \sigma^2)|\Omega| \\ &\quad - \|z + \gamma + \sigma^2\|_\infty \log \left(C_2 \|w_{n_k}\|_1 + \|A\|_1 C_1 J_0(v_{n_k}) + (\gamma + \sigma^2)|\Omega| \right), \end{aligned}$$

$$\geq \; C_2 \|w_{n_k}\|_1 - M - \|z + \gamma + \sigma^2\|_\infty \log(C_2 \|w_{n_k}\|_1 + M), \tag{3.70}$$

where M is an upper bound for $\|A\|_1 C_1 J_0(v_{n_k}) + (\gamma + \sigma^2)|\Omega|$, and $\|A\|_1$ is the operator norm induced by the norm on $L^1(\Omega)$. Thus

$$\liminf \left(C_2\|w_{n_k}\|_1 - M - \|z+\gamma+\sigma^2\|_\infty \log(C_2\|w_{n_k}\|_1 + M)\right) \leq \liminf T_\alpha(u_{n_k}) = K.$$

If $\|w_{n_k}\|_1 \to +\infty$, then the limit inferior on the left would equal $+\infty$, so $\liminf \|w_{n_k}\|_1 < +\infty$. Let $\{u_{n_{k_j}}\} \subset \{u_{n_k}\}$ be a subsequence such that $\|w_{n_{k_j}}\|_1 \to \liminf \|w_{n_k}\|_1$. Then $\|w_{n_{k_j}}\|_1$ is uniformly bounded.

Since $\|u_{n_{k_j}}\|_{BV} \leq \|w_{n_{k_j}}\|_1 + (C_1+1)J_0(v_{n_{k_j}})$, $\|u_{n_{k_j}}\|_{BV}$ is uniformly bounded, which implies that $\liminf \|u_n\|_{BV}$ is finite.

Thus it has been shown that if $\liminf T_\alpha(u_n) < +\infty$, then $\liminf \|u_n\|_{BV}$ is finite. Therefore $\|u_n\|_{BV} \to +\infty$ implies $T_\alpha(u_n) \to +\infty$. □

Existence and uniqueness of solutions of (3.63), (3.60) now follows immediately. Thus the following is result of this subsection

Theorem 3.4.7. *$R_\alpha(A, z+\gamma)$ is well-defined.*

Before continuing, let note that in the denoising case, i.e. when A is the identity operator, the existence and uniqueness of minimizers of $T_\alpha(u)$ was proved in [3].

3.4.2 $R_\alpha(A, z+\gamma)$ is Continuous

Recalling that A_n and z_n in (1.11) satisfy the same assumptions as A and z, the above arguments give that $u_{\alpha,n}$ (defined in (1.12)) exists, and is unique if A_n is invertible.

Theorem 3.4.8. *For each $n \in \mathbb{N}$, let u_α be the unique minimizer of T_α over $\mathcal{C}$ and $u_{\alpha,n}$ a minimizer of $T_{\alpha,n}$ over $\mathcal{C}$. Suppose, furthermore, that*

1. *for any sequence $\{u_n\} \subset L^p(\Omega)$,*

$$\lim_{n\to+\infty} T_{\alpha,n}(u_n) = +\infty \quad \textit{whenever} \quad \lim_{n\to+\infty} \|u_n\|_{BV} = +\infty; \tag{3.71}$$

2. *given $M > 0$ and $\epsilon > 0$, there exists N such that*

$$|T_{\alpha,n}(u) - T_\alpha(u)| < \epsilon \quad \textit{whenever} \quad n \geq N, \ \|u\|_{BV} \leq M. \tag{3.72}$$

Then

$$u_{\alpha,n} \to u_\alpha \ \textit{in} \ L^p(\Omega) \tag{3.73}$$

for $1 \leq p < d/(d-1)$. If $d \geq 2$ and $p = d/(d-1)$ convergence is weak, i.e.

$$u_{\alpha,n} \rightharpoonup u_\alpha. \tag{3.74}$$

Proof. Note that $T_{\alpha,n}(u_{\alpha,n}) \leq T_{\alpha,n}(u_\alpha)$. This and (3.72), yield

$$\limsup T_{\alpha,n}(u_{\alpha,n}) \leq T_\alpha(u_\alpha) < +\infty. \tag{3.75}$$

Thus by (3.71), $\{u_{\alpha,n}\}$ is BV-bounded. By Theorem 3.4.3, there exists a subsequence $\{u_{\alpha,n_j}\}$ that converges (weakly) to some $\hat{u} \in L^p(\Omega)$. Furthermore, by the weak lower semicontinuity of T_α, (3.72), and (3.75) it follows

$$\begin{aligned} T_\alpha(\hat{u}) &\leq \liminf T_\alpha(u_{\alpha,n_j}), \\ &= \liminf (T_\alpha(u_{\alpha,n_j}) - T_{\alpha,n_j}(u_{\alpha,n_j})) + \liminf T_{\alpha,n_j}(u_{\alpha,n_j}), \\ &\leq T_\alpha(u_\alpha). \end{aligned}$$

Since u_α is the unique minimizer of T_α, $\hat{u} = u_\alpha$. Since every convergent subsequence of $\{u_{\alpha,n}\}$ converges to u_α, hence $u_{\alpha,n} \to u_\alpha$ (weakly if $p = d/(d-1)$). □

The following corollary of Theorem 3.4.8 is the stability result for (3.63), (3.60).

Corollary 3.4.9. *$R_\alpha(Au; z + \gamma)$ is continuous with respect to $L^p(\Omega)$ for $1 < p < d/(d-1)$ and is weakly continuous with respect to $L^p(\Omega)$ for $p = d/(d-1)$.*

Proof. Without loss of generality, due to (3.61), let assume $\beta = 0$ and show that conditions *(i)* and *(ii)* from Theorem 3.4.8 hold. Also, all limits are assumed to be taken as $n \to +\infty$.

For condition *(i)*, let proceed as in the proof of Lemma 3.4.6. Taking $u_{\alpha,n} = v_{\alpha,n} + w_{\alpha,n}$, suppose that $\liminf T_{\alpha,n}(u_{\alpha,n}) = K < +\infty$, and let $\{u_{\alpha,n_k}\} \subset \{u_{\alpha,n}\}$ be a subsequence such that $T_{\alpha,n_k}(u_{\alpha,n_k}) \to K$.

Note that $T_{0,n_k}(u_{0,n_k})$ is uniformly bounded below, since using the analogue of inequality (3.69) together with the properties of $x - c\log x$ for $c > 0$ yields $T_{0,n_k}(u_{0,n_k}) \geq \|z_{n_k} + \sigma^2\|_\infty - \|z_{n_k} + \sigma^2\|_\infty \log \|z_{n_k} + \sigma^2\|_\infty$, which is uniformly bounded below since $\|z_{n_k} - (z+\gamma)\|_\infty \to 0$ and $z \in L^\infty(\Omega)$. Thus $T_{\alpha,n_k}(u_{\alpha,n_k}) \geq T_{0,n_k}(u_{0,n_k}) + \alpha J_0(v_{\alpha,n_k})$ implies that $J_0(v_{\alpha,n_k})$ is uniformly bounded.

Since $J_0(v_{\alpha,n_k})$ is bounded, from (3.70),

$$T_{\alpha,n_k}(u_{\alpha,n_k}) \;\; \geq \;\; C_2\|w_{\alpha,n_k}\|_1 - M - \|z_{n_k} + \sigma^2\|_\infty \log\left(C_2\|w_{\alpha,n_k}\|_1 + M\right), \qquad (3.76)$$

where M is the upper bound on $\|A_n\|_1 C_1 J_0(v_{\alpha,n}) + (\gamma + \sigma^2)|\Omega|$ obtained using the uniform boundedness of both $\|A_n\|$ (Banach-Steinhaus) and $J_0(v_{\alpha,n_k})$. Since $\|z_n + \sigma^2\|_\infty$ is uniformly bounded and $\liminf T_{\alpha,n_k}(u_{\alpha,n_k}) = K$, there exists a subsequence $\{u_{\alpha,n_{k_j}}\} \subset \{u_{\alpha,n_k}\}$ such that $\|w_{\alpha,n_{k_j}}\|_1$ is uniformly bounded. Thus $\|u_{\alpha,n_{k_j}}\|_{BV} \leq \|w_{\alpha,n_{k_j}}\|_1 + (C_1 + 1)J_0(v_{\alpha,n_{k_j}})$ implies that $\|u_{\alpha,n_{k_j}}\|_{BV}$ is uniformly bounded, so $\liminf \|u_{\alpha,n}\|_{BV}$ is finite. It has been shown that $\liminf T_{\alpha,n}(u_{\alpha,n}) < +\infty$ implies $\liminf \|u_{\alpha,n}\|_{BV}$ is finite, so $\|u_{\alpha,n}\|_{BV} \to +\infty$ implies $T_{\alpha,n}(u_{\alpha,n}) \to +\infty$.

For condition *(ii)*, note that, using Jensen's inequality and the properties of the logarithm,

$$\begin{aligned} |T_{\alpha,n}(u) - T_\alpha(u)| \;\; &= \;\; \left| \int_\Omega \left((A_n - A)u - (z_n + \sigma^2)\log(A_n u + \gamma + \sigma^2)\right)\, dx \right. \\ &\qquad \left. + \int_\Omega \left((z + \gamma + \sigma^2)\log(Au + \gamma + \sigma^2)\right)\, dx \right|, \\ &\leq \;\; \|A_n - A\|_1 \|u\|_1 \\ &\qquad + \|z_n - (z+\gamma)\|_\infty \log(\|A_n\|_1\|u\|_1 + (\gamma + \sigma^2)|\Omega|) \\ &\qquad\quad + \|z + \gamma + \sigma^2\|_\infty \log \left\| \frac{(Au + \gamma + \sigma^2)}{(A_n u + \gamma + \sigma^2)} \right\|_1 . \end{aligned} \qquad (3.77)$$

By assumption, $\|A_n - A\|_1, \|z_n - (z+\gamma)\|_\infty \to 0$. Furthermore, by the Banach-Steinhaus Theorem, $\|A_n\|_1$ is uniformly bounded, and since it has been assumed that $\|u\|_{BV}$ is bounded, by Theorem 3.4.3 $\|u\|_1$ is bounded as well. Thus the first two terms on the right-hand side in (3.77) tend to zero as $n \to +\infty$. For the third term note that

$$\left\| \frac{Au+\gamma+\sigma^2}{A_n u+\gamma+\sigma^2} - 1 \right\|_1 \leq \left\| \frac{1}{A_n u+\gamma+\sigma^2} \right\|_1 \|A_n - A\|_1 \, \|u\|_1,$$

which converges to zero since $\|1/(A_n u+\gamma+\sigma^2)\|_1$ is bounded and $\|A_n - A\|_1 \to 0$. Thus $\log(\|(Au+\gamma+\sigma^2)/(Au_n+\gamma+\sigma^2)\|_1) \to \log(1) = 0$, and hence

$$|T_{\alpha,n}(u) - T_\alpha(u)| \to 0. \tag{3.78}$$

The desired result now follows from Theorem 3.4.8. □

Finally, the main result of this subsection now follows directly from Theorem 3.4.7 and Corollary 3.4.9.

