2010

REGULARIZATION PARAMETER SELECTION METHODS FOR ILL POSED POISSON IMAGING PROBLEMS

John Goldes
The University of Montana

REGULARIZATION PARAMETER SELECTION METHODS FOR ILL-POSED POISSON IMAGING PROBLEMS

by

John Abraham Goldes

B.A. University of Montana, 2005

M.A. University Montana, 2007

presented in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy
in Mathematics, Applied Mathematics

The University of Montana
Missoula, MT

May 2010

Approved by:

Perry Brown, Associate Provost for Graduate Education
Graduate School

Johnathan Bardsley, Chair
Mathematics

Leonid Kalachev
Mathematics

Emily Stone
Mathematics

Jennifer Halfpap
Mathematics

Jesse Johnson
Computer Science

Goldes, John A. Ph.D., May 2010 Mathematics

Regularization Parameter Selection Methods for Ill-Posed Poisson Imaging Problems

Committee Chair: John Bardsley, Ph.D.

A common problem in imaging science is to estimate some underlying true image given noisy measurements of image intensity. When image intensity is measured by the counting of incident photons emitted by the object of interest, the data-noise is accurately modeled by a Poisson distribution, which motivates the use of Poisson maximum likelihood estimation. When the underlying model equation is ill-posed, regularization must be employed. I will present a computational framework for solving such problems, including statistically motivated methods for choosing the regularization parameter. Numerical examples will be included.

Contents

List of Figures

Chapter 1

Introduction

An inverse problem is a problem in which some unknown quantity must be estimated using measurements indirectly related to that quantity. A broad array of research falls under the umbrella of solving inverse problems. Inverse problems arise in geophysics, remote sensing, imaging science, and other fields. In groundwater modeling, for example, one estimates material parameters of an aquifer from measurements of pressure of the fluid that immerses the aquifer. Many inverse problems cannot be solved analytically and so computational methods play an important role in obtaining an estimate. It is often the case that a small amount of noise in the data leads to large errors in the estimates. Such problems are referred to as ill-posed. In order to deal with ill-posedness, techniques known as regularization methods have been developed.

1.1 Astronomical Imaging Example

A common problem in astronomical imaging is to estimate the image of an object in outer-space using pictures of that object recorded using a ground-based telescope. Turbulence in the atmosphere causes distortions in the planar wave-front resulting in a blurred image being recorded. The blurred image

can be described by the equation

$$\mathbf{z} = \mathbf{A}\mathbf{u}_e, \tag{1.1}$$

where $\mathbf{z} \in \mathbb{R}^N$ is a vector obtained by lexicographical ordering [34] of the $\sqrt{N} \times \sqrt{N}$ blurred image array, $\mathbf{A}$ is the forward model matrix, and $\mathbf{u}_e$ is a vector obtained by lexicographical ordering of the $\sqrt{N} \times \sqrt{N}$ underlying true image array. Equation (1.1) is a discretization of a continuous model often given by convolution:

$$z(x,y) = \int\int_{\Omega} a(x-x',y-y')u_e(x',y')dx'dy', \tag{1.2}$$

where a is the point-spread function, and $\Omega \in \mathbb{R}^2$ is the computational domain. The inverse problem is to estimate $\mathbf{u}_e$ given measurements of $\mathbf{z}$. The process of recording $\mathbf{z}$ introduces noise into the measurements and this needs to be taken into account when estimating $\mathbf{u}_e$. The blurred image $\mathbf{z}$ is recorded using a charged-coupled device (CCD) camera. A CCD camera consists of an array of pixels onto which photons fall and are counted. It is known that errors in counting processes are well-modeled by a Poisson distribution. A statistical model for the observations $\mathbf{z}$ is

$$\mathbf{z} = \text{Poiss}(\mathbf{A}\mathbf{u}_e + \gamma) + N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{1.3}$$

where $\text{Poiss}(\lambda)$ indicates a Poisson random vector with mean λ, and $N(\mathbf{0}, \sigma^2\mathbf{I})$ indicates a normally distributed random vector with mean $\mathbf{0}$ and covariance $\sigma^2\mathbf{I}$, where $\mathbf{I}$ denotes the identity matrix. γ models the background intensity, and the Gausssian term is a result of background noise in the recording electronics. See [33] for a more detailed description of this model.

1.2 Positron Emission Tomography Example

Positron emission tomography (PET) is a technique that is used to track the uptake of certain metabolites in an organism. In a typical case, a solution of glucose that has been tagged with a radioactive isotope is injected into a patient. When the isotope decays, a photon is emitted, which anihilates with

an electron, causing a pair of photons to propagate in opposite directions. If the two photons reach the detectors at either end of the connecting line within a short enough period of time, an event is recorded along that line, known as a line of response (LOR). The mean intensity along the ith LOR is modeled [32] by

$$I_i = e^{-\int_{L_i} \mu(s)ds} \int_{L_i} u_e(s)ds + \gamma_i, \tag{1.4}$$

where u_e is the target emission density function, γ_i is the expected number of erroneous counts along the ith LOR, and μ is the attenuation function. Note that the collection of all line integrals $\int_{L_i} u_e(s)ds$ defines the Radon transform of u_e. The exponential term represents the probability that an event along the ith LOR is recorded. In practice, PET data is discrete and the discretized version of equation (1.4) is given by

$$\mathbf{I}_i = [\mathbf{A}^{\text{emiss}}\mathbf{u}_e]_i + \gamma_i,$$

where $\mathbf{A}^{\text{emiss}} = \mathbf{G}\mathbf{A}^{\text{Radon}}$, with $\mathbf{G}$ denoting a diagonal matrix with $[\mathbf{G}]_{ii} = e^{-\int_{L_i} \mu(s)ds}$, and $\mathbf{A}^{\text{Radon}}$ denoting the discrete Radon transform matrix. $[\mathbf{A}^{\text{Radon}}]_{i,j}$ is the length of the intersection of the ith LOR with the jth computational grid element. Note that this model does not take into account detector efficiency or detector dead time. For a more detailed discussion of the model of PET data see [29].

Since PET data consists of photon counts, the noise in the data is well-modeled by a Poisson distribution. Hence the statistical model for PET data is given by

$$\mathbf{z} = \text{Poiss}(\mathbf{A}^{\text{emiss}}\mathbf{u}_e + \gamma). \tag{1.5}$$

1.3 Ill-Posed Poisson Likelihood Estimation

In the astronomical imaging and PET imaging examples the noise in the data follows a Poisson distribution. In many situations, this distribution is approximated by a Gaussian distribution. This is advantageous because methods for solving the resulting least-squares estimation problem have been studied extensively. However an improvement in the estimate of the target image can be obtained by

using the correct noise model [34]. In this situation the resulting estimation problem entails minimizing the negative logarithm of the Poisson likelihood function. This problem can be stated as:

$$\mathbf{u}_{\mathrm{ML}} = \operatorname{argmin}_{\mathbf{u}\geq 0}\{T_0(\mathbf{u};\mathbf{z}) \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} [\mathbf{A}\mathbf{u}]_i + \gamma_i - z_i \ln([\mathbf{A}\mathbf{u}]_i + \gamma_i)\}. \tag{1.6}$$

This convex minimization problem is nonnegatively constrained and has no closed-form solution. Iterative methods exist for solving such problems; for example the Richardson-Lucy [34], and the algorithm we will present in Chapter 2.

Since the estimation problem in both examples is ill-posed, regularization is required. If an iterative method is used to solve (1.6), then a regularization method can be formulated as a stopping rule [7]. Alternatively, a penalty term can be added to the minimization problem. The estimate of the target image is then found by solving

$$\mathbf{u}_\alpha = \operatorname{argmin}_{\mathbf{u}\geq \mathbf{0}}\{T_\alpha(\mathbf{u}) \stackrel{\mathrm{def}}{=} T_0(\mathbf{u};\mathbf{z}) + \alpha J(\mathbf{u})\}, \tag{1.7}$$

where α is the regularization parameter and $J(\mathbf{u})$ is the regularization function. A theoretical justification for employing this approach is given in [5, 11, 12]. In the PET imaging example, the Tikhonov regularized problem is known as the penalized maximum likelihood problem and has been extensively studied [3, 14–19, 22, 24, 27, 31, 36].

In Chapter 2, an algorithm is described for solving the convex, nonnegatively constrained minimization problem and convergence of the algorithm is proven for various functions J, given that the resulting function T_α satisfies certain properties.

In Chapter 3 the addition of a penalty term is motivated using a Bayesian framework. From that perspective, it is apparent that the form of $J(\mathbf{u})$ should reflect whatever prior knowledge is available about the unknown image. For example, if the unknown image is known to be smooth, then taking $J(\mathbf{u}) = \mathbf{u}^T\mathbf{L}\mathbf{u}$, where $\mathbf{L}$ is a discretization of the negative Laplacian operator, is an appropriate choice

[11]. In this case the regularization term is quadratic and $\mathbf{L}$ is referred to as the regularization matrix. If on the other hand edges are known to exist in the target image, regularization functions which allow for edge-preservation are of interest. Taking $J(\mathbf{u})$ to be the total variation function is one option for an edge-preserving penalty term [5, 31]. In Chapter 5, the construction of a quadratic regularization term that allows for edge-preservation is described, while [13] describes a more statistically rigorous procedure for constructing an edge-preserving quadratic regularization term. The advantage of a quadratic penalty term is that computing $\mathbf{u}_\alpha$ is more efficient than when total variation regularization is used.

In Chapter 3, it is also shown that for quadratic regularization terms in which the regularization matrix satisfies certain conditions and for the discrete approximation of the total variation function that is given in [6], the conditions necessary for the convergence of the nonnegatively constrained convex minimization algorithm presented in Chapter 2 are met.

The regularization parameter controls the contribution of the penalty term to the solution. Hence some method of selecting a value of α that yields a quality reconstruction is desired. In the case of least squares estimation, such methods are well-developed. However, those methods cannot be directly applied to problems in which the data noise model is Poisson. In Chapter 4, a Taylor series argument is used to approximate the negative-log of the Poisson likelihood with a weighted sum of squares term. This approximation is used to extend certain methods for selecting the value of α to the case of Poisson likelihood estimation [8].

The regularization parameter selection methods that will be considered are: generalized cross validation (GCV), unbiased predictive risk estimation (UPRE), and the discrepancy principle (DP). GCV selects the value of α that minimizes the GCV function, which is an approximation of leave-one-out cross validation function for large-scale problems [34, 35]. DP makes an approximation from which it follows that an appropriate value for α is that which yields a solution for which the sum of squares of the weighted residuals is equal to the mean of a χ^2 distribution [26, 34]. The UPRE method selects the value of α that minimizes an unbiased estimator of the predictive risk [34]. In Chapter 4 these methods are introduced in detail.

In Chapter 5, the methods and general framework are tested on examples in both the contexts of the astronomical and PET imaging. Multiple synthetic data sets are used and the results indicate that our framework yields good reconstructions.

We end with conclusions in Chapter 6

The work in this thesis is based on the papers [4, 8–10].

Chapter 2

The Optimization Algorithm

Material in this chapter is based on the work in [4].

This chapter contains a method for solving non-negatively constrained convex optimization problems as well as a proof of convergence of the method. Both the method and the proof of convergence are the subject of [4]. The optimization problem of interest has the form:

$$\min_{\mathbf{u}\in\Omega} T(\mathbf{u}), \tag{2.1}$$

where $\Omega = \{\mathbf{u} \in \mathbb{R}^n \mid u_i \geq 0, i = 1,\ldots,n\}$ and $T : \Omega \to \mathbb{R}$ is a function for which the following assumptions hold:

Assumption 1: T is coercive, twice continuously differentiable, and $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u} \in \Omega$;

Assumption 2: The gradient of T is Lipschitz continuous with Lipschitz constant L.

Here ∇T and $\nabla^2 T$ denote the gradient and Hessian, respectively, of T. Note that in [4], T was assumed to be strictly convex, while here the stronger assumption that $\nabla^2 T$ is positive definite over Ω is made.

Lemma 2.1.2 below establishes that if $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u} \in \Omega$ then T is strictly convex over Ω. T is coercive on Ω if the following is holds:

$$\|\mathbf{u}\|_2 \to \infty \quad \text{implies} \quad T(\mathbf{u}) \to \infty \quad \text{for } \mathbf{u} \in \Omega. \tag{2.2}$$

T is strictly convex on Ω if it has the property that for $\mathbf{u}_1, \mathbf{u}_2 \in \Omega$ and $\tau \in [0,1]$, the following inequality holds:

$$T(\tau\mathbf{u}_1 + (1-\tau)\mathbf{u}_2) < \tau T(\mathbf{u}_1) + (1-\tau)T(\mathbf{u}_2). \tag{2.3}$$

2.1 A Gradient Projection-Reduced Newton (GPRN) method

Here I present an iterative method for solving problems of form (2.1). This method has a nested iterative design in which one outer iteration consists of two stages. The first stage entails performing gradient projection iterations in order to identify the active set. The second stage uses conjugate gradient (CG) iterations to compute a Newton step on the free (inactive) variables.

2.1.1 Preliminaries

The projection of $\mathbf{u} \in \mathbb{R}^n$ onto Ω is given by

$$\mathscr{P}(\mathbf{u}) := \text{argmin}_{\mathbf{v}\in\Omega}\|\mathbf{u}-\mathbf{v}\| = \max\{\mathbf{u}, \mathbf{0}\}, \tag{2.4}$$

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

$$\mathscr{A}(\mathbf{u}) = \{i \mid u_i = 0\}, \tag{2.5}$$

and the inactive set $\mathscr{I}(\mathbf{u})$ is defined to be the complementary set of indices.

The reduced gradient of T at $\mathbf{u} \in \Omega$ is defined to be

$$\nabla_{\text{red}} T(\mathbf{u}) = \begin{cases} [\nabla T(\mathbf{u})]_i, & i \in \mathscr{I}(\mathbf{u}), \\ 0, & i \in \mathscr{A}(\mathbf{u}), \end{cases} \tag{2.6}$$

and the projected gradient of T at $\mathbf{u} \in \Omega$ is given by

$$\nabla_{\text{proj}} T(\mathbf{u}) = \begin{cases} [\nabla T(\mathbf{u})]_i, & i \in \mathscr{I}(\mathbf{u}) \text{ or } i \in \mathscr{A}(\mathbf{u}) \text{ and } \frac{\partial T(\mathbf{u})}{\partial u_i} < 0, \\ 0, & \text{otherwise.} \end{cases} \tag{2.7}$$

The reduced Hessian of T at $\mathbf{u} \in \Omega$ is given by

$$\nabla^2_{\text{red}} T(\mathbf{u}) = \begin{cases} [\nabla^2 T(\mathbf{u})]_{i,j}, & i \in \mathscr{I}(\mathbf{u}) \text{ and } j \in \mathscr{I}(\mathbf{u}), \\ 0, & \text{otherwise.} \end{cases} \tag{2.8}$$

2.1.2 The Gradient Projection Iteration

The first stage of the GPRN algorithm is defined by the gradient projection iteration. The gradient projection iteration is defined as follows: given $\mathbf{u}_k \in \Omega$, $\mathbf{u}_{k+1}$ is found by the following computations:

$$\mathbf{p}_k = -\nabla T(\mathbf{u}_k), \tag{2.9}$$

$$\lambda_k = \text{argmin}_{\lambda > 0}\{T(\mathscr{P}(\mathbf{u}_k + \lambda \mathbf{p}_k))\}, \tag{2.10}$$

$$\mathbf{u}_{k+1} = \mathscr{P}(\mathbf{u}_k + \lambda_k \mathbf{p}_k). \tag{2.11}$$

In the implementation of the gradient projection iteration, an inexact solution to subproblem (2.10) is obtained by using a projected backtracking line search. This is accomplished by using a line search algorithm to generate a sequence $\{\lambda_k^j\}_{j=0}^m$, which is terminated once λ_k^j satisfies

$$T(\mathbf{u}_k(\lambda_k^j)) \leq T(\mathbf{u}_k) - \frac{\mu}{\lambda_k^j} \|\mathbf{u}_k - \mathbf{u}_k(\lambda_k^j)\|^2, \tag{2.12}$$

where $\mu \in (0,1)$ and

$$\mathbf{u}_k(\lambda) = \mathscr{P}(\mathbf{u}_k + \lambda \mathbf{p}_k). \tag{2.13}$$

Then in (2.10), $\lambda_k \stackrel{\text{def}}{=} \lambda_k^m$. In the line search algorithm that generates the sequence $\{\lambda_k^j\}_{j=0}^m$, the initial step length parameter is given by

$$\lambda_k^0 = \frac{\|\mathbf{p}_k\|^2}{\langle \nabla^2 T(\mathbf{u}_k)\mathbf{p}_k, \mathbf{p}_k\rangle}. \tag{2.14}$$

At the jth line search iteration, if (2.12) is not satisfied by λ_k^{j-1}, then λ_k^j is computed as follows. Compute the solution of

$$\hat{\lambda}_k^j = \text{argmin}_\lambda \left\{ T(\mathbf{u}_k) - \|\mathbf{p}_k\|^2\lambda + \frac{T(\mathbf{u}_k(\lambda_k^{j-1})) + \lambda_k^{j-1}\|\mathbf{p}_k\|^2 - T(\mathbf{u}_k)}{(\lambda_k^{j-1})^2}\lambda^2 \right\}, \tag{2.15}$$

and then set

$$\lambda_k^j = \text{median}\{\lambda_k^{j-1}/100, \hat{\lambda}_k, \lambda_k^{j-1}/2\}. \tag{2.16}$$

A criterion for terminating the gradient projection iterations needs to be specified. A similar algorithm of Moré and Toraldo [25] gives a useful stopping rule; the iterations should be stopped once

$$T(\mathbf{u}_k) - T(\mathbf{u}_{k+1}) \le \gamma_{\text{GP}} \max_{l<k}\{T(\mathbf{u}_l) - T(\mathbf{u}_{l+1})\}, \tag{2.17}$$

where $0 < \gamma_{\text{GP}} < 1$ is fixed. In [4], $\gamma_{\text{GP}} = .1$, but its optimal value is problem dependent.

2.1.3 The Reduced Newton step

Though the convergence of gradient projection iterations has been established [23], interspersing the gradient projection iterations with steps computed from the reduced Newton system

$$\nabla^2_{\text{red}} T(\mathbf{u})\mathbf{p} = -\nabla_{\text{red}} T(\mathbf{u}_k), \tag{2.18}$$

results in an improved rate of convergence [34]. For large-scale problems solving (2.18) directly is inefficient. Instead a sequence $\{\mathbf{p}_k^j\}$ that converges to an approximate solution $\mathbf{p}_k$ of (2.18) is obtained by applying conjugate gradient [34] iterations to minimizing

$$q_k(\mathbf{p}) = T(\mathbf{u}_k) + \langle \nabla_{\text{red}} T(\mathbf{u}_k), \mathbf{p} \rangle + \frac{1}{2} \langle \nabla^2_{\text{red}} T(\mathbf{u}_k)\mathbf{p}, \mathbf{p} \rangle. \tag{2.19}$$

The stopping rule for the conjugate-gradient iterations is analogous to (2.17):

$$q_k(\mathbf{p}_k^j) - q_k(\mathbf{p}_k^{j+1}) \leq \gamma_{\text{CG}} \max_{l<j} \{q_k(\mathbf{p}_k^l) - q_k(\mathbf{p}_k^{l+1})\}, \tag{2.20}$$

where $0 < \gamma_{\text{CG}} < 1$ is fixed. If m_{CG} is the smallest integer that satisfies (2.20), then $\mathbf{p}_k$ is taken to be $\mathbf{p}_k^{m_{\text{CG}}}$.

After obtaining an approximate solution of (2.18), a backtracking line search is again performed with the decrease condition:

$$T(\mathbf{u}_k(\lambda_k^m)) \leq T(\mathbf{u}_k). \tag{2.21}$$

The Gradient Projection-Reduced Newton Iteration

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

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

(1) Compute gradient projection iterations $\{\mathbf{u}_{k,j}\}_{j=0}^{j_k}$, with $\mathbf{u}_{k,0} \stackrel{\text{def}}{=} \mathbf{u}_k$,
until either (2.17) is satisfied or $\text{GP}_{\max}$ iterations have been computed.

Step 2: Given $\mathbf{u}_k \stackrel{\text{def}}{=} \mathbf{u}_{k,j_k}$.

(1) Do CG iterations to approximately minimize the quadratic (2.19)
until either (2.20) 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 (2.21), and return $\mathbf{u}_{k+1} = \mathbf{u}_k(\lambda_k^m)$.

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

2.1.4 Proof of Convergence

Here the GPRN algorithm applied to (2.1) with T satisfying Assumptions 1 and 2 is shown to converge. The proof requires that the line search parameters generated within stage 1 of the algorithm be bounded. I will start by showing that this is so. First though, I need to state and prove some preliminary lemmas.

This first lemma is used to establish that a sufficient condition for the strict convexity of T is that $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u} \in \Omega$.

Lemma 2.1.1. *Given that T is continuously differentiable, T is strictly convex over a convex set Ω if and only if*

$$T(\mathbf{v}) > T(\mathbf{u}) + \nabla T(\mathbf{u})^T(\mathbf{v}-\mathbf{u}) \tag{2.22}$$

for all $\mathbf{u}, \mathbf{v} \in \Omega$.

Proof. First, assume that for all $\mathbf{u}, \mathbf{v} \in \Omega$,

$$T(\mathbf{v}) > T(\mathbf{u}) + \nabla T(\mathbf{u})^T(\mathbf{v}-\mathbf{u}).$$

Fix $\mathbf{x}_1, \mathbf{x}_2 \in \Omega$ and $\nu \in (0,1)$. Setting $\mathbf{u} = \nu\mathbf{x}_1 + (1-\nu)\mathbf{x}_2$ and alternatively $\mathbf{v} = \mathbf{x}_1$ or $\mathbf{v} = \mathbf{x}_2$ gives

$$T(\mathbf{x}_1) > T(\mathbf{x}) + \nabla T(\mathbf{x})^T(\mathbf{x}_1 - \mathbf{x}), \tag{2.23}$$

$$T(\mathbf{x}_2) > T(\mathbf{x}) + \nabla T(\mathbf{x})^T(\mathbf{x}_2 - \mathbf{x}). \tag{2.24}$$

Multiplying (2.23) by ν, (2.24) by $1-\nu$, and adding yields

$$\nu T(\mathbf{x}_1) + (1-\nu)T(\mathbf{x}_2) > T(\mathbf{u}) + \nabla T(\mathbf{u})^T(\nu\mathbf{x}_1 + (1-\nu)\mathbf{x}_2 - \mathbf{u}). \tag{2.25}$$

Now substituting $\mathbf{u} = \nu\mathbf{x}_1 + (1-\nu)\mathbf{x}_2$ into (2.25) gives

$$\nu T(\mathbf{x}_1) + (1-\nu)T(\mathbf{x}_2) > T(\nu\mathbf{x}_1 + (1-\nu)\mathbf{x}_2).$$

Thus T is strictly convex.

Next assume that T is strictly convex. Then for any $\mathbf{u}, \mathbf{v} \in \Omega$ and $\nu \in (0,1)$,

$$T(\nu\mathbf{v} + (1-\nu)\mathbf{u}) < \nu T(\mathbf{v}) + (1-\nu)T(\mathbf{u}).$$

This implies that

$$\frac{T(\mathbf{u} + \nu(\mathbf{v}-\mathbf{u})) - T(\mathbf{u})}{\nu} < T(\mathbf{v}) - T(\mathbf{u}),$$

and taking the limit as $\nu \to 0$ yields

$$\nabla T(\mathbf{u})^T(\mathbf{v}-\mathbf{u}) \leq T(\mathbf{v}) - T(\mathbf{u}), \tag{2.26}$$

which implies

$$T(\mathbf{x} + \nu(\mathbf{y}-\mathbf{x})) \geq T(\mathbf{x}) + \nabla T(\mathbf{x})^T[\nu(\mathbf{y}-\mathbf{x})]$$

for $\mathbf{x}, \mathbf{y} \in \Omega$) and $\nu \in (0,1)$. Now suppose Suppose for some $\mathbf{x}, \mathbf{y} \in \Omega$,

$$\nabla T(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) = T(\mathbf{y}) - T(\mathbf{x}).$$

Then

$$\begin{aligned} T(\mathbf{x}) + \nu\nabla T(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) &= \nu T(\mathbf{y}) + (1-\nu)T(\mathbf{x}) \\ &> T(\mathbf{x} + \nu(\mathbf{y}-\mathbf{x})) \\ &\geq T(\mathbf{x}) + \nabla T(\mathbf{x})^T[\nu(\mathbf{y}-\mathbf{x})], \end{aligned}$$

which is a contradiction. Therefore

$$\nabla T(\mathbf{u})^T(\mathbf{v}-\mathbf{u}) < T(\mathbf{v}) - T(\mathbf{u})$$

for all $\mathbf{u}, \mathbf{v} \in \Omega$.

□

Lemma 2.1.2. *Given that T is twice continuously differentiable, T is strictly convex over Ω if $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u} \in \Omega$.*

Proof. Taylor's theorem states that for some ν, $0 < \nu < 1$,

$$T(\mathbf{v}) = T(\mathbf{u}) + \nabla T(\mathbf{u})^T(\mathbf{v}-\mathbf{u}) + \frac{1}{2}(\mathbf{v}-\mathbf{u})^T \nabla^2 T(\mathbf{u}+\nu(\mathbf{v}-\mathbf{u}))(\mathbf{v}-\mathbf{u}). \tag{2.27}$$

If $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u} \in \Omega$, then

$$T(\mathbf{v}) > T(\mathbf{u}) + \nabla T(\mathbf{u})^T(\mathbf{v}-\mathbf{u}),$$

and in light of Lemma 2.1.1, T is strictly convex. □

This lemma will be used in some proofs of some of the other preliminary lemmas as well as in the proof of convergence.

Lemma 2.1.3. *For all $\mathbf{u}, \mathbf{v} \in \Omega$ and $\lambda \geq 0$,*

$$(\mathbf{v}-\mathbf{u}(\lambda))^T(\mathbf{u}(\lambda) - \mathbf{u} + \lambda \nabla T(\mathbf{u})) \geq 0, \tag{2.28}$$

where $\mathbf{u}(\lambda) = \mathscr{P}(\mathbf{u} - \lambda \nabla T(\mathbf{u}))$.

Proof. Let $\mathbf{u} \in \Omega$ and define

$$\mathscr{A}(\mathbf{u}(\lambda)) = \{i = 1, \ldots, n \mid u(\lambda)_i = 0\}. \tag{2.29}$$

Let $\mathscr{A}(\mathbf{u}(\lambda))'$ denote the complement of $\mathscr{A}(\mathbf{u}(\lambda))$. Given $\mathbf{v} \in \Omega$, define $\mathscr{A}(\mathbf{v})$ in a similar fashion. Note that for $i \in \mathscr{A}(\mathbf{u}(\lambda))'$, $u(\lambda)_i - u_i + \lambda \nabla T(\mathbf{u})_i) = 0$. Also for $i \in \mathscr{A}(\mathbf{u}(\lambda)) \cap \mathscr{A}(\mathbf{v}))'$, $u_i - \lambda \nabla T(\mathbf{u})_i \leq 0$ and

$$\begin{aligned} (u(\lambda)_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 &= (u_i + \lambda \nabla T(\mathbf{u})_i)^2 \\ &\leq (v_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2. \end{aligned}$$

Therefore it is the case that

$$\begin{aligned} \|\mathbf{u}(\lambda) - \mathbf{u} + \lambda \nabla T(\mathbf{u})\|^2 &= \sum_{i=1}^{n} (u(\lambda)_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 \\ &= \sum_{i \in \mathscr{A}(\mathbf{u}(\lambda))'} (u(\lambda)_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 + \sum_{i \in \mathscr{A}(\mathbf{u}(\lambda)) \cap \mathscr{A}(\mathbf{v})'} (u(\lambda)_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 \\ &\quad + \sum_{i \in \mathscr{A}(\mathbf{u}(\lambda)) \cap \mathscr{A}(\mathbf{v})} (u(\lambda)_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 \\ &\leq \sum_{i \in \mathscr{A}(\mathbf{u}(\lambda))'} (v_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 + \sum_{i \in \mathscr{A}(\mathbf{u}(\lambda)) \cap \mathscr{A}(\mathbf{v})'} (v_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 \\ &\quad + \sum_{i \in \mathscr{A}(\mathbf{u}(\lambda)) \cap \mathscr{A}(\mathbf{v})} (u(\lambda)_i - u_i + \lambda \nabla T(\mathbf{u})_i)^2 \\ &= \|\mathbf{v} - \mathbf{u} + \lambda \nabla T(\mathbf{u})\|^2. \end{aligned} \tag{2.30}$$

Equation (2.30) implies that

$$\|\mathbf{u}(\lambda) - \mathbf{u} + \lambda \nabla T(\mathbf{u})\| \leq \|\mathbf{v} - \mathbf{u} + \lambda \nabla T(\mathbf{u})\|. \tag{2.31}$$

It follows that the function

$$\phi(t) = \|(1-t)\mathbf{u}(\lambda) + t\mathbf{v} - \mathbf{u} + \lambda \nabla T(\mathbf{u})\|^2/2 \tag{2.32}$$

achieves a local minimum at $t = 0$. Therefore,

$$0 \leq \phi'(0) = (\mathbf{v} - \mathbf{u}(\lambda))^T(\mathbf{u}(\lambda - \mathbf{u} + \lambda \nabla T(\mathbf{u})). \tag{2.33}$$

□

Note that (2.28) can be rewritten as

$$(\mathbf{u} - \mathbf{u}(\lambda))^T(\mathbf{v} - \mathbf{u}(\lambda)) \leq \lambda \nabla T(\mathbf{u})^T(\mathbf{v} - \mathbf{u}(\lambda)). \tag{2.34}$$

The result of setting $\mathbf{v} = \mathbf{u}$ in (2.34) is the following corollary:

Corollary 2.1.4. *For all* $\mathbf{u} \in \Omega$ *and* $\lambda \geq 0$,

$$\|\mathbf{u} - \mathbf{u}(\lambda)\|^2 \leq \lambda \nabla T(\mathbf{u})^T(\mathbf{u} - \mathbf{u}(\lambda)). \tag{2.35}$$

This next lemma is needed in the proof that the set of all line search parameters generated in step 1 of the GPRN algorithm are bounded.

Lemma 2.1.5. *Let* $\mathbf{u} \in \Omega$ *Then sufficient decrease condition (2.12) holds for all* λ *such that*

$$0 \leq \lambda \leq \frac{2(1-\mu)}{L}, \tag{2.36}$$

where L is the Lipschitz constant for ∇T.