3.4.3 $R_\alpha(A, z+\gamma)$ is Convergent

The task in this section is to prove that $R_\alpha(A, z+\gamma)$ is convergent, that is, Condition 2 of Definition 2.2.1 holds. For this, assuming that A is invertible so that u_{exact} is the unique solution of $Au = z$.

Theorem 3.4.10. *$R_\alpha(Au; z+\gamma)$ is convergent with respect to $L^p(\Omega)$ for $1 \leq p < d/(d-1)$ and is weakly convergent for $p = d/(d-1)$.*

Proof. Suppose $\alpha_n \to 0$ at a rate such that

$$(T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n})))/\,\alpha_n \tag{3.79}$$

is bounded. Again, due to (3.61), it suffices to consider the $\beta = 0$ case. Then since $u_{\alpha_n,n}$

minimizes $T_{\alpha_n,n}$, yields

$$T_{\alpha_n,n}(u_{\alpha_n,n}) \leq T_{\alpha_n,n}(u_{\text{exact}}). \tag{3.80}$$

Since $\{z_n\}$ and $\{A_n\}$ are uniformly bounded and $A_n \to A$ in the $L^1(\Omega)$ operator norm, $\{T_{\alpha_n,n}(u_{\text{exact}})\}$ is a bounded sequence, and (3.80) implies that $\{T_{\alpha_n,n}(u_{\alpha_n,n})\}$ is therefore also a bounded sequence.

Subtracting $T_{0,n}(u_{0,n})$ from each term in (3.80), dividing by α_n, and using the decomposition $u_{\alpha_n,n} = v_{\alpha_n,n} + w_{\alpha_n,n}$ yields

$$\begin{aligned}(T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n}))/\,\alpha_n + J_0(v_{\alpha_n,n}) \leq\ & (T_{0,n}(u_{\text{exact}}) - T_{0,n}(u_{0,n}))/\,\alpha_n \\ & + J_0(u_{\text{exact}}).\end{aligned} \tag{3.81}$$

By (3.79), the right-hand side of (3.81) is bounded, implying the left hand side is bounded. Since $T_{0,n}(u_{\alpha_n,n}) - T_{0,n}(u_{0,n})$ is nonnegative, $J_0(v_{\alpha_n,n})$ is therefore also bounded. The boundedness of $T_{\alpha_n,n}(u_{\alpha_n,n})$ together with (3.76) imply that $\|w_{\alpha_n,n}\|_1$ is bounded. The BV-boundedness of $\{u_{\alpha_n,n}\}$ then follows from (3.67).

In order to show that $\|u_{\alpha_n,n} - u_{\text{exact}}\|_p \to 0$ ($u_{\alpha_n,n} \rightharpoonup u_{\text{exact}}$ for $p = d/(d-1)$) let proceed by showing that every subsequence of $\{u_{\alpha_n,n}\}$ contains a subsequence that converges to u_{exact}. Every subsequence $\{u_{\alpha_{n_j},n_j}\}$ of $\{u_{\alpha_n,n}\}$ is BV-bounded since $\{u_{\alpha_n,n}\}$ is, and by Theorem 3.4.3, it is reasonable to assume that $\{u_{\alpha_{n_j},n_j}\}$ converges strongly (weakly for $p = d/(d-1)$) to some $\hat{u} \in L^p(\Omega)$. Then

$$\begin{aligned} T_0(\hat{u}) \quad = \quad & \int_\Omega (A(\hat{u} - u_{\alpha_{n_j},n_j}) + (A - A_{n_j})u_{\alpha_{n_j},n_j} + (z_{n_j} - (z+\gamma))\log(A\hat{u} + \gamma + \sigma^2))\,dx \\ & - \int_\Omega (z_{n_j} + \sigma^2)\log((A_{n_j}u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2))\,dx + T_{0,n_j}(u_{\alpha_{n_j},n_j}),\end{aligned}$$

which, as in previous arguments, yields

$$\begin{aligned}|T_{0,n_j}(u_{\alpha_{n_j},n_j}) - T_0(\hat{u})| \leq & \left| \int_\Omega A\left(\hat{u} - u_{\alpha_{n_j},n_j}\right)\, dx \right| \\ & + \|z_{n_j} - (z+\gamma)\|_\infty \log(\|A\|_1 \|\hat{u}\|_1 + \gamma|\Omega|) \\ & + \|z_{n_j} + \sigma^2\|_\infty \log \|(A_{n_j} u_{\alpha_{n_j},n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2)\|_1 \\ & + \|A - A_{n_j}\|_1 \|u_{\alpha_{n_j},n_j}\|_1.\end{aligned}$$

Then for $1 \leq p \leq d/d-1$,

$$\|z_{n_j} - (z+\gamma)\|_\infty \log(\|A\|_1 \|\hat{u}\|_1 + (\gamma + \sigma^2)|\Omega|) \to 0,$$

since $\|z_{n_j} - (z+\gamma)\|_\infty \to 0$ and $\log(\|A\|_1 \|\hat{u}\|_1 + (\gamma+\sigma^2)|\Omega|)$ is constant, and

$$\|A - A_{n_j}\|_1 \|u_{\alpha_{n_j},n_j}\|_1 \to 0$$

since $\|A - A_{n_j}\|_1 \to 0$ and $\|u_{\alpha_{n_j},n_j}\|_1$ is uniformly bounded.

Since A is a bounded linear operator and Ω is a set of finite measure, $F(u) = \int_\Omega Au\ dx$ is a bounded linear functional on $L^p(\Omega)$. The weak convergence of $\{u_{\alpha_{n_j},n_j}\}$ then implies $\int_\Omega Au_{\alpha_{n_j},n_j}\ dx \to \int_\Omega A\hat{u}\ dx$, which yields $\int_\Omega A\left(\hat{u} - u_{\alpha_{n_j},n_j}\right)\ dx \to 0$.

Since A is compact, it is completely continuous, i.e. the weak convergence of $u_{\alpha_{n_j},n_j}$ to $\hat{u}$ implies that $\|Au_{\alpha_{n_j},n_j} - A\hat{u}\|_1 \to 0$ (cf. [4, Prop. 3.3]). Thus, since $\left\|\frac{1}{A\hat{u}+\gamma+\sigma^2}\right\|_1$ is bounded, and

$$\begin{aligned}\left\| \frac{A_{n_j} u_{\alpha_{n_j},n_j} + \gamma + \sigma^2}{A\hat{u} + \gamma + \sigma^2} - 1 \right\|_1 & \leq \left\| \frac{1}{A\hat{u} + \gamma + \sigma^2} \right\|_1 \|A_{n_j} u_{\alpha_{n_j},n_j} - A\hat{u}\|_1, \\ & \leq \left\| \frac{1}{A\hat{u} + \gamma + \sigma^2} \right\|_1 \\ & \quad \times \left(\left\|A_{n_j} - A\right\|_1 \|u_{\alpha_{n_j},n_j}\|_1 + \|Au_{\alpha_{n_j},n_j} - A\hat{u}\|_1 \right),\end{aligned}$$

hence $\|z_{n_j} + \sigma^2\|_\infty \log \|(A_{n_j} u_{\alpha_{n_j}, n_j} + \gamma + \sigma^2)/(A\hat{u} + \gamma + \sigma^2)\|_1 \to 0$. Therefore

$$T_0(\hat{u}) = \lim_{n_j \to +\infty} T_{0,n_j}(u_{\alpha_{n_j}, n_j}).$$

Invoking (3.81), (3.79), and (3.78), respectively, yields

$$\lim_{n_j \to +\infty} T_{0,n_j}(u_{\alpha_{n_j}, n_j}) = \lim_{n_j \to +\infty} T_{0,n_j}(u_{0,n_j}) = \lim_{n_j \to +\infty} T_{0,n_j}(u_{\text{exact}}) = T_0(u_{\text{exact}}).$$

Thus $T_0(\hat{u}) = T_0(u_{\text{exact}})$, and, since u_{exact} is the unique minimizer of T_0, $\hat{u} = u_{\text{exact}}$. Therefore $\{u_{\alpha_{n_j}, n_j}\}$ converges strongly (weakly for $p = d/(d-1)$) to u_{exact} in $L^p(\Omega)$. □

The results obtained in this subsection are summarized in the following theorem.

Theorem 3.4.11. *$R_\alpha(A, z + \gamma)$ defined in (3.63) defines a regularization scheme.*

Proof. By Corollary 3.4.9 and Theorem 3.4.7 $R_\alpha(A, z + \gamma)$ is well-defined and continuous and therefore satisfies Condition 1 of Definition 2.2.1. Theorem 3.4.10 then gives Condition 2, and hence R_α defines a regularization scheme. □

Chapter 4

Numerical Method

4.1 An Efficient Numerical Method

In this section, the focus is on the computational problem of interest, which was last seen in the Introduction:

$$\arg\min_{\mathbf{u}\geq\mathbf{0}} \left\{ T_\alpha(\mathbf{u}) \stackrel{\text{def}}{=} T_0(\mathbf{u}) + \frac{\alpha}{2} J(\mathbf{u}) \right\}, \tag{4.1}$$

where

$$T_0(\mathbf{u}) \stackrel{\text{def}}{=} \sum_{i=1}^{N^2} ([\mathbf{A}\mathbf{u}]_i + \gamma + \sigma^2) - \sum_{i=1}^{N^2} (z_i + \sigma^2) \log([\mathbf{A}\mathbf{u}]_i + \gamma + \sigma^2), \tag{4.2}$$

and $J(\mathbf{u})$ is the regularization function, which is assumed to be nonnegative, convex, and $\nabla J(\mathbf{u})$ Lipschitz continous with Lipschitz constant L. Here $\mathbf{A} \in \mathbb{R}^{N^2 \times N^2}$, $\mathbf{u}$, $\mathbf{z} \in \mathbb{R}^{N^2}$ are assumed to have been obtained from a discretization (e.g. mid-point quadrature plus a collocation of indices) of the underlying operator equation $z = Au$, followed by a lexicographical ordering of unknowns. Also $\mathbf{z}$ is assumed to contain random noise.

For any image collected by a $N \times N$ CCD camera, the noise contained in $\mathbf{z}$ follows a well-known distribution [31]. In particular, $\mathbf{z}$ is assumed to be a realization of the random vector

$$\hat{\mathbf{z}} = \text{Poiss}(\mathbf{A}\mathbf{u}) + \text{Poiss}(\gamma \cdot \mathbf{1}) + N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{4.3}$$

Here $\mathbf{1}$ is an $N^2 \times 1$ vector of all ones. (4.3) means that $\hat{\mathbf{z}}$ is the sum of three random vectors: the first two are Poisson with Poisson parameter vectors $\mathbf{Au}$ and $\gamma \cdot \mathbf{1}$ respectively, and the third is Normal with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}$. Following arguments found in the Introduction, a first approximation of (4.3) is

$$\hat{\mathbf{z}} + \sigma^2 \cdot \mathbf{1} = \text{Poiss}(\mathbf{Au} + \boldsymbol{\gamma} + \boldsymbol{\sigma}^2), \tag{4.4}$$

where $\boldsymbol{\gamma} = \gamma\mathbf{1}$ and $\boldsymbol{\sigma}^2 = \sigma^2\mathbf{1}$, which has probability density function

$$p_{\mathbf{z}}(\mathbf{z}|\mathbf{u}) := \prod_{i=1}^{N^2} \frac{([\mathbf{Au}]_i + \gamma + \sigma^2)^{z_i+\sigma^2} \exp[-([\mathbf{Au}]_i + \gamma + \sigma^2)]}{(z_i + \sigma^2)!}. \tag{4.5}$$

Even though Poisson random variables take on only discrete values, $p_{\mathbf{z}}(\mathbf{z}|\mathbf{u})$ should, in theory, be positive only for $\mathbf{z} \in \mathbb{Z}_+^{N^2}$. However to ease in both analysis and computation, $p_{\mathbf{z}}$ is a probability density defined on $\mathbb{R}_+^{N^2} \cup \{0\}$.

In the Bayesian approach, a prior probability density $p_{\mathbf{u}}(\mathbf{u})$ for $\mathbf{u}$ is also specified, and the posterior density

$$p_{\mathbf{u}}(\mathbf{u}|\mathbf{z}) := \frac{p_{\mathbf{z}}(\mathbf{z}|\mathbf{u})p_{\mathbf{u}}(\mathbf{u})}{p_{\mathbf{z}}(\mathbf{z})}, \tag{4.6}$$

given by Bayes' Law, is maximized with respect to $\mathbf{u}$. The maximizer of $p_{\mathbf{u}}(\mathbf{u};\mathbf{z})$ is called the maximum a posteriori (MAP) estimate. Since $p_{\mathbf{z}}(\mathbf{z})$ does not depend on $\mathbf{u}$, it is not needed in the computation of the MAP estimate and thus can be ignored. Maximizing (4.6) is equivalent to minimizing

$$\begin{aligned} T(\mathbf{u}) &= -\ln p_{\mathbf{z}}(\mathbf{z}|\mathbf{u}) - \ln p_{\mathbf{u}}(\mathbf{u}) \\ &= T_0(\mathbf{u}) - \ln p_{\mathbf{u}}(\mathbf{u}). \end{aligned} \tag{4.7}$$

with respect to $\mathbf{u}$. The $\ln(z_i+\sigma^2)$ term is dropped from the summation for notational simplicity since this term has no effect on the corresponding minimization problem. Comparing (4.7) and (4.1), it is clear that $-\ln p_{\mathbf{u}}(\mathbf{u})$ corresponds to the regularization term in the classical penalized likelihood approach to regularization. However in the Bayesian setting, $p_{\mathbf{u}}(\mathbf{u})$ is the probability density known as the *prior* from which the unknown $\mathbf{u}$ is assumed to arise. Thus the prior knowledge regarding the characteristics of $\mathbf{u}$ can be formulated in the form of a probability density $p_{\mathbf{u}}(\mathbf{u})$, this yields to a natural, and statistically rigorous, motivation for the regularization method. In this chapter, one of the goal is to provide a statistical interpretation for the use of standard Tikhonov, ℓ^2-norm of the gradient, and total variation regularization for the negative-log prior $-\ln p_{\mathbf{u}}(\mathbf{u})$.

4.2 Statistical Interpretations of Regularization

Standard Tikhonov regularization corresponds to the following choice of negative-log prior:

$$-\ln p_{\mathbf{u}}(\mathbf{u}) = \frac{\alpha}{2}\|\mathbf{u}\|_2^2.$$

This corresponds to the assumption that the prior for $\mathbf{u}$ is a zero-mean Gaussian random vector with covariance matrix $\alpha^{-1}\mathbf{I}$, which has the effect of penalizing reconstructions with large ℓ^2-norm.

For ℓ^2-norm of the gradient regularization, the negative-log penalty has the similar form

$$-\ln p_{\mathbf{u}}(\mathbf{u}) = \frac{\alpha}{2}\|\mathbf{D}\mathbf{u}\|_2^2,$$

where $\mathbf{D} = [\mathbf{\Gamma}_x, \mathbf{\Gamma}_y]^T$ is a discrete gradient operator. The discretization of the gradient that is used to obtain $\mathbf{D}$, or of the Laplacian to obtain $\mathbf{D}^T\mathbf{D}$, determines the form of the covariance matrix $\alpha^{-1}(\mathbf{D}^T\mathbf{D})^\dagger$, where "$\dagger$" denotes psuedo-inverse. In this case, following [21], the solution is assumed to be a "differentially Gaussian" random vector. The use of this regularization function has the effect of penalizing reconstructions that aren't smooth.

For total variation, $\mathbf{u}$ is a two-dimensional array. In [21], it is noted that for a large number of two-dimensional signals $[\mathbf{u}]_{i,j} = u_{i,j}$, the values of $u_{i+1,j} - u_{i,j}$ and $u_{i,j+1} - u_{i,j}$ tend to be realistically modeled by a Laplace distribution with mean 0, which has the form

$$p(x) = \frac{\alpha}{2} e^{-\alpha|x|}.$$

An analysis of 4000 digital images in [23] provides further support for this claim. The Laplace distribution has heavy tails, which means that the probability of $u_{i+1,j} - u_{i,j}$ and $u_{i,j+1} - u_{i,j}$ being large (such as is the case at a jump discontinuity in the image) is not prohibitively small.

The main result of [21] is that for one-dimensional signals, the total variation method of [26], with appropriate choice of regularization parameter, provides the MAP estimate given the assumptions that the measurement noise is i.i.d. Gaussian and that adjacent pixel differences are independent and satisfy a Laplacian distribution. For images of dimensions two and higher, however, an analogous result has not been given. This is due in large part to the fact that the one-dimensional Laplacian distribution does not extend in a straightforward manner – as do the Gaussian and Poisson distributions – to the multivariate case. To see this, recall that in [15] it is shown that if

$$[\mathbf{\Gamma u}]_{i,j} = ([\mathbf{\Gamma}_x \mathbf{u}]_{i,j}, [\mathbf{\Gamma}_y \mathbf{u}]_{i,j}) := (u_{i+1,j} - u_{i,j}, u_{i,j+1} - u_{i,j})$$

is an i.i.d. Laplacian random vector, its probability density function is given by

$$\frac{\alpha^2}{2\pi} K_0 \left(\alpha \sqrt{[\mathbf{\Gamma}_x \mathbf{u}]_{i,j}^2 + [\mathbf{\Gamma}_y \mathbf{u}]_{i,j}^2} \right),$$

where K_0 is the order zero, modified second kind Bessel function. Assuming independence, this yields the negative-log prior

$$-\ln p_{\mathbf{u}}(\mathbf{u}) = c - \sum_{i,j=1}^{N} \ln K_0 \left(\alpha \sqrt{[\mathbf{\Gamma}_x \mathbf{u}]_{i,j}^2 + [\mathbf{\Gamma}_y \mathbf{u}]_{i,j}^2} \right) \tag{4.8}$$

where c is a constant. Ignoring c and using the approximation

$$-\ln K_0(x) \approx x,$$

whose accuracy is illustrated in Figure 4.1, hence

$$-\ln p_{\mathbf{u}}(\mathbf{u}) \approx \alpha \sum_{i,j=1}^{N} \sqrt{[\boldsymbol{\Gamma}_x \mathbf{u}]_{i,j}^2 + [\boldsymbol{\Gamma}_y \mathbf{u}]_{i,j}^2},$$

which is the discrete total variation function.

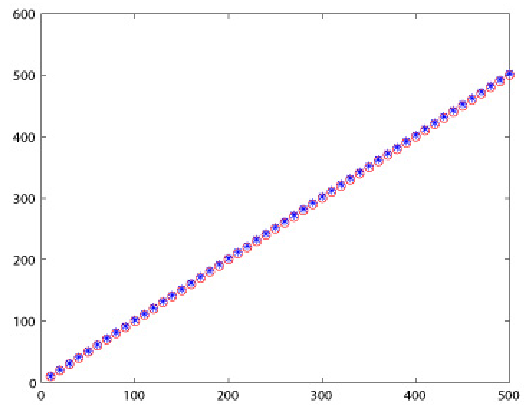

Figure 4.1: Plot of x (o) and $-\ln K_0(x)$ ($*$) on the interval $[0, 500]$.

4.3 An Analysis of the Posterior Density Function

For notational convenience, (4.7) is

$$T_\alpha(\mathbf{u}) = T_0(\mathbf{u}) + \alpha J(\mathbf{u}), \tag{4.9}$$

where, in the case of total variation, $J(\mathbf{u})$ is modify so that it has the form

$$J_\beta(\mathbf{u}) := \sum_{i,j=1}^{N} \sqrt{[\boldsymbol{\Gamma}_x \mathbf{u}]_{i,j}^2 + [\boldsymbol{\Gamma}_y \mathbf{u}]_{i,j}^2 + \beta}, \tag{4.10}$$

with $\beta > 0$ included to ensure the differentiability of J_β.

The gradient and Hessian of $T_\alpha(\mathbf{u})$ are given by

$$\begin{aligned} \nabla T_\alpha(\mathbf{u}) &= \nabla T_0(\mathbf{u}) + \alpha \nabla J(\mathbf{u}), \\ \nabla^2 T_\alpha(\mathbf{u}) &= \nabla^2 T_0(\mathbf{u}) + \alpha \nabla^2 J(\mathbf{u}). \end{aligned} \tag{4.11}$$

The gradient and Hessian of the Poisson likelihood functional $T_0(\mathbf{u})$ have expressions

$$\nabla T_0(\mathbf{u}) = \mathbf{A}^T \left(\frac{\mathbf{Au} - (\mathbf{z} - \gamma)}{\mathbf{Au} + \gamma + \sigma^2} \right), \tag{4.12}$$

$$\nabla^2 T_0(\mathbf{u}) = \mathbf{A}^T \text{diag} \left(\frac{\mathbf{z} + \sigma^2}{(\mathbf{Au} + \gamma + \sigma^2)^2} \right) \mathbf{A}, \tag{4.13}$$

where diag($\mathbf{v}$) is the diagonal matrix with $\mathbf{v}$ as its diagonal. Here, and in what follows, $\mathbf{x}/\mathbf{y}$ is used, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^N$, to denote Hadamard, or component-wise, division, and $\mathbf{x}^2 := \mathbf{x} \odot \mathbf{x}$, where " $\odot$ " denotes the Hadamard product.