Proof. Let $\mathbf{v} = \mathbf{u} - \mathbf{u}(\lambda)$. Then by the fundamental theorem of calculus,

$$T(\mathbf{u} - \mathbf{v}) - T(\mathbf{u}) = T(\mathbf{u}(\lambda)) - T(\mathbf{u}) = -\int_0^1 \nabla T(\mathbf{u} - t\mathbf{v})^T \mathbf{v} dt. \tag{2.37}$$

Adding and subtracting $\nabla T(\mathbf{u})^T \mathbf{v}$ yields

$$T(\mathbf{u}(\lambda)) = T(\mathbf{u}) + \nabla T(\mathbf{u})^T(\mathbf{u}(\lambda) - \mathbf{u}) - \int_0^1 (\nabla T(\mathbf{u} - t\mathbf{v}) - \nabla T(\mathbf{u}))^T \mathbf{v} dt,$$

which can be rewritten as

$$\lambda(T(\mathbf{u}) - T(\mathbf{u}(\lambda))) = \lambda \nabla T(\mathbf{u})^T(\mathbf{u} - \mathbf{u}(\lambda)) + \lambda \int_0^1 (\nabla T(\mathbf{u} - t\mathbf{v}) - \nabla T(\mathbf{u}))^T \mathbf{v} dt. \tag{2.38}$$

Note that

$$\begin{aligned} \left\| \int_0^1 (\nabla T(\mathbf{u} - t\mathbf{v}) - \nabla T(\mathbf{u}))^T \mathbf{v} dt \right\| &\leq \int_0^1 \|\nabla T(\mathbf{u} - t\mathbf{v}) - \nabla T(\mathbf{u})\| \|\mathbf{v}\| dt \\ &\leq \int_0^1 L\|\mathbf{v}\| t dt \\ &= L\|\mathbf{u} - \mathbf{u}(\lambda)\|^2/2, \end{aligned} \tag{2.39}$$

and so applying Corollary 2.1.4 to the inequality

$$\lambda(T(\mathbf{u}) - T(\mathbf{u}(\lambda))) \geq \lambda \nabla T(\mathbf{u})^T(\mathbf{u} - \mathbf{u}(\lambda)) - \lambda L\|\mathbf{u} - \mathbf{u}(\lambda)\|^2/2 \tag{2.40}$$

yields

$$\lambda(T(\mathbf{u}) - T(\mathbf{u}(\lambda))) \geq (1 - \lambda L/2)\|\mathbf{u} - \mathbf{u}(\lambda)\|^2, \tag{2.41}$$

which implies the desired result. □

The proof of convergence requires that the set of all line search parameters generated in step 1 of the GPRN algorithm be bounded, and that is the subject of this next result.

Lemma 2.1.6. *Let $\{\{\lambda_{k,j}\}_{j=0}^{j_k-1}\}_{k=0}^{\infty}$ be the set of line search parameters generated by the gradient projection iterations within step 1 of the GPRN algorithm. (Note that k denotes the outer iteration and*

j the inner gradient projection iterations). Then there exists constants β and M such that

$$0 < \beta < \lambda_{k,j} < M \quad for\ all\ j\ and\ k. \tag{2.42}$$

Proof. Let $\{\lambda_{k,j}^l\}_{l=0}^m$ be the set of line search parameters generated by the jth gradient projection iteration in step 1 of the kth outer iteration. Note that $\lambda_{k,j} = \lambda_{k,j}^m$ Recall that the initial step length parameter is taken to be

$$\lambda_{k,j}^0 = \frac{\|\mathbf{p}_{k,j}\|^2}{\langle \nabla^2 T(\mathbf{u}_{k,j})\mathbf{p}_{k,j}, \mathbf{p}_{k,j}\rangle}. \tag{2.43}$$

Note that since $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u} \in \Omega$, we have the inequalities

$$0 < \sigma_{\min,k,j}\|\mathbf{p}_{k,j}\|^2 \leq \langle \nabla^2 T(\mathbf{u}_{k,j})\mathbf{p}_{k,j}, \mathbf{p}_{k,j}\rangle \leq \sigma_{\max,k,j}\|\mathbf{p}_{k,j}\|^2, \tag{2.44}$$

where $\sigma_{\min,k,j}$ and $\sigma_{\max,k,j}$ are the minimum and maximum eigenvalues, respectively, of $\nabla^2 T(\mathbf{u}_{k,j})$. Equation (2.44) implies that

$$\sigma_{\max,k,j}^{-1} \leq \lambda_{k,j}^0 \leq \sigma_{\min,k,h}^{-1}. \tag{2.45}$$

Note that if (2.12) is satisfied for $\lambda_{k,j}^0$, then

$$\sigma_{\max,k,j}^{-1} \leq \lambda_{k,j} \leq \sigma_{\min,k,h}^{-1}, \tag{2.46}$$

and it is the case that $\lambda_{k,j} \geq \min\left\{\sigma_{\max,k,j}^{-1}, (1-\mu)/(50L)\right\}$. If $\lambda_{k,j}^0$ does not satisfy (2.12) then since $\lambda_{k,j}^0 > 0$, it must be the case that $\lambda_{k,j}^0 > \frac{2(1-\mu)}{L}$, or else Lemma 2.1.5 would indicate that (2.12) would be satisfied. Recall that in the line search algorithm, (2.16) specifies that $\lambda_{k,j}^l$ be chosen to be the median of $\{\lambda_{k,j}^{l-1}/100, \hat{\lambda}_{k,j}^{l-1}, \lambda_{k,j}^{l-1}/2\}$, where $\hat{\lambda}_{k,j}^{l-1}$ is given by (2.15). Therefore $\lambda_{k,j}^{l-1} > \lambda_{k,j}^l$ and $\lambda_{k,j}^l \leq \frac{\lambda_{k,j}^0}{2^l}$ for $l = 1, \ldots, m$, from which it follows that $\lambda_{k,j}^l \to 0$ as $l \to \infty$. This implies that there exists there exists some m for which $\lambda_{k,j}^m \leq 2(1-\mu)/L$ and so the line search algorithm will produce a value that satisfies (2.12) in a finite number of steps.

Suppose that the line search algorithm terminates at the mth iteration (meaning that $\lambda_{k,j}^m$ satisfies (2.12)

and $\lambda_{k,j}^{m-1}$ does not). Then Lemma 2.1.5 indicates that $\lambda_{k,j}^{m-1} > \frac{2(1-\mu)}{L}$ and so

$$\lambda_{k,j}^{m} \geq \frac{\lambda_{k,j}^{m-1}}{100} > \frac{1-\mu}{50L}. \tag{2.47}$$

Inequality (2.47) combined with the fact that the line search algorithm always produces a decreasing sequence yields the inequalities

$$\min\left\{\sigma_{\max,k,j}^{-1}, (1-\mu)/(50L)\right\} \leq \lambda_{k,j} \leq \sigma_{\min,k,j}^{-1}. \tag{2.48}$$

Now it needs to be shown that the set $\{\sigma_{\max,k,j}\}$ is bounded from above and the set $\{\sigma_{\min,k,j}\}$ has a lower bound that is greater than 0. First note that each gradient projection iteration yields a decrease in the value of T. The value of T is also decreased in step 2 of the outer iteration. Hence the coercivity T implies that the set $\{\mathbf{u}_{k,j}\}$ is bounded. Now assume that the set $\{\sigma_{\max,k,j}\}$ does not have an upper bound. Then there exists a sequence $\{\mathbf{u}_{(k,j)_i}\}_{i=1}^{\infty}$ for which $\sigma_{(\max,k,j)_i} \to \infty$. Because $\{\mathbf{u}_{k,j}\}$ is bounded there exists a convergent subsequence $\{(\mathbf{u}_{(k,j)_i})_p\} \to \hat{\mathbf{u}}$. The fact that $\nabla^2 T(\mathbf{u})$ is square implies that $\text{trace}(\nabla^2 T(\mathbf{u})) = \sum_{i=1}^{n} \sigma_i$, where $\{\sigma_i\}_{i=1}^{n}$ are the eigenvalues of $\nabla^2 T(\mathbf{u})$. $(\sigma_{(\max,k,j)_i})_p \to \infty$ implies that $\text{trace}(\nabla^2 T((\mathbf{u}_{(k,j)_i})_p)) \to \infty$. Now by assumption, the elements of $\nabla^2 T(\mathbf{u})$ are continuous and so it follows that $\text{trace}(\nabla^2 T(\mathbf{u}))$ is continuous. Therefore $\text{trace}(\nabla^2 T((\mathbf{u}_{(k,j)_i})_p)) \to \text{trace}(\nabla^2 T(\hat{\mathbf{u}}))$ contradicts $\text{trace}(\nabla^2 T((\mathbf{u}_{(k,j)_i})_p)) \to \infty$. Thus $\{\sigma_{\max,k,j}\}$ is bounded.

Now let M be the upper bound of the set $\{\sigma_{\max,k,j}\}$ and assume that $\{\sigma_{\min,k,j}\}$ is not bounded away from zero. Then there exists a sequence $\{\mathbf{u}_{(k,j)_i}\}_{i=1}^{\infty}$ for which $\sigma_{(\min,k,j)_i} \to 0$, and the fact that $\{\mathbf{u}_{k,j}\}$ is bounded means there is a convergent subsequence $\{(\mathbf{u}_{(k,j)_i})_p\} \to \hat{\mathbf{u}}$. Note that since $\det(\nabla^2 T(\mathbf{u})) = \Pi_{i=1}^{n} \sigma_i$ it is true that $\det(\nabla^2 T(\mathbf{u}_{k,j})) \leq M^{n-1} \sigma_{\min,k,j}$. $(\sigma_{(\min,k,j)_i})_p \to 0$ implies that $\det(\nabla^2 T((\mathbf{u}_{(k,j)_i})_p)) \to 0$. However the continuity of the elements of $\nabla^2 T(\mathbf{u})$, and hence $\det(\nabla^2 T(\mathbf{u}))$, give that $\det(\nabla^2 T((\mathbf{u}_{(k,j)_i})_p)) \to \det(\nabla^2 T(\hat{\mathbf{u}})) = 0$, which contradicts the assumption that $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u} \in \Omega$. □

The proof of convergence requires that the concept of a stationary point be defined.

Definition 2.1.7. $\overline{\mathbf{u}} \in \Omega$ *is a stationary point if for all* $\mathbf{y} \in \Omega$,

$$\langle \nabla T(\overline{\mathbf{u}}), \mathbf{y} - \overline{\mathbf{u}} \rangle \geq 0. \tag{2.49}$$

I will now state and prove a lemma that gives a sufficient condition for a solution to (2.1).

Lemma 2.1.8. *Given that T is strictly convex and twice continuously differentiable, $\overline{\mathbf{u}}$ is the unique solution to (2.1) if and only if it is a stationary point.*

Proof. Suppose that $\overline{\mathbf{u}}$ is a solution to (2.1) and let $\mathbf{v} \in \Omega$. Since Ω is convex the line segment joining $\overline{\mathbf{u}}$ and $\mathbf{v}$ lies entirely in Ω. So the function

$$\phi(\nu) = T(\overline{\mathbf{u}} + \nu(\mathbf{v} - \overline{\mathbf{u}}))$$

is defined for all $\nu \in [0,1]$ and has a local minimizer at $\nu = 0$. Therefore

$$0 \leq \phi'(0) = \nabla T(\overline{\mathbf{u}})^T (\mathbf{v} - \overline{\mathbf{u}}).$$

If there exists $\mathbf{w} \in \Omega$ such that $T(\mathbf{w}) = T(\overline{\mathbf{u}})$ and $\mathbf{w} \neq \overline{\mathbf{u}}$, then the strict convexity of T gives that $T((\overline{\mathbf{u}} + \mathbf{w})/2) < T(\overline{\mathbf{u}})$, which contradicts $\overline{\mathbf{u}}$ being a solution to (2.1).

Now suppose that $\overline{\mathbf{u}}$ is a stationary point. Lemma (2.1.1 and (2.49) give that

$$T(\mathbf{v}) > T(\overline{\mathbf{u}}) + \nabla T(\overline{\mathbf{u}})^T (\mathbf{v} - \overline{\mathbf{u}}) \geq T(\overline{\mathbf{u}}).$$

□

Now I can state the main result on the convergence of the GPRN algorithm.

Theorem 2.1.9. *The iterates $\{\mathbf{u}_k\}$ generated by GPRN converge to the unique solution to prob-*

lem (2.1) for any initial guess $\mathbf{u}_0 \in \Omega$.

Proof. The coercivity of T combined with the fact that $\{T(\mathbf{u}_k)\}_{k=0}^{\infty}$ is monotone decreasing and bounded from below implies that $\{\mathbf{u}_k\}$ is a bounded set and so has a convergent subsequence. Let $\{\mathbf{u}_{k_l}\}$ be the convergent sequence and $\overline{\mathbf{u}}$ be its limit. Then the continuity of T implies that $T(\mathbf{u}_{k_l}) \to T(\overline{\mathbf{u}})$. Since $\{T(\mathbf{u}_k)\}$ is monotone decreasing it must be the case that $T(\mathbf{u}_k) \to T(\overline{\mathbf{u}})$. Otherwise there would exist some m such that $T(\mathbf{u}_k) < T(\overline{\mathbf{u}})$ for all $k \geq m$. But then that would imply that for all $k_l > m$, $T(\mathbf{u}_{k_l}) < T(\mathbf{u}_m) < T(\overline{\mathbf{u}})$ which would contradict the fact that $\mathbf{u}_{k_l} \to \overline{\mathbf{u}}$.

Let $\mathbf{u}_{k_l,0}(= \mathbf{u}_{k_l})$ and $\mathbf{u}_{k_l,1}$ be defined as in Stage 1 of the GPRN iteration. Then $T(\mathbf{u}_{k_l,0}) > T(\mathbf{u}_{k_l,1})$ and so $T(\mathbf{u}_{k_l,1}) \to T(\overline{\mathbf{u}})$. Equation (2.12) gives that

$$\|\mathbf{u}_{k_l,0} - \mathbf{u}_{k_l,1}\|^2 \leq \frac{\lambda_{k_l,0}}{\mu}[T(\mathbf{u}_{k_l,0}) - T(\mathbf{u}_{k_l,1})],$$

and because Lemma 2.1.6 states that the $\lambda_{k_l,0}$'s are bounded above $T(\mathbf{u}_{k_l,0}) - T(\mathbf{u}_{k_l,1})$ converges to zero and so

$$\|\mathbf{u}_{k_l,0} - \mathbf{u}_{k_l,1}\| \to 0. \tag{2.50}$$

It follows from Lemma 2.1.3 that for all $\mathbf{y} \in \Omega$,

$$\begin{aligned}
\langle \nabla T(\mathbf{u}_{k_l}), \mathbf{u}_{k_l} - \mathbf{y} \rangle &= \langle \nabla T(\mathbf{u}_{k_l,0}), \mathbf{u}_{k_l,1} - \mathbf{y} \rangle + \langle \nabla T(\mathbf{u}_{k_l,0}), \mathbf{u}_{k_l,0} - \mathbf{u}_{k_l,1} \rangle \\
&\leq \frac{1}{\lambda_{k_l,0}} \langle \mathbf{u}_{k_l,0} - \mathbf{u}_{k_l,1}, \mathbf{u}_{k_l,1} - \mathbf{y} \rangle + \langle \nabla T(\mathbf{u}_{k_l,0}), \mathbf{u}_{k_l,0} - \mathbf{u}_{k_l,1} \rangle \\
&\leq \|\mathbf{u}_{k_l,0} - \mathbf{u}_{k_l,1}\| \cdot \left\| \frac{\mathbf{u}_{k_l,1} - \mathbf{y}}{\lambda_{k_l,0}} + \nabla T(\mathbf{u}_{k_l,0}) \right\|.
\end{aligned}$$

The $\lambda_{k_l,0}$'s are bounded below by Lemma 2.1.6. Since $\{\mathbf{u}_k\}$ is a bounded set and ∇T is Lipschitz continuous, ∇T is bounded on $\{\mathbf{u}_{k_l}\}$. Letting $k_l \to \infty$ it follows from (2.50) that $\langle \nabla T(\overline{\mathbf{u}}), \mathbf{u} - \mathbf{y} \rangle \leq 0$ for all $\mathbf{y} \in \Omega$. $\overline{\mathbf{u}}$ therefore satisfies the definition of a stationary point and by Lemma 2.1.8 $\overline{\mathbf{u}}$ is a solution of problem (2.1).