Note that for moderate to large values of σ^2, say $\sigma^2 \geq 3^2$, it is extremely unlikely for z_i to be negative. Then, since Poisson random variables take on only nonnegative integer values, the random vector $\mathbf{z} + \sigma^2 \mathbf{1}$ is also highly unlikely to have nonpositive components. Assuming that $\mathbf{A}$ is positive definite and that $\mathbf{Au} \geq \mathbf{0}$ whenever $\mathbf{u} \geq \mathbf{0}$, it immediately follows that $\nabla^2 T_0(\mathbf{u})$ is positive definite for all $\mathbf{u} \geq \mathbf{0}$, and hence T_0 is strictly convex on $\mathbf{u} \geq \mathbf{0}$.

The gradient and Hessian of J for Tikhonov regularization are given by $\nabla J(\mathbf{u}) = \mathbf{u}$ and $\nabla^2 J(\mathbf{u}) = \mathbf{I}$, respectively, whereas for ℓ^2-norm of the gradient regularization, they are given by $\nabla J(\mathbf{u}) = \mathbf{D}^T \mathbf{D} \mathbf{u}$ and $\nabla^2 J(\mathbf{u}) = \mathbf{D}^T \mathbf{D}$, respectively. Moreover, both choices clearly yield convex, Lipschitz continuous functions.

In the case of total variation the gradient and Hessian computations are somewhat involved.

$$\nabla J_\beta(\mathbf{u}) = \mathbf{L}_1(\mathbf{u})\mathbf{u}, \tag{4.14}$$

$$\nabla^2 J_\beta(\mathbf{u}) = \mathbf{L}_1(\mathbf{u}) + 2\mathbf{L}_2(\mathbf{u}), \tag{4.15}$$

where, if $\psi(t) := \sqrt{t+\beta}$, $\mathbf{\Gamma u}^2 := (\mathbf{\Gamma}_x\mathbf{u})^2 + (\mathbf{\Gamma}_y\mathbf{u})^2$, and $\mathbf{\Gamma}_{xy}\mathbf{u} := \mathbf{\Gamma}_x\mathbf{u} \odot \mathbf{\Gamma}_y\mathbf{u}$,

$$\mathbf{L}_1(\mathbf{u}) = \begin{bmatrix} \mathbf{\Gamma}_x \\ \mathbf{\Gamma}_y \end{bmatrix}^T \begin{bmatrix} \text{diag}(\psi'(\mathbf{\Gamma u}^2)) & \mathbf{0} \\ \mathbf{0} & \text{diag}(\psi'(\mathbf{\Gamma u}^2)) \end{bmatrix} \begin{bmatrix} \mathbf{\Gamma}_x \\ \mathbf{\Gamma}_y \end{bmatrix}, \tag{4.16}$$

$$\mathbf{L}_2(\mathbf{u}) = \begin{bmatrix} \mathbf{\Gamma}_x \\ \mathbf{\Gamma}_y \end{bmatrix}^T \begin{bmatrix} \text{diag}((\mathbf{\Gamma}_x\mathbf{u})^2 \odot \psi''(\mathbf{\Gamma u}^2)) & \text{diag}(\mathbf{\Gamma}_{xy}\mathbf{u} \odot \psi''(\mathbf{\Gamma u}^2)) \\ \text{diag}(\mathbf{\Gamma}_{xy}\mathbf{u} \odot \psi''(\mathbf{\Gamma u}^2)) & \text{diag}((\mathbf{\Gamma}_y\mathbf{u})^2 \odot \psi''(\mathbf{\Gamma u}^2)) \end{bmatrix} \begin{bmatrix} \mathbf{\Gamma}_x \\ \mathbf{\Gamma}_y \end{bmatrix}.$$

For a more detailed treatment of these computations see [33]. Since $\nabla^2 J_\beta(\mathbf{u})$ is positive semi-definite for all $\mathbf{u}$, and hence, J_β is a convex function. The Lipschitz continuity of ∇J_β follows from (4.15).

It is now time to prove that T_α has a unique minimizer on $\mathbf{u} \geq \mathbf{0}$. This follows if T_α is strictly convex and coercive on $\mathbf{u} \geq \mathbf{0}$ [33, Chapter 2]. Recall that T_α is coercive on $\mathbf{u} \geq \mathbf{0}$ provided

$$\|\mathbf{u}\|_2 \to \infty \qquad \text{implies} \qquad T_\alpha(\mathbf{u}) \to \infty.$$

Theorem 4.3.1. *Assume that* $\mathbf{z} + \sigma^2\mathbf{1} > \mathbf{0}$, $\mathbf{A}$ *is positive definite and* $\mathbf{Au} \geq \mathbf{0}$ *for all* $\mathbf{u} \geq \mathbf{0}$. *Then* T_α *is strictly convex and coercive on* $\mathbf{u} \geq \mathbf{0}$, *and hence has a unique nonnegative minimizer.*

Proof. First, noting that given the assumptions and discussion above, T_0 is strictly convex on $\mathbf{u} \geq \mathbf{0}$. It has been also argued that J is convex. Thus T_α is strictly convex on $\mathbf{u} \geq \mathbf{0}$.

The coercivity of T_α in all cases is proved using the following application of Jensen's inequality:

$$T_0(\mathbf{u}) \geq \|\mathbf{Au} + (\gamma + \sigma^2)\mathbf{1}\|_1 - \|\mathbf{z} + \sigma^2\|_\infty \ln \|\mathbf{Au} + (\gamma + \sigma^2)\mathbf{1}\|_1 \tag{4.17}$$

for $\mathbf{u} \geq \mathbf{0}$. Since $\mathbf{A}$ is positive definite $\|\mathbf{u}\|_2 \to \infty$ implies that $\|\mathbf{Au}\|_1 \to \infty$, which in turn implies, via (4.17), that $T_0(\mathbf{u}) \to \infty$. Coercivity then follows from the fact that $J(\mathbf{u}) \geq \mathbf{0}$ for all $\mathbf{u}$.

It results, finally, from the fact that $\mathbf{u} \geq \mathbf{0}$ is a convex set. □

4.3.1 A Weighted Least Squares Approximation of $T_0(\mathbf{u}, \mathbf{z})$

First, a computation of various derivatives of T_0 is given. The gradient and Hessian of T_0 with respect to $\mathbf{u}$ are given by

$$\begin{aligned} \nabla_{\mathbf{u}} T_0(\mathbf{u};\mathbf{z}) &= \mathbf{A}^T \left(\frac{\mathbf{Au} - (\mathbf{z} - \gamma)}{\mathbf{Au} + \gamma + \boldsymbol{\sigma}^2} \right), &(4.18)\\ \nabla^2_{\mathbf{uu}} T_0(\mathbf{u};\mathbf{z}) &= \mathbf{A}^T \mathrm{diag} \left(\frac{\mathbf{z} + \boldsymbol{\sigma}^2}{(\mathbf{Au} + \gamma + \boldsymbol{\sigma}^2)^2} \right) \mathbf{A}, &(4.19) \end{aligned}$$

where division – here and for the remainder of the manuscript – is computed component-wise. The gradient and Hessian of T_0 with respect to $\mathbf{z}$ are given by

$$\begin{aligned} \nabla_{\mathbf{z}} T_0(\mathbf{u};\mathbf{z}) &= -\log(\mathbf{Au} + \boldsymbol{\gamma} + \boldsymbol{\sigma}^2), &(4.20)\\ \nabla^2_{\mathbf{zz}} T_0(\mathbf{u};\mathbf{z}) &= \mathbf{0}. &(4.21) \end{aligned}$$

The second order mixed partial derivatives of T_0 are given by

$$\begin{aligned} \nabla^2_{\mathbf{uz}} T_0(\mathbf{u};\mathbf{z}) &= -\mathbf{A}^T \mathrm{diag} \left(\frac{\mathbf{1}}{\mathbf{Au} + \boldsymbol{\gamma} + \boldsymbol{\sigma}^2} \right), &(4.22)\\ \nabla^2_{\mathbf{zu}} T_0(\mathbf{u};\mathbf{z}) &= -\mathrm{diag} \left(\frac{\mathbf{1}}{\mathbf{Au} + \boldsymbol{\gamma} + \boldsymbol{\sigma}^2} \right) \mathbf{A}. &(4.23) \end{aligned}$$

Now, let $\mathbf{u}_e$ be the exact object and $\mathbf{z}_e \stackrel{\text{def}}{=} \mathbf{Au}_e + \boldsymbol{\gamma}$ the background shifted exact data. Then, letting $\mathbf{k} = \mathbf{z} - \mathbf{z}_e$ and $\mathbf{h} = \mathbf{u} - \mathbf{u}_e$ and expanding T_0 in a Taylor series about $\mathbf{u}_e$ and $\mathbf{z}_e$, gives from (4.18)-(4.23)

$$\begin{aligned} T_0(\mathbf{u};\mathbf{z}) &= T_0(\mathbf{u}_e + \mathbf{h}; \mathbf{z}_e + \mathbf{k}),\\ &= T_0(\mathbf{u}_e;\mathbf{z}_e) - \mathbf{k}^T \nabla_{\mathbf{z}} T_0(\mathbf{u}_e;\mathbf{z}_e) + \frac{1}{2}\mathbf{h}^T \nabla^2_{\mathbf{uu}} T_0(\mathbf{u}_e;\mathbf{z}_e)\mathbf{h}\\ &\quad + \frac{1}{2}\mathbf{k}^T \nabla^2_{\mathbf{uz}} T_0(\mathbf{u}_e;\mathbf{z}_e)\mathbf{h} + \frac{1}{2}\mathbf{h}^T \nabla^2_{\mathbf{zu}} T_0(\mathbf{u}_e;\mathbf{z}_e)\mathbf{k} + \mathcal{O}(\|\mathbf{h}\|_2^3, \|\mathbf{k}\|_2^3)\\ &= \frac{1}{2}(\mathbf{Au} - (\mathbf{z} - \boldsymbol{\gamma}))^T \mathrm{diag} \left(\frac{\mathbf{1}}{\mathbf{Au}_e + \boldsymbol{\gamma} + \boldsymbol{\sigma}^2} \right) (\mathbf{Au} - (\mathbf{z} - \boldsymbol{\gamma}))\\ &\quad + T_0(\mathbf{u}_e;\mathbf{z}) + \mathcal{O}(\|\mathbf{h}\|_2^3, \|\mathbf{k}\|_2^3). \end{aligned}$$

Thus the quadratic Taylor series approximation of $T_0(\mathbf{u};\mathbf{z})$ about the points $(\mathbf{u}_e;\mathbf{z}_e)$ is given by

$$T_0(\mathbf{u}_e;\mathbf{z}) + \frac{1}{2}(\mathbf{A}\mathbf{u} - (\mathbf{z}-\boldsymbol{\gamma}))^T \mathrm{diag}\left(\frac{\mathbf{1}}{\mathbf{A}\mathbf{u}_e + \boldsymbol{\gamma} + \boldsymbol{\sigma}^2}\right)(\mathbf{A}\mathbf{u} - (\mathbf{z}-\boldsymbol{\gamma})). \tag{4.24}$$

This Taylor series connection between the negative-log of the Poisson likelihood and (4.24) has, to the knowledge of the authors, not been noted elsewhere.

It is important to emphasize that the quadratic approximation (4.24) of T_0 will be accurate provided $\|\mathbf{h}\|_2$ and $\|\mathbf{k}\|_2$ are small *relative to* $\|\mathbf{u}_e\|_2$ *and* $\|\mathbf{z}_e\|_2$. Since this will hold in practice for typical data $\mathbf{z}$ and reasonable approximations of $\mathbf{u}$, it is reasonable to expect that (4.24) will be an accurate approximation of $T_0(\mathbf{u};\mathbf{z})$ in a region of $(\mathbf{u}_e;\mathbf{z}_e)$ that is not restrictively small.

In practice, if this approximation is to be used, $\mathbf{A}\mathbf{u}_e$ must in turn be approximated. The natural choice is to use $\mathbf{z}+\boldsymbol{\sigma}^2$ instead since $E(\mathbf{z}) = \mathbf{A}\mathbf{u}_e + \boldsymbol{\gamma}$, where E is the expected value of the random variable $\mathbf{z}$. This yields the following weighted least squares approximation of the Poisson likelihood function

$$T_{\mathrm{wls}}(\mathbf{u};\mathbf{z}) \stackrel{\mathrm{def}}{=} \frac{1}{2}\|\mathbf{C}^{-1/2}(\mathbf{A}\mathbf{u} - (\mathbf{z}-\boldsymbol{\gamma}))\|_2^2, \quad \text{where} \quad \mathbf{C} = \mathrm{diag}(\mathbf{z}+\boldsymbol{\sigma}^2). \tag{4.25}$$

Note that the constant term $T_0(\mathbf{u}_e;\mathbf{z})$ has been dropped since it does not effect the computations. In a large number of applications in which Poisson data is analyzed, T_0 is approximated by a weighted least squares function. The analysis above suggests that T_{wls} is the natural choice.

4.4 A Nonnegatively Constrained Convex Programming Method

In this section, the aim is to give an outline of a computationally efficient method for solving (4.1), following the approach set forth in [10], which was developed for use on Tikhonov-regularized Poisson likelihood estimation problems.

4.4.1 Preliminaries

The projection of a vector $\mathbf{u} \in \mathbb{R}^{N^2}$ onto the feasible set $\{\mathbf{u} \in \mathbb{R}^{N^2} \mid \mathbf{u} \geq \mathbf{0}\}$ can be conveniently expressed as

$$\mathcal{P}(\mathbf{u}) \stackrel{\text{def}}{=} \arg\min_{\mathbf{v}\in\Omega} ||\mathbf{v} - \mathbf{u}||_2 = \max\{\mathbf{u}, \mathbf{0}\},$$

where $\max\{\mathbf{u}, \mathbf{0}\}$ is the vector whose ith component is zero if $u_i < 0$ and is u_i otherwise. The active set for a vector $\mathbf{u} \geq \mathbf{0}$ is defined

$$\mathcal{A}(\mathbf{u}) = \{i \mid u_i = 0\},$$

and the complementary set of indices, $\mathcal{I}(\mathbf{u})$, is known as the inactive set.

The next step is to make some definitions that will be required in the discussion of the iterative method. The reduced gradient of T_α at $\mathbf{u} \geq \mathbf{0}$ is given by

$$[\nabla_{\text{red}} T_\alpha(\mathbf{u})]_i = \begin{cases} \frac{\partial T_\alpha(\mathbf{u})}{\partial u_i}, & i \in \mathcal{I}(\mathbf{u}) \\ 0, & i \in \mathcal{A}(\mathbf{u}), \end{cases}$$

the projected gradient of T_α by

$$[\nabla_{\text{proj}} T_\alpha(\mathbf{u})]_i = \begin{cases} \frac{\partial T_\alpha(\mathbf{u})}{\partial u_i}, & i \in \mathcal{I}(\mathbf{u}), \text{ or } i \in \mathcal{A}(\mathbf{u}) \text{ and } \frac{\partial T_\alpha(\mathbf{u})}{\partial u_i} < 0, \\ 0, & \text{otherwise}, \end{cases}$$

and the reduced Hessian by

$$[\nabla^2_{\text{red}} T_\alpha(\mathbf{u})]_{ij} = \begin{cases} \frac{\partial^2 T_\alpha(\mathbf{u})}{\partial u_i \partial u_j}, & \text{if } i \in \mathcal{I}(\mathbf{u}) \text{ and } j \in \mathcal{I}(\mathbf{u}) \\ \delta_{ij}, & \text{otherwise}. \end{cases}$$

Finally, $\mathbf{D}_\mathcal{I}(\mathbf{u})$ is defined to be the diagonal matrix with components

$$[\mathbf{D}_\mathcal{I}(\mathbf{u})]_{ii} = \begin{cases} 1, & i \in \mathcal{I}(\mathbf{u}) \\ 0, & i \in \mathcal{A}(\mathbf{u}), \end{cases} \tag{4.26}$$

and $\mathbf{D}_{\mathcal{A}}(\mathbf{u}) = \mathbf{I} - \mathbf{D}_{\mathcal{I}}(\mathbf{u})$. Note then that

$$\nabla_{\text{red}} T_\alpha(\mathbf{u}) = \mathbf{D}_{\mathcal{I}}(\mathbf{u})\, \nabla T_\alpha(\mathbf{u}), \tag{4.27}$$

$$\nabla^2_{\text{red}} T_\alpha(\mathbf{u}) = \mathbf{D}_{\mathcal{I}}(\mathbf{u})\, \nabla^2 T_\alpha(\mathbf{u})\, \mathbf{D}_{\mathcal{I}}(\mathbf{u}) \;+\; \mathbf{D}_{\mathcal{A}}(\mathbf{u}), \tag{4.28}$$

4.4.2 Gradient Projection Iteration

A key component of the iterative method introduced in [10] is the gradient projection iteration [25], which is presented now. Given $\mathbf{u}_k \geq \mathbf{0}$, $\mathbf{u}_{k+1}$ is computed via

$$\mathbf{p}_k = -\nabla T_\alpha(\mathbf{u}_k) \tag{4.29}$$

$$\lambda_k = \arg\min_{\lambda > 0} T_\alpha(\mathcal{P}(\mathbf{u}_k + \lambda \mathbf{p}_k)) \tag{4.30}$$

$$\mathbf{u}_{k+1} = \mathcal{P}(\mathbf{u}_k + \lambda_k \mathbf{p}_k) \tag{4.31}$$

In practice, subproblem (4.30) is solved inexactly using a projected backtracking line search. In the implementation used here, the initial step length parameter is

$$\lambda_k^0 = \frac{||\mathbf{p}_k||_2^2}{\langle \nabla^2 T_\alpha(\mathbf{u}_k)\mathbf{p}_k, \mathbf{p}_k \rangle}. \tag{4.32}$$

The quadratic backtracking line search algorithm found in [27] is then used to create a sequence of line search parameters $\{\lambda_k^j\}_{j=0}^m$, where m is the smallest positive integer such that the sufficient decrease condition

$$T_\alpha(\mathbf{u}_k(\lambda_k^j)) \leq T_\alpha(\mathbf{u}_k) - \frac{\mu}{\lambda_k^j} ||\mathbf{u}_k - \mathbf{u}_k(\lambda_k^j)||_2^2 \tag{4.33}$$

holds. Here $\mu \in (0, 1)$ and

$$\mathbf{u}_k(\lambda) = \mathcal{P}_\Omega(\mathbf{u}_k + \lambda \mathbf{p}_k). \tag{4.34}$$

The approximate solution of (4.30) is then taken to be λ_k^m.

The proof of a convergence result for gradient projection iterations now follows.

Theorem 4.4.1. *Under the same hypotheses as Theorem 4.3.1, the gradient projection iteration applied to (4.1) yields a sequence* $\{\mathbf{u}_k\}$ *that converges to the unique solution* $\mathbf{u}^*$*.*

Proof. First, let show that $\nabla T_\alpha(\mathbf{u})$ is Lipshitz continuous.

$$\begin{aligned}\|\nabla T_\alpha(\mathbf{u}) - \nabla T_\alpha(\mathbf{v})\|_2 &= \left\|\mathbf{A}^T\left(\frac{\mathbf{A}\mathbf{u}+\gamma-\mathbf{z}}{\mathbf{A}\mathbf{u}+\gamma+\sigma^2} - \frac{\mathbf{A}\mathbf{v}+\gamma-\mathbf{z}}{\mathbf{A}\mathbf{v}+\gamma+\sigma^2}\right) + \alpha(J(\mathbf{u}) - J(\mathbf{v}))\right\|_2 \\ &\leq \|\mathbf{A}\|_2 F(\mathbf{u},\mathbf{v}) + \alpha L\|\mathbf{u}-\mathbf{v}\|_2,\end{aligned}$$

where L is the Lipschitz constant for J and

$$\begin{aligned}F(\mathbf{u},\mathbf{v}) &= \left\|\frac{(\mathbf{A}(\mathbf{u}-\mathbf{v}))\odot(\mathbf{z}+\sigma^2)}{(\mathbf{A}\mathbf{u}+\gamma+\sigma^2)\odot(\mathbf{A}\mathbf{v}+\gamma+\sigma^2)}\right\|_2 \\ &\leq \|\mathbf{A}\|_2\left\|\frac{\mathbf{z}+\sigma^2}{(\gamma+\sigma^2)^2}\right\|_2 \|\mathbf{u}-\mathbf{v}\|_2.\end{aligned}$$

Hence,

$$\|\nabla T_\alpha(\mathbf{u}) - \nabla T_\alpha(\mathbf{v})\| \leq \left(\|\mathbf{A}\|^2\left\|\frac{\mathbf{z}+\sigma^2}{(\gamma+\sigma^2)^2}\right\| + \alpha L\right)\|\mathbf{u}-\mathbf{v}\|,$$

establishing that ∇T_α is Lipschitz continuous.

Then by [25, Theorem 5.4.5] the gradient projection iteration (4.29)-(4.31) is well defined. Moreover, every limit point $\bar{\mathbf{u}}$ of the gradient projection iterates satisfies $\nabla_{\text{proj}} T_\alpha(\overline{\mathbf{u}}) = \mathbf{0}$ [25, Theorem 5.4.6].