It remains to be shown that $\mathbf{u}_k \to \bar{\mathbf{u}}$. Given $\mathbf{v} \in \mathbb{R}^n$ such that $\bar{\mathbf{u}} + \mathbf{v} \in \Omega$, it follows from Taylor's theorem that there exists $\nu \in (0,1)$ for which

$$T(\bar{\mathbf{u}} + \mathbf{v}) = T(\bar{\mathbf{u}}) + \langle \nabla T(\bar{\mathbf{u}}), \mathbf{v} \rangle + \frac{1}{2} \langle \nabla^2 T(\bar{\mathbf{u}} + \nu \mathbf{v})\mathbf{v}, \mathbf{v} \rangle. \tag{2.51}$$

Let $\nu_k \in (0,1)$ be the value of ν which satisfies equation (2.51) when $\mathbf{v} = \mathbf{u}_k - \bar{\mathbf{u}}$. It follows that

$$\begin{aligned} T(\mathbf{u}_k) - T(\bar{\mathbf{u}}) &= \langle \nabla T(\bar{\mathbf{u}}), \mathbf{u}_k - \bar{\mathbf{u}} \rangle + \frac{1}{2} \langle \nabla^2 T(\bar{\mathbf{u}} + \nu_k(\mathbf{u}_k - \bar{\mathbf{u}}))(\mathbf{u}_k - \bar{\mathbf{u}}), \mathbf{u}_k - \bar{\mathbf{u}} \rangle & (2.52) \\ &\geq \frac{1}{2} \langle \nabla^2 T(\bar{\mathbf{u}} + \nu(\mathbf{u}_k - \bar{\mathbf{u}}))(\mathbf{u}_k - \bar{\mathbf{u}}), \mathbf{u}_k - \bar{\mathbf{u}} \rangle & (2.53) \\ &\geq \frac{\sigma_{\min}}{2} \|\mathbf{u}_k - \bar{\mathbf{u}}\|^2, & (2.54) \end{aligned}$$

where equation (2.53) follows from Definition 2.1.7, and $\sigma_{\min}$ is defined as follows. Consider the set $\{\bar{\mathbf{u}} + \nu_k(\mathbf{u}_k - \bar{\mathbf{u}})\}$ and define $\sigma_{\min,\nu_k}$ to be the minimum eigenvalue of $\nabla^2 T(\bar{\mathbf{u}} + \nu_k(\mathbf{u}_k - \bar{\mathbf{u}}))$. $\sigma_{\min}$ is then defined to be the $\inf\{\sigma_{\min,\nu_k}\}$. The same reasoning that was used in the proof of Lemma 2.1.6 to show that the set $\sigma_{\min,k,j}$ was bounded away from zero can be used to show that $\{\sigma_{\min,\nu_k}\}$ is also bounded away from zero. Hence $\sigma_{\min} > 0$. Thus $T(\mathbf{u}_k) \to T(\bar{\mathbf{u}})$ implies $\mathbf{u}_k \to \bar{\mathbf{u}}$.

□

Chapter 3

Regularization Functions

This chapter contains material from [4, 9].

Astronomical and medical imaging data consists of the photon counts that are recorded at each pixel in an $n_x \times n_y$ array of pixels. Let $\mathbf{z}$ be the vector obtained by the lexicographical column ordering of the data. The photon counts are noisy and the data model for $\mathbf{z}$ is given by

$$\mathbf{Z} = \text{Poiss}(\mathbf{A}\mathbf{u}_e + \gamma), \tag{3.1}$$

where $\text{Poiss}(\lambda)$ is a Poisson-distributed independent random vector with mean λ, $\mathbf{u}_e \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n\times n}$, $n = n_x n_y$ and $\gamma = \gamma\mathbf{1}$ with $\gamma > 0$. $\mathbf{u}_e$ is the vector obtained by lexicographical column ordering of the underlying true image array. The probability mass function for data $\mathbf{z}$ from (3.1) is

$$p(\mathbf{z} \mid \mathbf{u}) = \prod_{i=1}^{n} \frac{e^{-[\mathbf{A}\mathbf{u}]_i + \gamma_i}([\mathbf{A}\mathbf{u}]_i + \gamma_i)^{z_i}}{z_i!}, \tag{3.2}$$

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

Given data $\mathbf{z}$ that is a realization of the random vector $\mathbf{Z}$ it is of interest to estimate $\mathbf{u}_e$. The maximum

likelihood estimate $\mathbf{u}_{\mathrm{ML}}$ is found by maximizing $p(\mathbf{z};\mathbf{u})$ in (3.2) with respect to $\mathbf{u}$ with the constraint that $\mathbf{u} \geq \mathbf{0}$. Alternatively $\mathbf{u}_{\mathrm{ML}}$ is found by solving

$$\mathbf{u}_{\mathrm{ML}} = \mathrm{argmin}_{\mathbf{u}\geq 0}\{T_0(\mathbf{u};\mathbf{z}) \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} [\mathbf{A}\mathbf{u}]_i + \gamma_i - z_i \ln([\mathbf{A}\mathbf{u}]_i + \gamma_i)\}. \tag{3.3}$$

Note that $T_0(\mathbf{z};\mathbf{u})$ is equal to $-\ln p(\mathbf{z} \mid \mathbf{u})$ plus an additive constant.

The solution of (3.3) can be unstable with respect to the noise in $\mathbf{z}$ if $\mathbf{A}$ is ill-conditioned. Regularization is therefore required, and Bayes Law provides statistical motivation for formulating a regularized problem. In this context, $\mathbf{u}_e$ is assumed to be a realization of a random vector $\mathbf{U}$. Given a probability density $p(\mathbf{u})$ for $\mathbf{U}$ the posterior density is given by

$$p(\mathbf{u} \mid \mathbf{z}) = \frac{p(\mathbf{z} \mid \mathbf{u})p(\mathbf{u})}{p(\mathbf{z})}. \tag{3.4}$$

The maximum a posteriori (MAP) estimate $\mathbf{u}_{\mathrm{MAP}}$ is found by maximizing (3.4) with respect to $\mathbf{u}$. This is equivalent to minimizing $T(\mathbf{u}) = T_0(\mathbf{u};\mathbf{z}) - \ln p(\mathbf{u})$. The function $-\ln p(\mathbf{u})$ corresponds to the penalty term in a deterministic setting. The probability density $p(\mathbf{u})$ is determined by the distribution, known as the prior, that $\mathbf{u}_e$ is assumed to arise from. $\mathbf{u}_{\mathrm{MAP}}$ is therefore the solution of

$$\mathbf{u}_{\mathrm{MAP}} = \mathrm{argmin}_{\mathbf{u}\geq 0} \left\{ T_\alpha(\mathbf{u}) \stackrel{\mathrm{def}}{=} T_0(\mathbf{u};\mathbf{z}) - \ln p(\mathbf{u}) \right\}. \tag{3.5}$$

The existence of $\mathbf{u}_{\mathrm{MAP}}$ requires that the cost function T_α possess certain properties on its domain, namely coercivity and weak lower semicontinuity.

Definition 3.0.10. *A function $T : \mathscr{H} \to \mathbb{R}$ is coercive if $T(\mathbf{u}_k) \to \infty$ for any sequence $\{\mathbf{u}_k\}_{k=0}^{\infty}$ of elements of $\mathscr{H}$ for which $\|\mathbf{u}_k\|_2 \to \infty$ as $k \to \infty$.*

The definition of weak lower semicontinuity requires that weak convergence be defined.

Definition 3.0.11. *A sequence $\{f_n\}$ in a Hilbert space $\mathscr{H}$ converges weakly to f_*, denoted by $f_n \rightharpoonup f_*$, if for all $f \in \mathscr{H}$, $\lim_{n\to\infty} \langle f_n - f_*, f \rangle_{\mathscr{H}} = 0$.*

Definition 3.0.12. *A functional $T : \mathscr{H} \to \mathbb{R}$ is weakly lower semicontinuous if*

$$T(f_*) \leq \liminf_{n\to\infty} T(f_n) \quad \text{whenever} \quad f_n \rightharpoonup f_*. \tag{3.6}$$

In finite-dimensional spaces weak convergence is equivalent to strong convergence and continuity implies weak lower semicontinuity. Also convex functionals are weakly lower semicontinuous [34].

A theorem can now be stated concerning the existence of a minimizer of T_α [34, Chapter 2]

Theorem 3.0.13. *Suppose that $T : \mathscr{H} \to \mathbb{R}$ is weakly lower semicontinuous and coercive and that Ω is a closed and convex set. Then T has a minimizer over Ω and the minimizer is unique if T is strictly convex.*

Proof. Suppose that $\{\mathbf{u}_k\}$ is a sequence for which $T(\mathbf{u}_k) \to T_* \stackrel{\text{def}}{=} \inf_{\mathbf{u}_k\in\Omega} T(\mathbf{u}_k)$. That $\{\mathbf{u}_k\}$ is a bounded sequence follows from the coercivity of T. Bounded sequences in Hilbert spaces have weakly convergent subsequences [34]. Let $\{\mathbf{u}_{k_j}\}$ denote the weakly convergent subsequence of $\{\mathbf{u}_k\}$ and $\mathbf{u}_*$ its limit and note that $\mathbf{u}_* \in \Omega$ because Ω, being closed and convex, is weakly closed [37]. It follows from the weak lower semicontinuity of T that

$$T(\mathbf{u}_*) \leq \liminf T(\mathbf{u}_{k_j}) = \lim T(\mathbf{u}_k) = T_*,$$

and hence $T(\mathbf{u}_*) = T_*$. Now, suppose T is strictly convex. Then $\mathbf{u}_*$ is unique because if for some $\mathbf{v} \neq \mathbf{u}_*$, $T(\mathbf{v}) = T_*$ the strict convexity of T implies that $T_* > T((\mathbf{v}+\mathbf{u}_*)/2)$ which contradicts the definition of T_*. □

In the next two sections I will discuss two different options for the form of $-\ln p(\mathbf{u})$ and show that under certain assumptions, each option yields a cost function which has a unique minimizer on Ω.

3.1 Tikhonov Regularization

One option for the form of $-\ln p(\mathbf{u})$ that I will examine is to assume that

$$-\ln p(\mathbf{u}) = \alpha/2\langle \mathbf{C}\mathbf{u}, \mathbf{u}\rangle, \tag{3.7}$$

where $\mathbf{C}$ is symmetric positive semi-definite. This is equivalent to assuming that $\mathbf{U}$ arises from a Gaussian with mean $\mathbf{0}$ and covariance $\alpha^{-1}\mathbf{C}^{\dagger}$, where "†" denotes pseudo-inverse. This option yields a regularization term equivalent to that which is found in Tikhonov regularization. The choice of the matrix $\mathbf{C}$ must be justified by the prior knowledge of $\mathbf{u}_e$ that is available. Also necessary for our analysis is the assumption that $\mathbf{A}$ and $\mathbf{C}$ have non-intersecting null spaces.

The most common choice, $\mathbf{C} = \mathbf{I}$, where $\mathbf{I}$ is the $n \times n$ identity matrix, yields a cost function which penalizes proposed reconstructions with a large ℓ^2 norm. If $\mathbf{u}_e$ is known to be smooth then an appropriate choice is $\mathbf{C} = \mathbf{L}$, where $\mathbf{L}$ is a discretization of the negative Laplacian operator $-\nabla^2$. This choice yields a cost function which penalizes proposed non-smooth reconstructions.

A third choice for the form of $\mathbf{C}$ is presented in [4] and should be used when it is known a priori that $\mathbf{u}_e$ is smooth with the exception of discontinuities, the location of which are known (at least approximately in practice). This form is given by

$$\mathbf{C} = \mathbf{D}_1^T \Lambda \mathbf{D}_1 + \mathbf{D}_2^T \Lambda \mathbf{D}_2, \tag{3.8}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are discretizations of the horizontal and vertical derivatives respectively, and Λ is a diagonal matrix with $[\Lambda]_{i,i} = 1$ if the ith pixel is away from an edge and $[\Lambda]_{i,i} < 1$ if the ith pixel is adjacent to an edge. The problem of constructing Λ is discussed in the presentation of the numerical experiments in Chapter 5. Note that these three choices for $\mathbf{C}$ are all positive semi-definite [11].

3.1.1 Analysis of the cost function for the case of a quadratic regularization term

In this section, I will show that cost functions of the form

$$T(\mathbf{u}) = T_0(\mathbf{u};\mathbf{z}) + \frac{\alpha}{2}\mathbf{u}^T\mathbf{C}\mathbf{u}, \tag{3.9}$$

where $T_0(\mathbf{u};\mathbf{z})$ is given in (3.3), $\alpha > 0$ is the regularization parameter, and $\mathbf{C} \in \mathbb{R}^{n\times n}$ is a positive semi-definite matrix, are strictly convex and coercive on Ω. Note this is the cost function that results when $-\ln p(\mathbf{u})$ in (4.1) is given by equation (3.7).

The gradient and Hessian of $T_0(\mathbf{u};\mathbf{z})$ with respect to $\mathbf{u}$ are given, respectively, by

$$\nabla T_0(\mathbf{u};\mathbf{z}) = \mathbf{A}^T\left(\frac{\mathbf{A}\mathbf{u}-(\mathbf{z}-\gamma)}{\mathbf{A}\mathbf{u}+\gamma}\right), \tag{3.10}$$

$$\nabla^2 T_0(\mathbf{u};\mathbf{z}) = \mathbf{A}^T \mathrm{diag}\left(\frac{\mathbf{z}}{(\mathbf{A}\mathbf{u}+\gamma)^2}\right)\mathbf{A}, \tag{3.11}$$

where $\mathrm{diag}(\mathbf{v})$ is the diagonal matrix with its diagonal given by $\mathbf{v}$. For $\mathbf{v},\mathbf{w} \in \mathbb{R}^n$, the notation $\mathbf{v}/\mathbf{w}$ denotes component-wise division and $\mathbf{v}^2$ denotes the vector obtained from squaring the components of $\mathbf{v}$. Note then that $\nabla T(\mathbf{u}) = \nabla T_0(\mathbf{u};\mathbf{z}) + \alpha\mathbf{C}\mathbf{u}$ and $\nabla^2 T(\mathbf{u}) = \nabla^2 T_0(\mathbf{u};\mathbf{z}) + \alpha\mathbf{C}$.

The assumptions are made that $\mathbf{z} > \mathbf{0}$, $\mathbf{A}\mathbf{u} \in \Omega$ when $\mathbf{u} \in \Omega$, and that the intersection of the null spaces of $\mathbf{A}$ and $\mathbf{C}$ is trivial. Then it follows that for $\mathbf{v},\mathbf{u} \in \Omega$,

$$\begin{aligned}\langle \nabla^2 T(\mathbf{u})\mathbf{v},\mathbf{v}\rangle &= \left\langle \left(\mathbf{A}^T \operatorname{diag}\left(\frac{\mathbf{z}}{(\mathbf{Au}+\gamma)^2}\right)\mathbf{A}+\mathbf{C}\right)\mathbf{v},\mathbf{v}\right\rangle \\ &\geq \min_{i=1,\ldots,n}\left\{\frac{z_i}{([\mathbf{Au}]_i+\gamma)^2}\right\}\|\mathbf{Av}\|_2^2+\langle \mathbf{Cv},\mathbf{v}\rangle \qquad (3.12)\\ &> 0, \qquad (3.13)\end{aligned}$$

since $\mathbf{A}$ and $\mathbf{C}$ have non-intersecting null spaces.. Thus $\nabla^2 T(\mathbf{u})$ is positive definite for all $\mathbf{u}\in\Omega$ which implies (see Lemma 1.1.3) that T is a strictly convex function on Ω.

In addition to strict convexity, it is desirable that T also have the property of coercivity on Ω. Note that since $\mathbf{Au}\in\Omega$ when $\mathbf{u}\in\Omega$, $|[\mathbf{Au}]_i| = [\mathbf{Au}]_i$. It follows that for $\mathbf{u}\in\Omega$

$$T(\mathbf{u}) \geq \|\mathbf{Au}+\gamma\|_1 - \|\mathbf{z}\|_\infty \sum_{i=1}^{n} \ln([\mathbf{Au}]_i+\gamma_i) + \frac{\alpha}{2}\mathbf{u}^T\mathbf{Cu}. \qquad (3.14)$$

Now for $c>0$, $-c\ln x$ is a convex function of x and so Jensen's inequality implies that

$$-\|\mathbf{z}\|_\infty \sum_{i=1}^{n} \ln([\mathbf{Au}]_i+\gamma_i) \geq -n\|\mathbf{z}\|_\infty \ln\|\mathbf{Au}+\gamma\|_1. \qquad (3.15)$$

Inequalities (3.14), (3.15) give that

$$T(\mathbf{u}) \geq \|\mathbf{Au}+\gamma\|_1 - n\|\mathbf{z}\|_\infty \ln\|\mathbf{Au}+\gamma\|_1 + \frac{\alpha}{2}\mathbf{u}^T\mathbf{Cu}, \qquad (3.16)$$

and the assumption that $\mathbf{A}$ and $\mathbf{C}$ have non-intersecting null spaces implies that $\sup\{\|\mathbf{Au}\|_1, \mathbf{u}^T\mathbf{Cu}\}\to\infty$ as $\|\mathbf{u}\|_2\to\infty$. Furthermore $x - c\ln x\to\infty$ as $x\to\infty$. Thus $T(\mathbf{u})\to\infty$ as $\|\mathbf{u}\|_2\to\infty$.

The convergence of the iterative method that is used to compute $\mathbf{u}_{\mathrm{MAP}}$ requires that ∇T be Lipschitz continuous on Ω. Let $\mathbf{u},\mathbf{v}\in\Omega$ and note that

$$\begin{aligned} \|\nabla T(\mathbf{u}) - \nabla T(\mathbf{v})\|_2 &= \left\| \mathbf{A}^T \left(\frac{\mathbf{Au}+\gamma-\mathbf{z}}{\mathbf{Au}+\gamma} - \frac{\mathbf{Av}+\gamma-\mathbf{z}}{\mathbf{Av}+\gamma} \right) + \alpha \mathbf{C}(\mathbf{u}-\mathbf{v}) \right\|_2 \\ &\leq \|\mathbf{A}\|_2 F(\mathbf{u},\mathbf{v}) + \alpha \sigma_{\max}(\mathbf{C}) \|\mathbf{u}-\mathbf{v}\|_2, \end{aligned} \tag{3.17}$$

where $\sigma_{\max}(\mathbf{C})$ denotes the maximum eigenvalue of $\mathbf{C}$ and

$$\begin{aligned} F(\mathbf{u},\mathbf{v}) &= \left\| \frac{(\mathbf{A}(\mathbf{u}-\mathbf{v})) \odot \mathbf{z}}{(\mathbf{Au}+\gamma) \odot (\mathbf{Av}+\gamma)} \right\|_2 \\ &\leq \|\mathbf{A}\|_2 \left\| \frac{\mathbf{z}}{\gamma^2} \right\|_2 \|\mathbf{u}-\mathbf{v}\|_2. \end{aligned} \tag{3.18}$$

Inequality (3.17) combined with inequality (3.18) implies that

$$\|\nabla T(\mathbf{u}) - \nabla T(\mathbf{v})\|_2 \leq \left(\|\mathbf{A}\|_2^2 \left\| \frac{\mathbf{z}}{\gamma^2} \right\|_2 + \alpha \sigma_{\max}(\mathbf{C}) \right) \|\mathbf{u}-\mathbf{v}\|_2.$$

Thus ∇T is Lipschitz continuous. Hence the optimization method of Chapter 2 is convergent for Tikhonov regularization.

3.2 Total Variation Regularization

Another option for the form of $-\ln p(\mathbf{u})$ is to assume that $-\ln p(\mathbf{u}) = \alpha J(\mathbf{u})$, where $J(\mathbf{u})$ is a discretization of an approximation of the total variation (TV) function. This assumption is equivalent to assuming that $\mathbf{U}$ arises from the TV prior. Here

$$J(\mathbf{u}) \stackrel{\text{def}}{=} \frac{1}{2} \sum_{i=1}^{n} \psi([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2), \tag{3.19}$$

where $\psi(t) := \sqrt{t+\beta}$. β is a small positive parameter that is included to ensure that $J(\mathbf{u})$ is differentiable [34]. Total variation regularization should be used when $\mathbf{u}_e$ is known to be "blocky", i.e. piecewise constant, with the exception of jump discontinuities, and the length of the curves on which the discontinuities occur is relatively small.

3.2.1 Analysis of the cost function for the case of total variation regularization

When total variation regularization is used the cost function that must be minimized is

$$T(\mathbf{u}) = T_0(\mathbf{u};\mathbf{z}) + \alpha J(\mathbf{u}), \tag{3.20}$$

where $J(\mathbf{u})$ is given by (3.19). Using arguments similar to those in subsection 3.1.1 it can be shown that T is strictly convex and coercive on Ω. Such calculations require that the gradient and Hessian of J be known.

Given $\mathbf{u} \in \Omega$, to compute the gradient of J at $\mathbf{u}$ note that for $\mathbf{v} \in \Omega$,

$$\begin{aligned} \frac{d}{d\tau} J(\mathbf{u}+\tau\mathbf{v}) \mid_{\tau=0} &= \sum_{i=1}^{n} \psi'([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2)([\mathbf{D}_1\mathbf{u}]_i[\mathbf{D}_1\mathbf{v}]_i + [\mathbf{D}_2\mathbf{u}]_i[\mathbf{D}_2\mathbf{v}]_i) \\ &= \langle \mathrm{diag}(\psi'(\mathbf{D}\mathbf{u}^2))\mathbf{D}_1\mathbf{u}, \mathbf{D}_1\mathbf{v}\rangle + \langle \mathrm{diag}(\psi'(\mathbf{D}\mathbf{u}^2))\mathbf{D}_2\mathbf{u}, \mathbf{D}_2\mathbf{v}\rangle, \end{aligned} \tag{3.21}$$

where $\mathbf{D}\mathbf{u}^2 := (\mathbf{D}_1\mathbf{u})^2 + (\mathbf{D}_2\mathbf{u})^2$ and $\psi'(\mathbf{D}\mathbf{u}^2)$ is the vector whose ith component is $\psi'([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2)$. Here $\mathbf{v}^2 = \mathbf{v} \odot \mathbf{v}$. Hence the gradient of J has the form

$$\nabla J(\mathbf{u}) = \mathbf{L}_1(\mathbf{u})\mathbf{u}, \tag{3.22}$$

where

$$\mathbf{L}_1(\mathbf{u}) = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \end{bmatrix}^T \begin{bmatrix} \operatorname{diag}(\psi'(\mathbf{D}\mathbf{u}^2)) & \mathbf{0} \\ \mathbf{0} & \operatorname{diag}(\psi'(\mathbf{D}\mathbf{u}^2)) \end{bmatrix} \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \end{bmatrix}. \tag{3.23}$$

Given $\mathbf{v}, \mathbf{w} \in \Omega$, the Hessian of J at $\mathbf{u}$ can be computed as follows:

$$\begin{aligned} \frac{\partial^2}{\partial\tau\partial\xi} J(\mathbf{u}+\tau\mathbf{v}+\xi\mathbf{w})\Big|_{\tau=\xi=0} &= \sum_{i=1}^{n} \psi'([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2)([\mathbf{D}_1\mathbf{w}]_i[\mathbf{D}_1\mathbf{v}]_i + [\mathbf{D}_2\mathbf{w}]_i[\mathbf{D}_2\mathbf{v}]_i) \\ &\quad + \sum_{i=1}^{n} \psi''([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2)(2([\mathbf{D}_1\mathbf{u}]_i^2[\mathbf{D}_1\mathbf{w}]_i[\mathbf{D}_1\mathbf{v}]_i + [\mathbf{D}_1\mathbf{u}]_i[\mathbf{D}_2\mathbf{u}]_i[\mathbf{D}_1\mathbf{w}]_i[\mathbf{D}_2\mathbf{v}]_i) \\ &\quad + 2([\mathbf{D}_2\mathbf{u}]_i^2[\mathbf{D}_2\mathbf{w}]_i[\mathbf{D}_2\mathbf{v}]_i + [\mathbf{D}_2\mathbf{u}]_i[\mathbf{D}_1\mathbf{u}]_i[\mathbf{D}_2\mathbf{w}]_i[\mathbf{D}_1\mathbf{v}]_i)) \\ &= \langle \operatorname{diag}(\psi'(\mathbf{D}\mathbf{u}^2))\mathbf{D}_1\mathbf{w}, \mathbf{D}_1\mathbf{v}\rangle + \langle \operatorname{diag}(\psi'(\mathbf{D}\mathbf{u}^2))\mathbf{D}_2\mathbf{w}, \mathbf{D}_2\mathbf{v}\rangle \\ &\quad + 2\langle \operatorname{diag}(\mathbf{D}_1\mathbf{u}^2\psi''(\mathbf{D}\mathbf{u}^2))\mathbf{D}_1\mathbf{w}, \mathbf{D}_1\mathbf{v}\rangle + 2\langle \operatorname{diag}(\mathbf{D}_{12}\mathbf{u}\psi''(\mathbf{D}\mathbf{u}^2))\mathbf{D}_1\mathbf{w}, \mathbf{D}_2\mathbf{v}\rangle \\ &\quad + 2\langle \operatorname{diag}(\mathbf{D}_{12}\mathbf{u}\psi''(\mathbf{D}\mathbf{u}^2))\mathbf{D}_2\mathbf{w}, \mathbf{D}_1\mathbf{v}\rangle + 2\langle \operatorname{diag}(\mathbf{D}_2\mathbf{u}^2\psi''(\mathbf{D}\mathbf{u}^2))\mathbf{D}_1\mathbf{w}, \mathbf{D}_1\mathbf{v}\rangle, \end{aligned}$$

where $\psi''(\mathbf{D}\mathbf{u}^2)$ is the vector whose ith component is given by $\psi''([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2)$ and

$\mathbf{D}_{12}\mathbf{u} = \mathbf{D}_1\mathbf{u} \odot \mathbf{D}_2\mathbf{u}$. It follows that the Hessian of J can be written as

$$\nabla^2 J(\mathbf{u}) = \mathbf{L}_1(\mathbf{u}) + 2\mathbf{L}_2(\mathbf{u}), \tag{3.24}$$

where

$$\mathbf{L}_2 = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \end{bmatrix}^T \begin{bmatrix} \operatorname{diag}((\mathbf{D}_1\mathbf{u})^2 \odot \psi''(\mathbf{D}\mathbf{u}^2)) & \operatorname{diag}(\mathbf{D}_{12}\mathbf{u} \odot \psi''(\mathbf{D}\mathbf{u}^2)) \\ \operatorname{diag}(\mathbf{D}_{12}\mathbf{u} \odot \psi''(\mathbf{D}\mathbf{u}^2)) & \operatorname{diag}((\mathbf{D}_1\mathbf{u})^2 \odot \psi''(\mathbf{D}\mathbf{u}^2)) \end{bmatrix} \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \end{bmatrix}. \tag{3.25}$$

Note that for $\psi(t) = \sqrt{t+\beta}$, $\psi'(t) = (t+\beta)^{-1/2}/2$ and $\psi''(t) = -(t+\beta)^{-3/2}/4$. Now, for any

$\mathbf{v}, \mathbf{u} \in \Omega$, $\mathbf{v} \neq \mathbf{0}$,

$$\begin{aligned}
\langle \nabla^2 J(\mathbf{u})\mathbf{v}, \mathbf{v} \rangle &= \mathbf{v}^T (\mathbf{L}_1(\mathbf{u}) + 2\mathbf{L}_2(\mathbf{u}))\mathbf{v} \\
&= \sum_{i=1}^{n} [\mathbf{D}_1 \mathbf{v}]_i^2 \left(\frac{1}{2\sqrt{[\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta}} - \frac{[\mathbf{D}_1\mathbf{u}]_i^2}{2([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta)^{\frac{3}{2}}} \right) \\
&\quad - \sum_{i=1}^{n} [\mathbf{D}_1 \mathbf{v}]_i [\mathbf{D}_2 \mathbf{v}]_i \frac{[\mathbf{D}_1\mathbf{u}]_i [\mathbf{D}_2\mathbf{u}]_i}{([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta)^{\frac{3}{2}}} \\
&\quad + \sum_{i=1}^{n} [\mathbf{D}_2 \mathbf{v}]_i^2 \left(\frac{1}{2\sqrt{[\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta}} - \frac{[\mathbf{D}_2\mathbf{u}]_i^2}{2([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta)^{\frac{3}{2}}} \right) \\
&= \frac{1}{2} \left(\sum_{i=1}^{n} \frac{[\mathbf{D}_1 \mathbf{v}]_i^2 ([\mathbf{D}_2\mathbf{u}]_i^2 + \beta)}{([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta)^{\frac{3}{2}}} - \sum_{i=1}^{n} \frac{2[\mathbf{D}_1 \mathbf{v}]_i [\mathbf{D}_2 \mathbf{v}]_i [\mathbf{D}_1\mathbf{u}]_i [\mathbf{D}_2\mathbf{u}]_i}{([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta)^{\frac{3}{2}}} \right. \\
&\quad \left. + \sum_{i=1}^{n} \frac{[\mathbf{D}_2 \mathbf{v}]_i^2 ([\mathbf{D}_1\mathbf{u}]_i^2 + \beta)}{([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta)^{\frac{3}{2}}} \right) \\
&= \frac{1}{2} \sum_{i=1}^{n} \frac{([\mathbf{D}_1 \mathbf{v}]_i [\mathbf{D}_2\mathbf{u}]_i - [\mathbf{D}_2 \mathbf{v}]_i [\mathbf{D}_1\mathbf{u}]_i)^2 + \beta([\mathbf{D}_1\mathbf{v}]_i^2 + [\mathbf{D}_2\mathbf{v}]_i^2)}{([\mathbf{D}_1\mathbf{u}]_i^2 + [\mathbf{D}_2\mathbf{u}]_i^2 + \beta)^{\frac{3}{2}}} \\
&\geq 0, \qquad (3.26)
\end{aligned}$$

Thus $\nabla^2 J(\mathbf{u})$ is positive semi-definite for all $\mathbf{u} \in \Omega$.