Now, $\{T_\alpha(\mathbf{u}_k)\}$ is a decreasing sequence that is bounded below by $T_\alpha(\mathbf{u}^*)$, where $\mathbf{u}^*$ is the unique solution of (4.1) given by Theorem 4.3.1. Hence it converges to some $\overline{T}_\alpha \geq T_\alpha(\mathbf{u}^*)$. Since T_α is coercive, $\{\mathbf{u}_k\}$ is bounded, and hence there exists a subsequence $\{\mathbf{u}_{k_j}\}$ converging to some $\overline{\mathbf{u}}$. By the results mentioned in the first paragraph of the proof, $\nabla_{\text{proj}} T_\alpha(\overline{\mathbf{u}}) = \mathbf{0}$, and since T_α is strictly convex, $\overline{\mathbf{u}} = \mathbf{u}^*$. Thus $T_\alpha(\mathbf{u}_k) \to T_\alpha(\mathbf{u}^*)$. The strict convexity and coercivity of T_α then implies [29, Corollary 27.2.2] that $\mathbf{u}_k \to \mathbf{u}^*$. □

4.4.3 The Reduced Newton Step and Conjugate Gradient

In practice, the gradient projection iteration is very slow to converge. However, a robust method with good convergence properties results if gradient projection iterations are interspersed with steps computed from the reduced Newton system

$$\nabla^2_{\text{red}} T_\alpha(\mathbf{u}_k)\, \mathbf{p} = -\nabla_{\text{red}} T_\alpha(\mathbf{u}_k). \tag{4.35}$$

Approximate solutions of (4.35) can be efficiently obtained using conjugate gradient iteration (CG) [28] applied to the problem of minimizing

$$q_k(\mathbf{p}) = T_\alpha(\mathbf{u}_k) + \langle \nabla_{\text{red}} T_\alpha(\mathbf{u}_k), \mathbf{p}\rangle + \frac{1}{2}\langle \nabla^2_{\text{red}} T_\alpha(\mathbf{u}_k)\, \mathbf{p}, \mathbf{p}\rangle. \tag{4.36}$$

The result is a sequence $\{\mathbf{p}_k^j\}$ that converges to the minimizer of (4.36). Even with rapid CG convergence, for large-scale problems it is important to choose effective stopping criteria to reduce overall computational cost. The following stopping criterion from Moré and Toraldo [27] is very effective:

$$q_k(\mathbf{p}_k^{j-1}) - q_k(\mathbf{p}_k^j) \le \gamma_{CG} \max\{q_k(\mathbf{p}_k^{i-1}) - q_k(\mathbf{p}_k^i) \,|\, i = 1, \dots, j-1\}, \tag{4.37}$$

where $0 < \gamma_{CG} < 1$. Then the approximate solution of (4.36) is taken to be the $\mathbf{p}_k^{m_{\text{CG}}}$ where m_{CG} is the smallest integer such that (4.37) is satisfied.

With $\mathbf{p}_k := \mathbf{p}_k^{m_{\text{CG}}}$, applying again a projected backtracking line search, only this time, the much less stringent acceptance criteria

$$T_\alpha(\mathbf{u}_k(\lambda_k^m)) < T_\alpha(\mathbf{u}_k). \tag{4.38}$$

is used.

4.4.4 Sparse Preconditioner

Since there is a conjugate gradient stage in the above algorithm, it is natural to introduce a preconditioner, yielding the proconditioned CG (PCG) iteration. The way in which the implementation of PCG for (4.1) that differs from those given in [6, 10] is in the definition of the preconditioner. However, the analysis nonetheless follow the general idea set forth in [10]; that is, a sparse preconditioner is used.

In practice, this amounts to create a sparse Hessian and use it as preconditioner. In particular, noting that

$$\nabla^2 T_\alpha(\mathbf{u}) = \mathbf{A}^T \operatorname{diag}(\mathbf{w}(\mathbf{u}))\ \mathbf{A} + \alpha \nabla^2 J(\mathbf{u}),$$

yields to the following definion of the sparse preconditioner

$$\mathbf{M}(\mathbf{u}) = \widehat{\mathbf{A}}^T \operatorname{diag}(\mathbf{w}(\mathbf{u}))\ \widehat{\mathbf{A}} + \alpha \nabla^2 J(\mathbf{u}), \tag{4.39}$$

where $\widehat{\mathbf{A}}$ is obtained by zeroing out all the "small" entries of $\mathbf{A}$ in the following fashion

$$[\widehat{\mathbf{A}}]_{ij} = \begin{cases} [\mathbf{A}]_{ij} & \text{if } [\mathbf{A}]_{ij} \geq \tau; \\ 0 & \text{otherwise.} \end{cases}$$

The truncation parameter τ is selected so that it this proportional to the largest entry of $\mathbf{A}$

$$\tau = r \max_{ij} [\mathbf{A}]_{ij}, \quad 0 < r < 1. \tag{4.40}$$

Note that in the cases of interest here, $\nabla^2 J(\mathbf{u})$ in (4.39) is sparse, so that $\mathbf{M}$ is a sparse matrix.

Now, implementing PCG at outer iteration k requires that solving a linear systems of the form

$$\mathbf{M}_{\text{red}}(\mathbf{u}_k)\mathbf{y} = \mathbf{z}$$

efficiently is reachable, where

$$\mathbf{M}_{\text{red}}(\mathbf{u}_k) = \mathbf{D}_{\mathcal{I}}(\mathbf{u}_k)\mathbf{M}(\mathbf{u}_k)\mathbf{D}_{\mathcal{I}}(\mathbf{u}_k) + \mathbf{D}_{\mathcal{A}}(\mathbf{u}_k)$$

for $\mathbf{y}, \mathbf{z} \in \mathbb{R}^{N^2}$. To do this, a way is to compute a Cholesky factorization of $\mathbf{M}_{\text{red}}(\mathbf{u}_k)$ at each outer iteration prior to beginning the inner PCG iterations.

4.4.5 The Numerical Algorithm

In the first stage of the algorithm a stopping criteria is needed for the gradient projection iterations. Borrowing from Moré and Toraldo [27] , gradient projection stops when

$$T_\alpha(\mathbf{u}_{k-1}) - T_\alpha(\mathbf{u}_k) \leq \gamma_{GP} \max\{T_\alpha(\mathbf{u}_{i-1}) - T_\alpha(\mathbf{u}_i) \,|\, i = 1, \ldots, k-1\}, \tag{4.41}$$

where $0 < \gamma_{GP} < 1$.

Gradient Projection-Reduced Newton-CG (GPRNCG) Algorithm

Step 0: Select initial guess $\mathbf{u}_0$, and set $k = 0$.

Step 1: Given $\mathbf{u}_k$.

(1) Take gradient projection steps until either (4.41) is satisfied or $\text{GP}_{\max}$ iterations have been computed. Return updated $\mathbf{u}_k$.

Step 2: Given $\mathbf{u}_k$.

(1) Do PCG iterations to approximately minimize the quadratic (4.36) until either (4.37) is satisfied or $\text{CG}_{\max}$ iterations have been computed. Return $\mathbf{p}_k = \mathbf{p}_k^{m_{\text{CG}}}$.

(2) Find λ_k^m that satisfies (4.38), and return $\mathbf{u}_{k+1} = \mathbf{u}_k(\lambda_k^m)$.

(3) Update $k := k + 1$ and return to Step 1.

Since at each outer GPRNCG iteration at least one gradient projection step, with sufficient decrease condition (4.38), is taken, Theorem 4.4.1 yields to the following result.

Theorem 4.4.2. *Under the same hypotheses as Theorem 4.3.1, the iterates* $\{\mathbf{u}_k\}$ *generated by GPRNCG are guaranteed to converge to the unique solution of problem (4.1).*

Proof. The GPRNCG iteration is well-defined since the gradient projection iteration (Step 1) is well-defined, and since the truncated CG search direction $\mathbf{p}_k$ given in Step 2.1 is guaranteed to be a descent direction [28]. In the proof of Theorem 4.4.1, [25, Theorem 5.4.6] is used. The analogous result holds for GPRNCG: every limit point of the sequence $\{\mathbf{u}_k\}$ generated by GPRNCG satisfies $\nabla_{\text{proj}} T_\alpha(\mathbf{u}) = \mathbf{0}$. The proof requires a minor modification of the proof of [25, Theorem 5.4.6]. Following the arguments given in the last paragraph of the proof of Theorem 4.4.1, the result follows. □

The Lagged-Diffusivity Modification for Total Variation

In the case of total variation, GPRNCG is much more efficient if the full Hessian is replaced by $\nabla^2 T_0(\mathbf{u}) + \alpha \mathbf{L}_1(\mathbf{u})$ (where $\mathbf{L}_1(\mathbf{u})$ is defined in (4.16)) within Step 2 of the algorithm, i.e. in (4.36) [6]. The name of this method is, gradient projection–reduced lagged diffusivity iteration (GPRLD).

4.5 Numerical Experiments

Finally, to finish the chapter, some numerical experiments that demonstrate the effectiveness of the various regularization methods in conjunction with the negative-log of the Poisson likelihood function. The tests are performed using the 64×64 simulated satellite seen on the left-hand side in Figure 4.2. Generating corresponding blurred noisy data requires a discrete PSF $\mathbf{a}$, which is computed using the Fourier optics [20] PSF model

$$\mathbf{a} = \left| \texttt{fft2}\left(\mathbf{p} \odot e^{\hat{\imath} \boldsymbol{\phi}} \right) \right|^2,$$

where $\mathbf{p}$ is the $N \times N$ indicator array for the telescopes pupil; " $\odot$ " denotes Hadamard (component-wise) product; $\boldsymbol{\phi}$ is the $N \times N$ array that represents the aberrations in the incoming wavefronts of light; $\hat{\imath} = \sqrt{-1}$; and `fft2` denotes the two-dimensional discrete Fourier

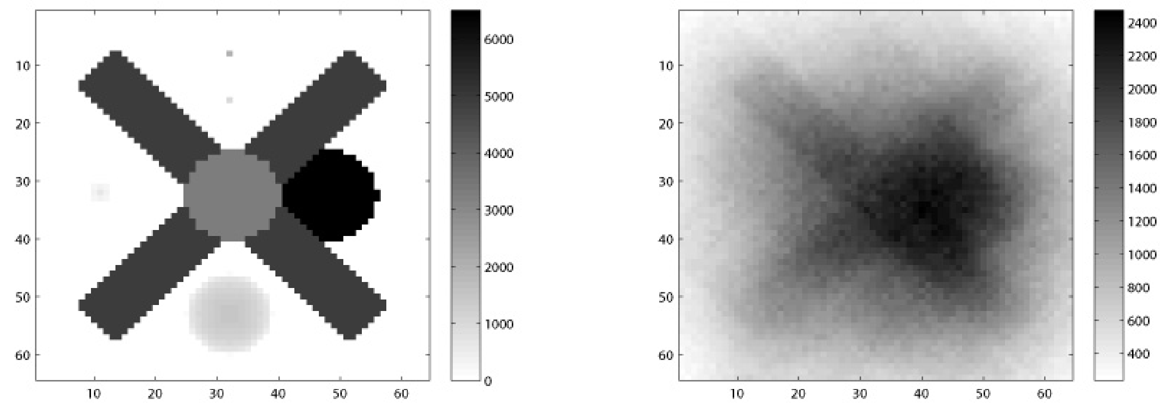

Figure 4.2: *On the left is the true object* $\mathbf{u}_{\text{true}}$. *On the right, is the blurred, noisy image* $\mathbf{z}$.