As in the case of quadratic regularization the assumptions that $\mathbf{Au} \in \Omega$ if $\mathbf{u} \in \Omega$ and $\mathbf{z} > \mathbf{0}$ are made. Assuming that the intersection of the null spaces of $\mathbf{A}$, and $\mathbf{D}_1$ and $\mathbf{D}_2$ are trivial we have that for $\mathbf{u}, \mathbf{v} \in \Omega$,

$$\begin{aligned}
\langle \nabla^2 T(\mathbf{u})\mathbf{v}, \mathbf{v} \rangle &= \langle \mathbf{A}^T \frac{\mathbf{z}}{(\mathbf{Au} + \gamma)^2} \mathbf{Av}, \mathbf{v} \rangle + \langle (\mathbf{L}_1(\mathbf{u}) + 2\mathbf{L}_2(\mathbf{u}))\mathbf{v}, \mathbf{v} \rangle \\
&\geq \min_{i=1,\ldots,n} \left\{ \frac{z_i}{([\mathbf{Au}]_i + \gamma)^2} \right\} \|\mathbf{Av}\|^2 + \langle \nabla^2 J(\mathbf{u})\mathbf{v}, \mathbf{v} \rangle \\
&> 0, \qquad (3.27)
\end{aligned}$$

To establish coercivity of T first note that

$$\begin{aligned} J(\mathbf{u}) &\geq \frac{1}{2}\sum_{i=1}^{n}\sqrt{[\mathbf{D}_1\mathbf{u}]_i^2+[\mathbf{D}_2\mathbf{u}]_i^2} \\ &\geq \frac{1}{2\sqrt{2}}\sum_{i=1}^{n}|[\mathbf{D}_1\mathbf{u}]_i|+||\mathbf{D}_2\mathbf{u}|_i| \qquad (3.28) \\ &= \frac{1}{2\sqrt{2}}(\|\mathbf{D}_1\mathbf{u}\|_1+\|\mathbf{D}_2\mathbf{u}\|_1), \qquad (3.29) \end{aligned}$$

where (3.28) is a consequence of Jensen's inequality applied to the concave function $\sqrt{x}$. So for $\mathbf{u}\in\Omega$ it is the case that

$$T(\mathbf{u}) \geq \|\mathbf{A}\mathbf{u}+\gamma\|_1 - n\|\mathbf{z}\|_\infty \ln\|\mathbf{A}\mathbf{u}+\gamma\|_1 + \frac{\alpha}{2\sqrt{2}}(\|\mathbf{D}_1\mathbf{u}\|_1+\|\mathbf{D}_2\mathbf{u}\|_1). \qquad (3.30)$$

The invertibility of $\mathbf{D}_1$ and $\mathbf{D}_2$ imply that $\max\{\|\mathbf{A}\mathbf{u}\|_1,\|\mathbf{D}_1\mathbf{u}\|_1,\|\mathbf{D}_2\mathbf{u}\|_1\} \xrightarrow[\|\mathbf{u}\|_1\to\infty]{} \infty$ and inequality (3.30) implies that $T(\mathbf{u})\to\infty$ when $\|\mathbf{u}\|_1\to\infty$. Thus T is coercive.

As in the case when the regularization term is quadratic, total variation regularization also yields a cost function for which the gradient is Lipschitz continuous. The Lipschitz continuity of ∇T_0 follows from inequalities (3.17) and (3.18). It remains to be shown that $\nabla J(\mathbf{u})$ is Lipschitz continuous. Given $\mathbf{u},\mathbf{v}\in\Omega$ note that

$$\begin{aligned} \|\nabla J(\mathbf{u})-\nabla J(\mathbf{v})\|_2 &= \|\mathbf{L}_1(\mathbf{u})\mathbf{u}-\mathbf{L}_1(\mathbf{v})\mathbf{v}\|_2 \\ &= \|(\mathbf{D}_1^T\psi'(\mathbf{u})\mathbf{D}_1+\mathbf{D}_2^T\psi'(\mathbf{u})\mathbf{D}_2)\mathbf{u}-(\mathbf{D}_1^T\psi'(\mathbf{v})\mathbf{D}_1+\mathbf{D}_2^T\psi'(\mathbf{v})\mathbf{D}_2)\mathbf{v}\|_2 \\ &\leq \|\mathbf{D}_1^T\psi'(\mathbf{u})\mathbf{D}_1\mathbf{u}-\mathbf{D}_1^T\psi'(\mathbf{v})\mathbf{D}_1\mathbf{v}\|_2+\|\mathbf{D}_2^T\psi'(\mathbf{u})\mathbf{D}_2\mathbf{u}-\mathbf{D}_2^T\psi'(\mathbf{v})\mathbf{D}_2\mathbf{v}\| \\ &\leq \|\mathbf{D}_1^T\|_2\|\psi'(\mathbf{u})\mathbf{D}_1\mathbf{u}-\psi'(\mathbf{v})\mathbf{D}_1\mathbf{v}\|_2+\|\mathbf{D}_2^T\|_2\|\psi'(\mathbf{u})\mathbf{D}_2\mathbf{u}-\psi'(\mathbf{v})\mathbf{D}_2\mathbf{v}\| \\ &\leq \|\mathbf{D}_1^T\|_2(\|\psi'(\mathbf{u})\mathbf{D}_1(\mathbf{u}-\mathbf{v})\|_2+\|(\psi'(\mathbf{u})-\psi'(\mathbf{v}))\mathbf{D}_1\mathbf{v}\|_2)+\|\mathbf{D}_2^T\|_2\|\psi'(\mathbf{u})\mathbf{D}_2(\mathbf{u}-\mathbf{v})\|_2 \\ &\quad +\|\mathbf{D}_2^T\|_2\|(\psi'(\mathbf{u})-\psi'(\mathbf{v}))\mathbf{D}_2\mathbf{v}\|_2. \qquad (3.31) \end{aligned}$$

The definition of ψ' can be used to obtain the following:

$$\begin{aligned}
\|(\psi'(\mathbf{u})-\psi'(\mathbf{v}))\mathbf{D}_1\mathbf{v}\|_1 &= \tfrac{1}{2} \sum_{j=1}^{n} \left| \frac{1}{\sqrt{[\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta}} - \frac{1}{\sqrt{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta}} \right| |[\mathbf{D}_1\mathbf{v}]_j| \\
&= \frac{1}{2}\sum_{j=1}^{n} \left| \frac{\sqrt{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta}-\sqrt{[\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta}}{\sqrt{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta}\sqrt{[\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta}} \right| |[\mathbf{D}_1\mathbf{v}]_j| \\
&= \frac{1}{2}\sum_{j=1}^{n} \frac{\left|\sqrt{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta}-\sqrt{[\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta}\right|}{\sqrt{([\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta)}\sqrt{([\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta)}} |[\mathbf{D}_1\mathbf{v}]_j| \\
&= \frac{1}{2}\sum_{j=1}^{n} \left\| \frac{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2-([\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2)}{\sqrt{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta}+\sqrt{[\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta}} \right| \\
&\quad \cdot \sqrt{\frac{[\mathbf{D}_1\mathbf{v}]_j^2}{([\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta)([\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta)}} \\
&\leq \frac{1}{2\sqrt{\beta}}\sum_{j=1}^{n} \left| \frac{(|[\mathbf{D}_1\mathbf{v}]_j|-|[\mathbf{D}_1\mathbf{u}]_j|)(|[\mathbf{D}_1\mathbf{v}]_j|+|[\mathbf{D}_1\mathbf{u}]_j|)}{\sqrt{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta}+\sqrt{[\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta}} \right. \\
&\quad \left. + \frac{(|[\mathbf{D}_2\mathbf{v}]_j|-|[\mathbf{D}_2\mathbf{u}]_j|)(|[\mathbf{D}_2\mathbf{v}]_j|+|[\mathbf{D}_2\mathbf{u}]_j|)}{\sqrt{[\mathbf{D}_1\mathbf{v}]_j^2+[\mathbf{D}_2\mathbf{v}]_j^2+\beta}+\sqrt{[\mathbf{D}_1\mathbf{u}]_j^2+[\mathbf{D}_2\mathbf{u}]_j^2+\beta}} \right| \\
&\leq \frac{1}{\sqrt{\beta}}(\|\mathbf{D}_1(\mathbf{u}-\mathbf{v})\|_1+\|\mathbf{D}_2(\mathbf{u}-\mathbf{v})\|_1).
\end{aligned}$$

It follows from the equivalence of norms on $\mathbb{R}^n$ that there exists some constant $C>0$ such that

$$\|(\psi'(\mathbf{u})-\psi'(\mathbf{v}))\mathbf{D}_1\mathbf{v}\|_2 \leq \frac{C}{\sqrt{\beta}}(\|\mathbf{D}_1(\mathbf{u}-\mathbf{v})\|_2+\|\mathbf{D}_2(\mathbf{u}-\mathbf{v})\|_2).$$

Therefore, inequality (3.31) implies that

$$\begin{aligned}
\|\nabla J(\mathbf{u}) - \nabla J(\mathbf{v})\|_2 &\leq \|\mathbf{D}_1^T\|_2 \left(\|\psi'(\mathbf{u})\mathbf{D}_1(\mathbf{u}-\mathbf{v})\|_2 + \frac{C}{\sqrt{\beta}}(\|\mathbf{D}_1(\mathbf{u}-\mathbf{v})\|_2 + \|\mathbf{D}_2(\mathbf{u}-\mathbf{v})\|_2) \right) \\
&\quad + \|\mathbf{D}_2^T\|_2 \left(\|\psi'(\mathbf{u})\mathbf{D}_2(\mathbf{u}-\mathbf{v})\|_2 + \frac{C}{\sqrt{\beta}}(\|\mathbf{D}_1(\mathbf{u}-\mathbf{v})\|_2 + \|\mathbf{D}_2(\mathbf{u}-\mathbf{v})\|_2) \right) \\
&\leq \|\mathbf{D}_1^T\|_2 \left(\frac{1}{2\sqrt{\beta}}\|\mathbf{D}_1\|_2 + \frac{C}{\sqrt{\beta}}(\|\mathbf{D}_1\|_2 + \|\mathbf{D}_2\|_2) \right) \|\mathbf{u}-\mathbf{v}\|_2 \\
&\quad + \|\mathbf{D}_2^T\|_2 \left(\frac{1}{2\sqrt{\beta}}\|\mathbf{D}_2\|_2 + \frac{C}{\sqrt{\beta}}(\|\mathbf{D}_1\|_2 + \|\mathbf{D}_2\|_2) \right) \|\mathbf{u}-\mathbf{v}\|_2, \qquad (3.32)
\end{aligned}$$

which establishes the Lipschitz continuity of ∇J. Hence the optimization method of Chapter 2 is convergent in the total variation case.

Chapter 4

Regularization Parameter Selection Methods

The chapter contains material from [8, 10].

Suppose that given a realization $\mathbf{z}$ of an independent Poisson random vector $\mathbf{Z}$ with parameter vector $\mathbf{A}\mathbf{u}_e + \gamma$, where $\mathbf{u}_e \in \Omega \stackrel{\text{def}}{=} \{\mathbf{u} \in \mathbb{R}^n \mid \mathbf{u} \geq \mathbf{0}\}$, $\mathbf{A} \in \mathbb{R}^{n\times n}$ is an ill-conditioned matrix such that $\mathbf{A}\mathbf{u} \in \Omega$ when $\mathbf{u} \in \Omega$, and $\gamma \in \Omega$, a problem of interest is to estimate $\mathbf{u}_e$. Such an estimate may be obtained by solving

$$\mathbf{u}_\alpha = \text{argmin}_{\mathbf{u}\in\Omega} \left\{ T_\alpha(\mathbf{u}) \stackrel{\text{def}}{=} T_0(\mathbf{u};\mathbf{z}) + \frac{\alpha}{2}\mathbf{u}^T\mathbf{C}\mathbf{u} \right\}. \tag{4.1}$$

where $\alpha > 0$ is the regularization parameter, $\mathbf{C}$ is the regularization matrix, and T_0 is the negative log of the Poisson likelihood function, given by

$$T_0(\mathbf{u};\mathbf{z}) = \sum_{i=1}^{n} [\mathbf{A}\mathbf{u}]_i + \gamma_i - z_i \ln([\mathbf{A}\mathbf{u}]_i + \gamma_i). \tag{4.2}$$

In Chapter 3, different forms of $\mathbf{C}$ are discussed, and in Chapter 2, an algorithm is presented that can

be used to compute $\mathbf{u}_\alpha$. In this chapter, methods for selecting the value of α are discussed. Existing methods for selecting α in the context of least squares estimation will be extended to the case of Poisson likelihood estimation using a quadratic approximation of T_0. The quadratic approximation will be derived and then three methods, the discrepancy principle, generalized cross validation, and unbiased predictive risk estimation, will be extended to selecting a value of α in (4.1).

4.1 Quadratic Approximation of T_0

Many methods exist for selecting a value of α when the fit-to-data functional is a sum of squares of the difference between the data and the prediction. These methods cannot be applied directly to solving (4.1). However a Taylor series argument can be made to obtain a quadratic approximation of T_0 and that can be used to extend existing parameter selection methods to (4.1).