transform. The $64^2 \times 64^2$ blurring matrix $\mathbf{A}$ is then defined by

$$\mathbf{Au} = \texttt{ifft2}\left(\hat{\mathbf{a}} \odot (\texttt{fft2}(\mathbf{u}))\right), \qquad \hat{\mathbf{a}} = \texttt{fft2}(\texttt{fftshift}(\mathbf{a})),$$

where `ifft2` is the inverse discrete Fourier transform and `fftshift` swaps the first and third and the second and fourth quadrants of the array $\mathbf{a}$. Then $\mathbf{A}$ is block Toeplitz with Toeplitz blocks (BTTB) [33]. For efficient computations, $\mathbf{A}$ is embedded in a $128^2 \times 128^2$ block circulant with circulant block (BCCB) matrix, which can be diagonalized by the two-dimensional discrete Fourier and inverse discrete Fourier transform matrices [33]. Data $\mathbf{z}$ with a signal-to-noise ratio of approximately 35 is then generated using the statistical model (4.3) with $\sigma^2 = 25$ and $\gamma = 10$ – physically realistic values for these parameters. To generate Poisson noise, the `poissrnd` function in MATLAB's Statistics Toolbox is used. The corresponding blurred, noisy data $\mathbf{z}$ is given on the right hand side in Figure 4.2.

Let present now the reconstructed images obtained using standard Tikhonov regularization, total variation regularization, and regularization by the ℓ^2-norm of the gradient.

4.5.1 Tikhonov Regularization

For Tikhonov regularization, GPRNCG is applied to problem (4.1). The regularization parameter $\alpha = 4\times 10^{-6}$ was chosen so that the solution error $\|\mathbf{u}_\alpha - \mathbf{u}_{\text{exact}}\|_2$ is minimized. Noting that this will not necessarily yield the "optimal" regularization parameter since slightly larger values of α, though resulting in a larger solution error, may also yield reconstructions that are more physically correct (in this case smoother). However, the objective in this dissertation is only to show that this method works in practice. The question of optimal regularization parameter choice is left for a later work. The choice of GPRNCG parameters included is $\text{GP}_{\text{max}} = 1$, since more gradient projection iterations did not appreciably improve the convergence properties of GPRNCG, $\text{CG}_{\text{max}} = 50$, and $\gamma_{\text{CG}} = 0.25$. GPRNCG iterations stops after a 10 orders of magnitude decrease in the norm of $\nabla_{\text{proj}} T_\alpha(\mathbf{u}_k)$. In this example, this stopping criteria was satisfied after only 12 GPRNCG iterations. The reconstruction is given on the bottom in Figure 4.4.

One of the benefits of using the Poisson likelihood in place of least squares is that it is sensitive to changes in the low intensity regions of images. This is illustrated by Figure 4.4 where a cross section of $\mathbf{u}_{\text{exact}}$, $\mathbf{z}$ and $\mathbf{u}_\alpha$ are plotted corresponding to the 32nd row of the respective arrays. The low intensity feature, which can also be seen in the two dimensional images in Figure 4.2, is reconstructed with reasonable accuracy using the Poisson likelihood. The high frequency artifacts in the high intensity regions, however, are not desirable. This observation coincides with those made by others (see e.g., [32]); namely, that the Poisson likelihood is sometimes less effective than least squares in regions of an image that are high intensity and very smooth. For general interest, Richardson-Lucy iteration is also applied, stopping iterations once $\|\mathbf{u}_k - \mathbf{u}_{\text{exact}}\|_2$ was minimized. Interestingly, the resulting reconstruction was visually indistinguishable (cross sections included) from that obtained using GPRNCG with $\alpha = 4 \times 10^{-6}$. This supports the observation that both methods can be characterized as regularized Poisson likelihood estimation schemes and hence should yield similar results. Also, the energy of the GPRNCG and Richardson-Lucy reconstructions is the same. Finally, the plot of the reconstruction obtained using GPRNCG with $\alpha = 4\times 10^{-5}$ is also presented. Note that though the solution error is larger for this choice of α, the reconstruction is smoother. In [10], the object studied is a synthetically generated star-field – for which the Poisson likelihood is particularly well-suited – and the data is generated using statistical model (4.3), but with a significantly lower signal-to-noise of approximately 4.5. The method presented here also works well on this example.

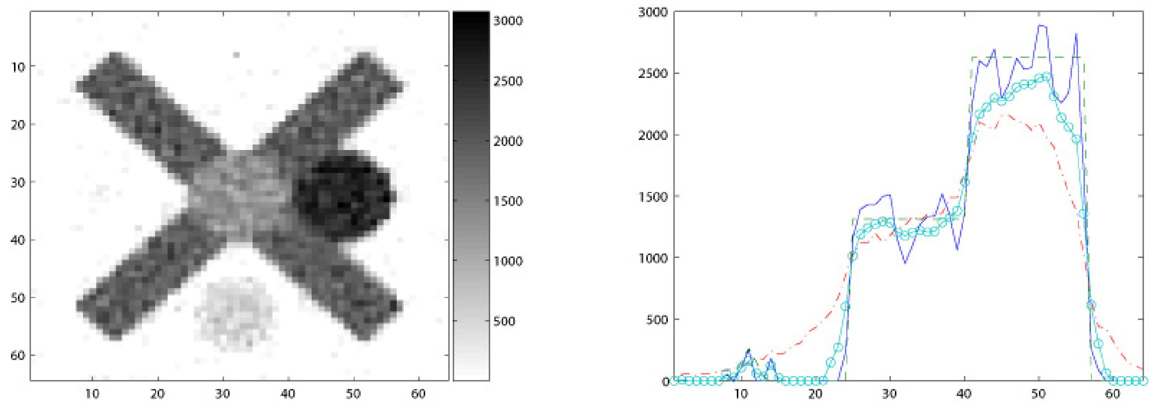

Figure 4.3: *On the left is the reconstruction obtained by GPRNCG applied to (4.1) with* $J(\mathbf{u}) = \frac{1}{2}\|\mathbf{u}\|_2^2$ *and* $\alpha = 4 \times 10^{-6}$. *On the right is are plots of the 32nd row of the* 64×64 *arrays* $\mathbf{u}_{\text{exact}}$, $\mathbf{z}$ *and* $\mathbf{u}_\alpha$. *The dashed line corresponds to the true object; the dash-dotted line corresponds to the blurred, noisy data* $\mathbf{z}$*; the solid line corresponds to* $\mathbf{u}_\alpha$ *with* $\alpha = 4 \times 10^{-6}$*; and the line with circles to* $\mathbf{u}_\alpha$ *with* $\alpha = 4 \times 10^{-5}$.

4.5.2 ℓ^2-Norm of the Gradient Regularization

For ℓ^2-norm of the gradient regularization, let apply GPRNCG to (4.1) with regularization parameter $\alpha = 10^{-6}$, which was chosen so that the solution error $\|\mathbf{u}_\alpha - \mathbf{u}_{\text{exact}}\|_2$ was near to minimal ($\alpha = 2 \times 10^{-7}$ minimized the solution error) but which yielded a reconstruction that was noticeable effected by the smoothing properties of the regularization function. This method of choosing α is admittedly ad hoc. However, the objective here is only to show that the method works in practice and that reconstructions are indeed smooth. The optimization parameters for GPRNCG iterations are the one used in the Tikhonov case. However, the iterations stop after a 9 orders of magnitude decrease in the norm of projected gradient of T_α. The reconstruction is given on the left in Figure 4.3. To demonstrate the effect of the Laplacian regularization, on the right in Figure 4.3, the plot of the 32nd row of of the true image and the reconstructions with $\alpha = 1 \times 10^{-6}$ and $\alpha = 2 \times 10^{-7}$, which minimizes the solution error, is provided.

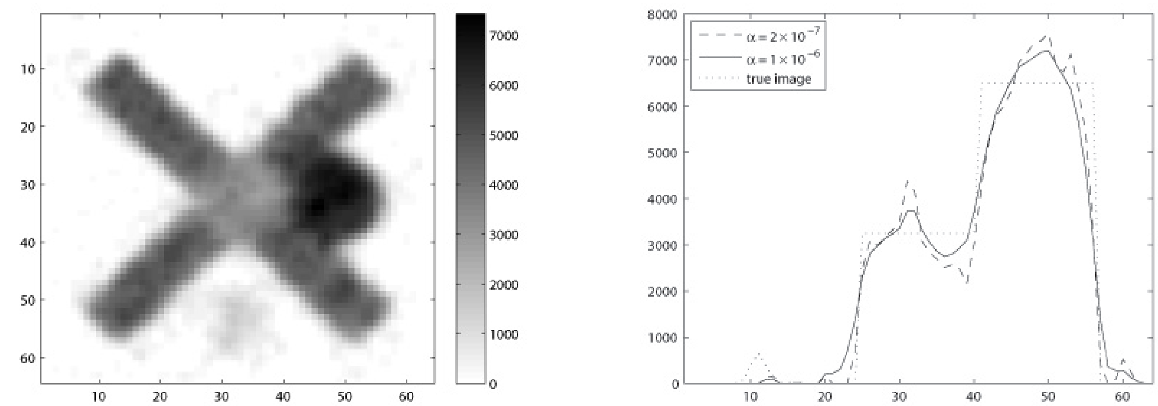

Figure 4.4: *On the left is the reconstruction obtained by GPRNCG applied to (4.1) with* $J(\mathbf{u}) = \frac{1}{2}\|\mathbf{Du}\|_2^2$ *and* $\alpha = 1 \times 10^{-6}$. *On the right is the 32nd row of the true image and the reconstructions with* $\alpha = 1 \times 10^{-6}$ *and* $\alpha = 2 \times 10^{-7}$, *which minimizes the solution error.*