Computing a Taylor series expansion of T_0 requires that various derivatives of T_0 be computed. The gradient and Hessian of T_0 with respect to $\mathbf{u}$ are given by

$$\nabla_{\mathbf{u}} T_0(\mathbf{u};\mathbf{z}) = \mathbf{A}^T \left(\frac{\mathbf{A}\mathbf{u} - (\mathbf{z} - \gamma)}{\mathbf{A}\mathbf{u} + \gamma} \right), \tag{4.3}$$

$$\nabla^2_{\mathbf{uu}} T_0(\mathbf{u};\mathbf{z}) = \mathbf{A}^T \operatorname{diag}\left(\frac{\mathbf{z}}{(\mathbf{A}\mathbf{u} + \gamma)^2} \right) \mathbf{A}, \tag{4.4}$$

where division and the square are understood to be computed component-wise and $\operatorname{diag}(\mathbf{v})$ denotes the diagonal matrix with diagonal given by $\mathbf{v}$. The gradient and Hessian of T_0 with respect to $\mathbf{z}$ are given by

$$\nabla_{\mathbf{z}} T_0(\mathbf{u};\mathbf{z}) = -\ln(\mathbf{A}\mathbf{u} + \gamma), \tag{4.5}$$

$$\nabla^2_{\mathbf{zz}} T_0(\mathbf{u};\mathbf{z}) = \mathbf{0}. \tag{4.6}$$

The mixed partials of T_0 are given by

$$\nabla^2_{\mathbf{uz}} = -\mathbf{A}^T \text{diag}\left(\frac{\mathbf{1}}{(\mathbf{Au}+\gamma)^2}\right), \tag{4.7}$$

$$\nabla^2_{\mathbf{zu}} = -\text{diag}\left(\frac{\mathbf{1}}{(\mathbf{Au}+\gamma)^2}\right)\mathbf{A}. \tag{4.8}$$

$$\tag{4.9}$$

Define the background-shifted exact data to be $\mathbf{z}_e = \mathbf{Au}_e + \gamma$. Letting $\mathbf{h} = \mathbf{u} - \mathbf{u}_e$ and $\mathbf{k} = \mathbf{z} - \mathbf{z}_e$ and by constructing a Taylor series expansion of $T_0(\mathbf{u};\mathbf{z})$ about $(\mathbf{u}_e;\mathbf{z}_e)$, it follows from (4.3)-(4.8) that

$$\begin{aligned}
T_0(\mathbf{u};\mathbf{z}) &= T_0(\mathbf{u}_e+\mathbf{h};\mathbf{z}_e+\mathbf{k}), \\
&= T_0(\mathbf{u}_e;\mathbf{z}_e) + \mathbf{k}^T\nabla_{\mathbf{z}}T_0(\mathbf{u}_e;\mathbf{z}_e) + \frac{1}{2}\mathbf{h}^T\nabla^2_{\mathbf{uu}}T_0(\mathbf{u}_e;\mathbf{z}_e)\mathbf{h} \\
&\quad + \frac{1}{2}\mathbf{h}^T\nabla^2_{\mathbf{uz}}T_0(\mathbf{u}_e;\mathbf{z}_e)\mathbf{k} + \frac{1}{2}\mathbf{k}^T\nabla^2_{\mathbf{zu}}T_0(\mathbf{u}_e;\mathbf{z}_e)\mathbf{h} \qquad (4.10) \\
&\quad + \mathscr{O}(\|\mathbf{h}\|_2^3, \|\mathbf{h}\|_2^2\|\mathbf{k}\|_2, \|\mathbf{h}\|_2\|\mathbf{k}\|_2^2, \|\mathbf{k}\|_2^3) \\
&= \sum_{i=1}^{n}\{[\mathbf{Au}_e]_i - [\mathbf{z}_e]_i \ln[\mathbf{Au}_e]_i\} - (\mathbf{z}-\mathbf{z}_e)^T\ln(\mathbf{Au}_e) \\
&\quad + \frac{1}{2}(\mathbf{Au}-\mathbf{Au}_e)^T\text{diag}\left(\frac{\mathbf{1}}{\mathbf{Au}_e+\gamma}\right)(\mathbf{Au}-\mathbf{Au}_e) \\
&\quad - \frac{1}{2}(\mathbf{z}-\mathbf{z}_e)^T\text{diag}\left(\frac{\mathbf{1}}{\mathbf{Au}_e+\gamma}\right)(\mathbf{Au}-\mathbf{Au}_e) \\
&\quad - \frac{1}{2}(\mathbf{Au}-\mathbf{Au}_e)^T\text{diag}\left(\frac{\mathbf{1}}{\mathbf{Au}_e+\gamma}\right)(\mathbf{z}-\mathbf{z}_e) \\
&\quad + \mathscr{O}(\|\mathbf{h}\|_2^3, \|\mathbf{h}\|_2^2\|\mathbf{k}\|_2, \|\mathbf{h}\|_2\|\mathbf{k}\|_2^2, \|\mathbf{k}\|_2^3) \qquad (4.11) \\
&= T_0(\mathbf{u}_e;\mathbf{z}) + \frac{1}{2}(\mathbf{Au}-(\mathbf{z}-\gamma))^T\text{diag}\left(\frac{\mathbf{1}}{\mathbf{z}_e}\right)(\mathbf{Au}-(\mathbf{z}-\gamma)) \\
&\quad + \mathscr{O}(\|\mathbf{h}\|_2^3, \|\mathbf{h}\|_2^2\|\mathbf{k}\|_2, \|\mathbf{h}\|_2\|\mathbf{k}\|_2^2, \|\mathbf{k}\|_2^2). \qquad (4.12)
\end{aligned}$$

The equality in (4.12) is obtained from (4.11) by adding and subtracting the term $(\mathbf{z}-\mathbf{z}_e)^T\frac{\mathbf{1}}{\mathbf{z}_e}(\mathbf{z}-\mathbf{z}_e)$.

Therefore the second order Taylor series expansion of $T_0(\mathbf{u};\mathbf{z})$ around the point $(\mathbf{u}_e;\mathbf{z}_e)$ consists of a sum of a term that is constant with respect to $\mathbf{u}$ and a quadratic term:

$$\frac{1}{2}\mathbf{r}^T \operatorname{diag}\left(\frac{\mathbf{1}}{\mathbf{z}_e}\right)\mathbf{r}, \tag{4.13}$$

where $\mathbf{r} = \mathbf{A}\mathbf{u} - (\mathbf{z} - \gamma)$.

Because in practice $\mathbf{u}_e$ is unknown, the quadratic approximation given in (4.13) cannot be used directly. This motivates an application of the mean value theorem. First, for a fixed $\alpha > 0$, define $\mathbf{z}_\alpha = \mathbf{A}\mathbf{u}_\alpha + \gamma$, where $\mathbf{u}_\alpha$ is computed from (4.1), and let $\mathbf{k}_\alpha = \mathbf{z}_\alpha - \mathbf{z}_e$. Then considering the ith component of $\frac{\mathbf{1}}{\mathbf{z}_e}$ to be a function of $[\mathbf{k}_\alpha]_i$ and noting that $\mathbf{A}\mathbf{u} + \gamma > \mathbf{0}$, the mean value theorem can be used to rewrite (4.13) in the following manner:

$$\begin{aligned}\frac{1}{2}\mathbf{r}^T\left(\mathbf{r}\odot\frac{\mathbf{1}}{\mathbf{z}_e}\right) &= \frac{1}{2}\mathbf{r}^T\left(\mathbf{r}\odot\frac{\mathbf{1}}{\mathbf{z}_\alpha - \mathbf{k}_\alpha}\right)\\ &= \frac{1}{2}\mathbf{r}^T\left(\mathbf{r}\odot\left(\frac{\mathbf{1}}{\mathbf{z}_\alpha} + \operatorname{diag}\left(\frac{\mathbf{1}}{(\mathbf{A}\mathbf{u}+\gamma-\hat{\mathbf{k}_\alpha})^2}\right)\mathbf{k}_\alpha\right)\right),\end{aligned} \tag{4.14}$$

where $[\hat{\mathbf{k}_\alpha}]_i$ lies in the interval with endpoints 0 and $[\mathbf{k}_\alpha]_i$. Since $\mathbf{r} = \mathbf{A}\mathbf{h} - \mathbf{k}$ it is the case that

$$\mathbf{r}^T \operatorname{diag}\left(\frac{\mathbf{r}}{(\mathbf{A}\mathbf{u}_\alpha + \gamma - \hat{\mathbf{k}_\alpha})^2}\right)\mathbf{l} = \mathscr{O}(\|\mathbf{h}\|_2^2\|\mathbf{k}_\alpha\|_2, \|\mathbf{h}\|_2\|\mathbf{k}\|_2\|\mathbf{k}_\alpha\|_2, \|\mathbf{k}\|_2^2\|\mathbf{k}_\alpha\|_2). \tag{4.15}$$

Thus (4.12), (4.14), and (4.15) yield the following approximation

$$\begin{aligned}T_0(\mathbf{u};\mathbf{z}) &= T_0(\mathbf{u}_e;\mathbf{z}) + T_0^{\mathrm{WLS}}(\mathbf{u};\mathbf{z})\\ &\quad + \mathscr{O}(\|\mathbf{h}\|_2^3, \|\mathbf{h}\|_2^2\|\mathbf{k}\|_2, \|\mathbf{h}\|_2\|\mathbf{k}\|_2^2, \|\mathbf{k}\|_2^2)\\ &\quad + \mathscr{O}(\|\mathbf{h}\|_2^2\|\mathbf{k}_\alpha\|_2, \|\mathbf{h}\|_2\|\mathbf{k}\|_2\|\mathbf{k}_\alpha\|_2, \|\mathbf{k}\|_2^2\|\mathbf{k}_\alpha\|_2).\end{aligned} \tag{4.16}$$

where

$$T_0^{\mathrm{WLS}}(\mathbf{u};\mathbf{z}) = \frac{1}{2}\left\|\mathbf{Z}_\alpha^{-1/2}(\mathbf{A}\mathbf{u} - (\mathbf{z} - \gamma))\right\|_2^2, \tag{4.17}$$

and $\mathbf{Z}_\alpha = \operatorname{diag}(\mathbf{z}_\alpha)$. We use approximation (4.16 to motivate three regularization parameter selection

methods.

4.2 Regularization Parameter Selection Methods

4.2.1 The Discrepancy Principle Method

The discrepancy principle (DP) is motivated by the idea that if $\mathbf{u}_\alpha$ is close to $\mathbf{u}_e$ then the expected discrepancy between $\mathbf{u}_\alpha$ and the data, $E(T_0(\mathbf{u}_\alpha;\mathbf{Z}))$, should be approximately equal to the expected discrepancy between $\mathbf{u}_e$ and the data, $E(T_0(\mathbf{u}_e;\mathbf{Z}))$. Note that here $\mathbf{Z}$ is the random vector of which the data $\mathbf{z}$ is a realization and so $T_0(\mathbf{u};\mathbf{Z})$ is a random variable. Since

$$E(T_0(\mathbf{u}_e;\mathbf{Z})) = \sum_{i=1}^{k}([\mathbf{A}\mathbf{u}_e]_i + \gamma_i) - ([\mathbf{A}\mathbf{u}_e]_i + \gamma_i)\ln([\mathbf{A}\mathbf{u}_e]_i + \gamma_i),$$

it cannot be evaluated directly because $\mathbf{u}_e$ is unknown in practice. The quadratic approximation (4.16) along with another approximation is used to compute an estimate of $E(T_0(\mathbf{u}_e;\mathbf{Z}))$ that does not depend on $\mathbf{u}_e$. Note that from (4.16) and (4.17) it follows that

$$E(T_0(\mathbf{u};\mathbf{Z})) \approx T_0(\mathbf{u}_e;\mathbf{z}_e) + E(T_0^{\mathrm{WLS}}(\mathbf{u};\mathbf{Z})), \tag{4.18}$$

and so if $\mathbf{u}_\alpha \approx \mathbf{u}_e$ then

$$E(T_0^{\mathrm{WLS}}(\mathbf{u}_\alpha;\mathbf{Z})) \approx E(T_0^{\mathrm{WLS}}(\mathbf{u}_e;\mathbf{Z})). \tag{4.19}$$

An estimate of $E(T_0^{\mathrm{WLS}}(\mathbf{u}_e;\mathbf{Z}))$ must now be obtained. In order to do this, the approximation

$$\mathbf{Z} - \gamma = \mathbf{A}\mathbf{u}_e + \eta, \tag{4.20}$$

where η is a Gaussian random vector with mean $\mathbf{0}$ and covariance matrix given by $\mathrm{diag}(\mathbf{A}\mathbf{u}_e + \gamma)$, is

made. Let

$$\mathbf{r}(\mathbf{u};\mathbf{Z}) \stackrel{\text{def}}{=} \text{diag}(\mathbf{A}\mathbf{u}+\gamma)^{-1/2}(\mathbf{A}\mathbf{u}-(\mathbf{Z}-\gamma)), \tag{4.21}$$

and note that that $\mathbf{r}(\mathbf{u}_e;\mathbf{Z})$ is a Gaussian random vector with mean $\mathbf{0}$ and variance $\mathbf{I}_n$, where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. It follows from a standard statistical result that the random variable $\|\mathbf{r}(\mathbf{u}_e;\mathbf{Z})\|_2^2$ has a chi-squared distribution with n degrees of freedom, for which the mean is n and the variance $2n$. Thus the discrepancy principle indicates that acceptable values of α are those that yield values of $\|\mathbf{r}(\mathbf{u}_\alpha;\mathbf{z})\|^2$ that are likely to be realizations of a $\chi^2(n)$ random variable and so the following criterion for acceptable values of α can be formulated: a value of α is appropriate if $\mathbf{r}(\mathbf{u}_\alpha;\mathbf{z})$ is within two standard deviations of n; that is if

$$n - 2\sqrt{2n} \leq 2T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}) \leq n + 2\sqrt{2n}, \tag{4.22}$$

where $\mathbf{u}_\alpha$ is computed from (4.1). A specific value of α can be chosen by approximately solving the nonlinear equation

$$2T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}) = n. \tag{4.23}$$

4.2.2 The Generalized Cross Validation Method

The regularization parameter selection method of leave-one-out cross validation selects the value of α that minimizes

$$\text{CV}(\alpha) = \frac{1}{n}\sum_{k=1}^{n}([\mathbf{A}\mathbf{u}_\alpha^k]_k + \gamma) - z_k \ln([\mathbf{A}\mathbf{u}_\alpha^k]_k + \gamma), \tag{4.24}$$

where

$$\mathbf{u}_\alpha^k = \text{argmin}_{\mathbf{u}\in\Omega}\left\{\sum_{i\neq k}([\mathbf{A}\mathbf{u}]_i + \gamma_i) - z_i \ln([\mathbf{A}\mathbf{u}]_i + \gamma_i) + \frac{\alpha}{2}\mathbf{u}^T\mathbf{C}\mathbf{u}\right\}. \tag{4.25}$$

For large-scale problems minimizing (4.24) is not feasible. In the method of generalized cross validation (GCV) for regularized least squares the minimizer is found of an approximation, referred to

as the GCV function, of CV(α) for which optimization is much more efficient. The GCV method is extended to (4.1) by making use of the quadratic approximation of T_0 given in (4.16).