4.5.3 Total Variation Regularization

In this case, the data $\mathbf{z}$ was generated using (4.3) with a signal-to-noise ratio of approximately 30. With the blurred, noisy data in hand, the object is estimated by solving (4.1) using GPRLD with $\text{GP}_{\max} = 1$ (note that then a value for γ_{GP} is not needed), $\gamma_{\text{CG}} = 0.25$ with preconditioning and 0.1 without, and $\text{CG}_{\max} = 40$, which is only ever satisfied if preconditioning is *not* used. Iterations stop once

$$\|\nabla_{\text{proj}} T_\alpha(\mathbf{u}_k)\|_2 / \|\nabla_{\text{proj}} T_\alpha(\mathbf{u}_0)\|_2 < \text{GradTol}, \tag{4.42}$$

where $\text{GradTol} = 10^{-5}$. These parameter values are chosen in order to balance computational efficiency with good convergence properties of the method. The initial guess was $\mathbf{u}_0 = \mathbf{1}$, and the regularization parameter was taken to be $\alpha = 5 \times 10^{-5}$. This choice of parameter approximately minimizes $\|\mathbf{u}_\alpha - \mathbf{u}_{\text{exact}}\|_2$. The reconstruction is given in Figure 4.5. In order to compare this result with those obtained using other methods, a comparison of plots of cross

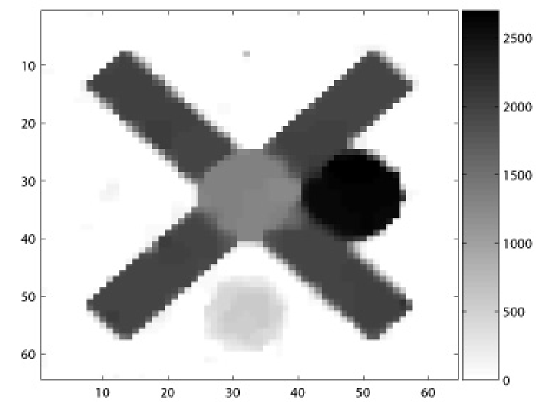

Figure 4.5: *On the left is the reconstruction obtained by GPRLD applied to (4.1) with* $J(\mathbf{u}) = \mathrm{TV}(\mathbf{u})$ *and* $\alpha = 5 \times 10^{-5}$.

sections of different reconstructions corresponding to the 32nd row and 32nd column; note that the sub-objects are all centered on one of these two cross-sections. In our first comparison, the picture of plot reconstructions obtained using the approach presented in this thesis, RL, and the projected Newton method applied to the problem of minimizing the Tikhonov regularized Poisson likelihood function over $\{\mathbf{u} \mid \mathbf{u} \geq \mathbf{0}\}$. For the latter method, $\mathrm{CG}_{\max} = 50$ is used, $\mathrm{GradTol} = 10^{-8}$, and initial guess $\mathbf{u}_0 = \mathbf{1}$. The regularization parameter - chosen as above - was $\alpha = 2 \times 10^{-6}$. The RL reconstruction was taken to be the iterate that minimizes $\|\mathbf{u}_k - \mathbf{u}_{\text{exact}}\|$. The results are given in Figure 4.6. The total variation reconstruction is visually superior to the others, with the exception of the Gaussian with the high peak in the left-hand plot. This is not surprising given the fact that it has been observed that standard Poisson estimation is particularly effective at reconstructing objects with high intensity, but small support, such as a star.

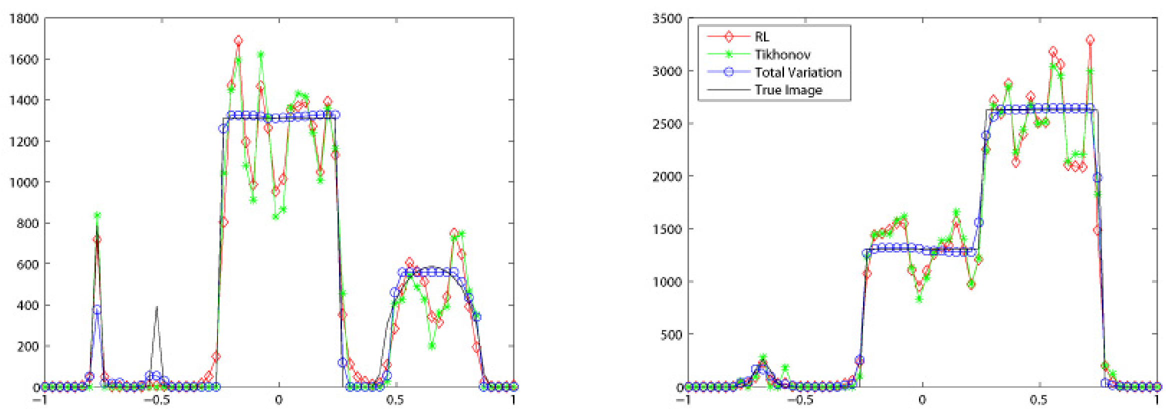

Figure 4.6: *Comparison of Cross-Sections of Reconstructions Obtained with a Variety of Poisson Likelihood Methods. The left-hand plot corresponds to column 32 and the right-hand plot corresponds to row 32 of the respective reconstructions. The true image is the solid line; the projected lagged-diffusivity reconstruction is the line with circles; the Tikhonov solution is the line with stars; and the Richardson-Lucy solution is the line with diamonds.*

Chapter 5

Conclusion

The main results of this thesis are the proofs that $R_\alpha(A; z+\gamma)$ defines a regularization scheme, where R_α is defined by (2.9), (1.12) for four different regularization functionals J : $\frac{1}{2}||u||_2^2$, $\frac{1}{2}||G(x)\nabla u||_2^2$, $\frac{1}{2}||\tau u||_2^2$ and $\int_\Omega \sqrt{|\nabla u|^2 + \beta}dx$.

Following the theoretical analysis, is the computational method of [10]. This method is efficient and very effective for nonnegatively constrained Poisson likelihood estimation problems. It is also proved in this thesis that under reasonable circumstances, its iterations converge to the unique minimizer of (4.1). In addition the demonstration the effectiveness of the approach in general, and of the computational method in particular, on simulated astronomical data generated using statistical model (1.2) is given.

The plan for the immediate future is to investigate a parameter selection method for α in the four methods above.

Bibliography

[1] *Wikipedia, entry on Charge-Couple-Device.*

[2] R. Acar and C.R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems **10** (1994), 1217–1229.

[3] T. Asaki, R. Chartrand, and T. Le, *A Variational Approach to Reconstructing Images Corrupted by Poisson Noise*, UCLA CAM Report (2005November), 05–49.

[4] John B. Conway, *A Course in Functional Analysis*, Second, Springer, 1990.

[5] Gerald B. Folland, *Introduction To Partial Differential Equations*, Princeton University Press, Princeton, 1995.

[6] J. M. Bardsley, *An Efficient Computational Method for Total Variation-Penalized Poisson Likelihood Estimation*, accepted in Inverse Problems and Imaging.

[7] J. M. Bardsley and N'djekornom Laobeul, *An analysis of regularization by diffusion for ill-posed poisson-likelihood estimation*, accepted in Inverse Problems in Science and Engineering.

[8] J. M. Bardsley and N'djekornom Laobeul, *Tikhonov Regularized Poisson Likelihood Estimation: Theoritical Justification and a Computational Method*, Inverse Problems in Science and Engineering.

[9] J. M. Bardsley and A. Luttman, *Total Variation-Penalized Poisson Likelihood Estimation for Ill-Posed Problems*, accepted in Advances in Computational Mathematics, Special Issue on Mathematical Imaging.

[10] J. M. Bardsley and C. R. Vogel, *A Nonnnegatively Constrained Convex Programming Method for Image Reconstruction*, SIAM Journal on Scientific Computing **25(4)** (2004), 1326–1343.

[11] Haïm Brezis, *Analyse Fonctionnelle Théorie et applications*, DUNOD, 2006.

[12] A. Chambolle and P.-L. Lions, *Image recovery via total variation minimization and related problems*, Numerische Mathematik **76, Number 2** (1997April), 167–188.

[13] Richard Cottle, *On the Convexity of Quadratic Forms Over Convex Sets*, Operations Research **15 No. 1** (1967), 170–172.

[14] Angus E. Taylor, *Introduction to Funtional Analysis*, Wiley Sons Inc, New York, 1961.

[15] Torbjørn Eltoft and Taesu Kim, *On the Multivariate Laplace Distribution*, IEEE Signal Processing Letters **13 (5)** (2006May), 300–303.

[16] Lawrence Evans, *Partial Differential Equations*, American Mathematical Society, 1998.

[17] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, 1992.

[18] W. Feller, *An Introduction to Probability Theory and Its Applications*, Wiley, New York, 1971.

[19] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Monographs in Mathematics **80** (1984).

[20] J. W. Goodman, *Introduction to Fourier Optics*, 2nd ed., McGraw-Hill, 1996.

[21] M. Green, *Statistics of images, the TV algorithm of Rudin-Osher-Fatemi for image denoising, and an improved denoising algorithm*, CAM Report (2002October), 02–55.

[22] Christian Hansen, James G. Nagy, Dianne P., and O$'$Leary, *Deblurring Images. Matrices, Spectra, and Filtering*, SIAM, 2006.

[23] Jinggang Huang and David Mumford, *Statistics of Natural Images and Models*, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (1999), 541–547.

[24] J. P. Kaipio, V. Kolehmainen, M. Vauhkonen, and E. Somersalo, *Inverse Problem with Structural Prior Information*, Inverse Problems **15** (1999), 713–729.

[25] C. T. Kelley, *Iterative Methods for Optimization*, Philadelphia, 1999.

[26] S. Osher L. I. Rudin and E. Fatemi, *Nonlinear Total Variation Based Noise Removal Algorithms*, Physica D **60(1-4)** (1992), 259 –268.

[27] J. J. Moré and G. Toraldo, *On the Solution of Large Quadratic Programming Problems with Bound Constraints*, SIAM Journal on Optimization **1** (1991), 93–113.

[28] J. Nocedal and S. Wright, *Numerical Optimization*, Springer, 1999.

[29] R. T. Rockafellar, *Convex Analysis* (1970).

[30] D. L. Snyder, C. W. Helstrom A. D. Lanterman, M. Faisal, and R. L. White, *Compensation for readout noise in CCD images*, Journal of the Optical Society of America A **12** (1995), 272–283.

[31] D. L. Snyder, A. M. Hammoud, and R. L. White, *Image recovery from data acquired with a charge-coupled-device camera*, Journal of the Optical Society of America A **10** (1993), 1014–1023.

[32] R. Vio, J. Bardsley, and W. Wamsteker, *Least-Squares methods with Poissonian noise: an analysis and a comparison with the Richardson-Lucy algorithm*, Astronomy and Astrophysics **436** (2005), 741–755.

[33] C. R. Vogel, *Computational Methods for Inverse Problems*, Philadelphia, 2002.

[34] Arch W. Naylor and George R. Sell, *Linear Operator Theory in Engineering and Science*, Springer-Verlag, 1982.

[35] E. Zeidler, *Applied Functional Analysis*, Springer-Verlag, New York.