In order to derive the GCV function some preliminary definitions and arguments are needed. The regularization operator $\mathbf{A}_\alpha$ is commonly defined to be the operator for which $\mathbf{u}_\alpha = \mathbf{A}_\alpha \mathbf{z}$. The derivation of the GCV function for regularized least squares problems makes use of the fact that the regularization operator in those problems is linear. For (4.1) however, $\mathbf{A}_\alpha$ is nonlinear and hence a linear approximation $\mathbf{A}_\alpha$ satisfying $\mathbf{u}_\alpha \approx \mathbf{A}_\alpha \mathbf{Z}_\alpha^{-1/2}(\mathbf{z}-\gamma)$ is needed. To obtain such an approximation, first define the matrix $\mathbf{D}_\alpha$ to be the diagonal matrix with nonzero entries given by $[\mathbf{D}_\alpha]_{ii} = 0$ if $[\mathbf{u}_\alpha]_i = 0$ and $[\mathbf{D}_\alpha]_{ii} = 1$ if $[\mathbf{u}_\alpha]_i > 0$ and note that since T_α is strictly convex, $\mathbf{u}_\alpha$ is the solution of the equation

$$\mathbf{D}_\alpha \nabla T_\alpha(\mathbf{D}_\alpha \mathbf{u}) = \mathbf{0} \tag{4.26}$$

which has minimum norm [28]. After applying approximation (4.16), equation (4.26) becomes

$$\mathbf{D}_\alpha \mathbf{A}^T \mathbf{Z}_\alpha^{-1}(\mathbf{A}\mathbf{D}_\alpha \mathbf{u} - (\mathbf{z}-\gamma)) + \alpha \mathbf{D}_\alpha \mathbf{C}\mathbf{D}_\alpha \mathbf{u} = \mathbf{0}, \tag{4.27}$$

for which the minimum-norm solution is

$$(\mathbf{D}_\alpha(\mathbf{A}^T \mathbf{Z}_\alpha^{-1}\mathbf{A} + \alpha \mathbf{C})\mathbf{D}_\alpha)^\dagger \mathbf{D}_\alpha \mathbf{A}^T \mathbf{Z}_\alpha^{-1}(\mathbf{z}-\gamma). \tag{4.28}$$

This motivates the following expression for $\mathbf{A}_\alpha$:

$$\mathbf{A}_\alpha = (\mathbf{D}_\alpha(\mathbf{A}^T \mathbf{Z}_\alpha^{-1}\mathbf{A} + \alpha \mathbf{C})\mathbf{D}_\alpha)^\dagger \mathbf{D}_\alpha \mathbf{A}^T \mathbf{Z}_\alpha^{-1/2}. \tag{4.29}$$

Now let $\mathbf{z}^k$ be defined by

$$[\mathbf{z}^k]_i = \begin{cases} z_i & i \neq k, \\ \left[\mathbf{A}\mathbf{u}_\alpha^k\right]_k + \gamma_k & i = k, \end{cases} \tag{4.30}$$

and note that $\mathbf{u}_\alpha^k \approx \mathbf{A}_\alpha \mathbf{Z}_\alpha^{-1/2}(\mathbf{z}^k - \gamma)$. Then

$$\begin{aligned}
\frac{[\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_\alpha]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_\alpha^k]_k}{[\mathbf{Z}_\alpha^{-1/2}\mathbf{z}]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{z}^k]_k} &= \frac{[\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha\mathbf{Z}_\alpha^{-1/2}(\mathbf{z}-\gamma)]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha\mathbf{Z}_\alpha^{-1/2}(\mathbf{z}^k-\gamma)]_k}{[\mathbf{Z}_\alpha^{-1/2}\mathbf{z}]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{z}^k]_k} \\
&= \frac{\sum_{i=1}^n([\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha]_{k,i}[\mathbf{Z}_\alpha^{-1/2}(\mathbf{z}-\gamma)]_i - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha]_{k,i}[\mathbf{Z}_\alpha^{-1/2}(\mathbf{z}^k-\gamma)]_i)}{[\mathbf{Z}_\alpha^{-1/2}\mathbf{z}]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{z}^k]_k} \\
&= [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha]_{k,k},
\end{aligned}$$

and, since

$$\begin{aligned}
1 - \frac{[\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_\alpha]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_\alpha^k]_k}{[\mathbf{Z}_\alpha^{-1/2}\mathbf{z}]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{z}^k]_k} &= \frac{[\mathbf{Z}_\alpha^{-1/2}(\mathbf{z}-\gamma)]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_\alpha]_k}{[\mathbf{Z}_\alpha^{-1/2}\mathbf{z}]_k - [\mathbf{Z}_\alpha^{-1/2}\mathbf{z}^k]_k} \\
&= \frac{[\mathbf{r}(\mathbf{u}_\alpha;\mathbf{z})]_k}{[\mathbf{r}(\mathbf{u}_\alpha^k;\mathbf{z})]_k},
\end{aligned}$$

it is the case that

$$[\mathbf{r}(\mathbf{u}_\alpha^k;\mathbf{z})]_k = \frac{[\mathbf{r}(\mathbf{u}_\alpha;\mathbf{z})]_k}{1 - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha]_{k,k}}. \tag{4.31}$$

Now (4.16) can be used to rewrite (4.24) as

$$\begin{aligned}
\mathrm{CV}(\alpha) &\approx \frac{1}{n}T_0(\mathbf{u}_e;\mathbf{z}) + \frac{1}{2n}\sum_{k=1}^n [\mathbf{r}(\mathbf{u}_\alpha^k;\mathbf{z})]_k^2 \\
&= \frac{1}{n}T_0(\mathbf{u}_e;\mathbf{z}) + \frac{1}{2n}\sum_{k=1}^n \frac{([\mathbf{r}(\mathbf{u}_\alpha;\mathbf{z})]_k)^2}{(1 - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha]_{k,k})^2},
\end{aligned} \tag{4.32}$$

where (4.32) is a result of (4.31). The computation of (4.32) can still be prohibitively expensive for large-scale problems and so the approximation

$$1 - [\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha]_{k,k} \approx \frac{1}{n}\mathrm{trace}(\mathbf{I}_n - \mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha) \tag{4.33}$$

is used. Therefore the GCV approximation for (4.1) is given by

$$\mathrm{GCV}(\alpha) = \frac{nT_0^{\mathrm{WLS}}(\mathbf{u}_\alpha)}{\mathrm{trace}(\mathbf{I}_n - \mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha)^2}. \tag{4.34}$$

The fact that $\mathbf{u}_\alpha$, and hence $\mathbf{A}_\alpha$ and $\mathbf{Z}_\alpha$, is computed directly from (4.1) is a key difference between (4.34) and the GCV function used for regularized least squares problems; otherwise the forms of the two approximations are very similar. The GCV method selects the value of α that solves

$$\alpha_{\text{GCV}} = \text{argmin}_{\alpha>0}\text{GCV}(\alpha). \tag{4.35}$$

Randomized Trace Estimation The presence of $\mathbf{D}_\alpha$ and $\mathbf{Z}_\alpha$ in the expression for $\mathbf{A}_\alpha$ as well as the size of $\mathbf{A}_\alpha$ can cause the exact evaluation of the term $\text{trace}(\mathbf{I}_n - \mathbf{A}\mathbf{A}_\alpha)$ to be impractical. Instead an estimate can be used that is computationally cheaper to evaluate. Note that if $\mathbf{W}$ is an $n \times 1$ random vector with mean $\mathbf{0}$ and covariance $\mathbf{I}_n$, then for $\mathbf{B} \in \mathbb{R}^{n\times n}$,

$$\begin{aligned} E(\mathbf{W}^T\mathbf{B}\mathbf{W}) &= E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} W_i W_j \mathbf{B}_{i,j}\right) \\ &= \sum_{i=1}^{n} \mathbf{B}_{i,i} E(W_i^2) \\ &= \text{trace}(\mathbf{B}). \end{aligned}$$

Thus $\mathbf{W}^T\mathbf{B}\mathbf{W}$ is an unbiased estimator of $\text{trace}(\mathbf{B})$. This fact motivates estimating $\text{trace}(\mathbf{I}_n - \mathbf{A}\mathbf{A}_\alpha)$ by taking a realization $\mathbf{w}$ of $\mathbf{W}$ and evaluating $t(\alpha) = \mathbf{w}^T(\mathbf{I}_n - \mathbf{A}\mathbf{A}_\alpha)\mathbf{w}$. The variance of $t(\alpha)$ is minimized when $\mathbf{W}$ is an independent random vector whose components take on the values of 1 and -1 each with probability 0.5 [34].

4.2.3 The Unbiased Predictive Risk Estimator Method

The Unbiased Predictive Risk Estimator (UPRE) method entails the minimization of an estimator of the expected value of the *predictive risk* $T_0(\mathbf{u}_\alpha; \mathbf{z}_e)$. The predictive risk cannot be evaluated because $\mathbf{z}_e$ is unknown.

As in the cases of the DP and GCV methods, the UPRE method for choosing α in regularized least

squares problems is extended to (4.1) by employing the quadratic approximation of T_0 given in (4.16). The first step is to use (4.16) to write the predictive risk as

$$T_0(\mathbf{u}_\alpha;\mathbf{z}_e) \approx T_0(\mathbf{u}_e;\mathbf{z}_e) + T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}_e). \tag{4.36}$$

It is also the case that

$$T_0(\mathbf{u}_\alpha;\mathbf{z}) \approx T_0(\mathbf{u}_e;\mathbf{z}_e) + T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}_e). \tag{4.37}$$

The expected value of the predictive risk and $E(T_0(\mathbf{u};\mathbf{z}))$ are then given by

$$E(T_0(\mathbf{u}_\alpha;\mathbf{z}_e)) \approx T_0(\mathbf{u}_e;\mathbf{z}_e) + E(T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}_e)), \tag{4.38}$$

$$E(T_0(\mathbf{u}_\alpha;\mathbf{z})) \approx T_0(\mathbf{u}_e;\mathbf{z}_e) + E(T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}_e)). \tag{4.39}$$

Arguments mimicking the reasoning found in [34, Section 7.1] can now be made. The following lemma is needed.

Lemma 4.2.1. *Let* $\mathbf{u} \in \mathbb{R}^n$, $\mathbf{B} \in \mathbb{R}^{n\times n}$, *and* η *be an* $n \times 1$ *random vector with mean* $\mathbf{0}$ *and covariance* $\sigma^2\mathbf{I}_n$. *Then*

$$E(\|\mathbf{u}+\mathbf{B}\eta\|^2) = \|\mathbf{u}\|^2 + \sigma^2\text{trace}(\mathbf{B}^T\mathbf{B}). \tag{4.40}$$

Proof.

$$\begin{aligned} E(\|\mathbf{u}+\mathbf{B}\eta\|^2) &= E(\mathbf{u}^T\mathbf{u}) + 2E(\mathbf{u}^T\mathbf{B}\eta) + E[(\mathbf{B}\eta)^T\mathbf{B}\eta] \\ &= \|\mathbf{u}\|^2 + 2E[(\mathbf{B}^T\mathbf{u})^T\eta] + E(\eta^T\mathbf{B}^T\mathbf{B}\eta) \\ &= \|\mathbf{u}\|_2^2 + \sum_{j=1}^{n}[\mathbf{B}^T\mathbf{u}]_j E(\nu_j) + \sum_{i=1}^{n}\sum_{j=1}^{n}[\mathbf{B}^T\mathbf{B}]_{i,j}E(\nu_i\nu_j). \end{aligned}$$

The result follows from the properties of η. □

Let $\mathbf{p}_\alpha = \mathbf{Z}_\alpha^{-1/2}(\mathbf{A}\mathbf{u}_\alpha - (\mathbf{z}_e - \gamma))$ and as in the derivation of the DP method approximate $\mathbf{Z}$ with (4.20).

Then

$$\mathbf{p}_\alpha = (\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha - \mathbf{I}_n)\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_e + \mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha\mathbf{Z}_\alpha^{-1/2}\eta.$$

If $\mathbf{u}_\alpha \approx \mathbf{u}_e$ then $\mathbf{Z}_\alpha^{-1/2} \approx \text{diag}(\mathbf{A}\mathbf{u}_e + \gamma)^{-1/2}$ and so the approximation

$$\text{cov}(\mathbf{Z}_\alpha^{-1/2}\eta) = \mathbf{I}_n \tag{4.41}$$

makes sense. It follows from (4.41) and Lemma 4.2.1 that

$$\begin{aligned} E(T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}_e)) &= \frac{1}{2}E(\|\mathbf{p}_\alpha\|^2) \\ &= \frac{1}{2}\|(\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha - \mathbf{I}_n)\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_e\|^2 + \frac{1}{2}\text{trace}((\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha)^2). \end{aligned} \tag{4.42}$$

It is also the case that

$$\mathbf{r}(\mathbf{u}_\alpha;\mathbf{z}) = (\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha - \mathbf{I}_n)\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_e + (\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha - \mathbf{I}_n)\mathbf{Z}_\alpha^{-1/2}\eta,$$

and again applying (4.41) and Lemma 4.2.1 yields

$$\begin{aligned} E(T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z})) &= \frac{1}{2}E(\|\mathbf{r}(\mathbf{u}_\alpha;\mathbf{z})\|^2) \\ &= \frac{1}{2}\|(\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha - \mathbf{I}_n)\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{u}_e\|^2 + \frac{1}{2}\text{trace}((\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha)^2) - \text{trace}(\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha) \\ &\quad + \frac{n}{2}. \end{aligned} \tag{4.43}$$

Thus

$$E(T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}_e)) = E(T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z})) + \text{trace}(\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha) - \frac{n}{2}, \tag{4.44}$$

and the UPRE is defined to be

$$\text{UPRE}(\alpha) = T_0^{\text{WLS}}(\mathbf{u}_\alpha;\mathbf{z}) + \text{trace}(\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha) - \frac{n}{2}. \tag{4.45}$$

Note that $\text{UPRE}(\alpha)$ is an unbiased estimator of the expected value of the predictive risk. The UPRE

method selects the value of α that solves

$$\alpha_{\text{UPRE}} = \text{argmin}_{\alpha>0}\text{UPRE}(\alpha). \tag{4.46}$$

In practice randomized trace estimation is used to estimate the term trace$(\mathbf{Z}_\alpha^{-1/2}\mathbf{A}\mathbf{A}_\alpha)$ in (4.45).

Chapter 5

Numerical Results

This chapter contains material first presented in papers [4, 8–10]. The codes used to generate these results, and which implement the methods of the previous chapters can be found at [2].

The nonnegatively constrained minimization algorithm presented in Chapter 2, regularization operators presented in Chapter 3, and the parameter selection methods presented in Chapter 4 were tested on the astronomical imaging and positron emission tomography (PET) examples that were presented in the introduction.

5.1 Astronomical Imaging Example

5.1.1 Statement of Problem

Recall that in astronomical imaging, the problem is to estimate the true image $\mathbf{u}_e$ given observations $\mathbf{z}$ for which the data-noise model is

$$\mathbf{z} = \text{Poiss}(\mathbf{A}\mathbf{u}_e + \gamma) + N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{5.1}$$

where γ gives the background intensity and σ^2 is the variance of the instrument readout noise. Here the observed image array and true image array are both $n \times n$. The forward-model matrix $\mathbf{A}$ is therefore $N \times N$, with $N = n^2$ and it arises from the discrete convolution of an image with an $n \times n$ point spread function (PSF) $\mathbf{a}$. $\mathbf{a}$ is computed using the Fourier optics PSf model [34]:

$$\mathbf{a} = \left| \mathscr{F}^{-1}\{\mathbf{p} \odot e^{\hat{\imath}\phi}\} \right|^2, \tag{5.2}$$

where $\mathscr{F}^{-1}$ denotes the two-dimensional inverse Fourier transform, $\mathbf{p}$ is the $n \times n$ indicator array for an annulus, $\odot$ signifies component-wise multiplication, $\hat{\imath} = \sqrt{-1}$, and ϕ is the $n \times n$ array that represents the distortion in the planar wavefronts of light resulting from turbulence in the atmosphere. ϕ is obtained using the Kolmogorov turbulence model [30].

In the case of periodic boundary conditions, multiplication by the matrix $\mathbf{A}$ can be carried out much more efficiently using fast fourier transforms (FFTs). In this case, multiplication by $\mathbf{A}$ can be written as

$$\mathbf{Au} = \texttt{vec}((\texttt{ifft2}(\hat{\mathbf{a}} \odot \texttt{fft2}(\mathbf{u})))) \quad \hat{\mathbf{a}} = \texttt{fft2}(\texttt{fftshift}(\mathbf{a})),$$

where `vec` column stacks $n \times n$ arrays to obtain $n^2 \times 1$ vectors, `fft2` and `ifft2` are the discrete two-dimensional Fourier and inverse Fourier transforms respectively, and `fftshift` swaps the first and third and second and fourth quadrants of the array $\mathbf{a}$. Multiplication by $\mathbf{A}$ can still be carried out using FFT's when zero boundary conditions are assumed, though a slightly different formulation results [34].

The computational framework was tested using a standard synthetic data example developed at the US Air Force Phillips Laboratory, Lasers and Imaging Directorate, Kirtland Air Force Base, New Mexico. The image is a simulation of a satellite as seen through a ground-based telescope. The 256×256 true image is shown on the left in Figure 5.1. Additionally, star field data was simulated and used to perform numerical tests. The star field data is plotted on a log scale to aid in visualization and can be seen on the right in Figure 5.1.

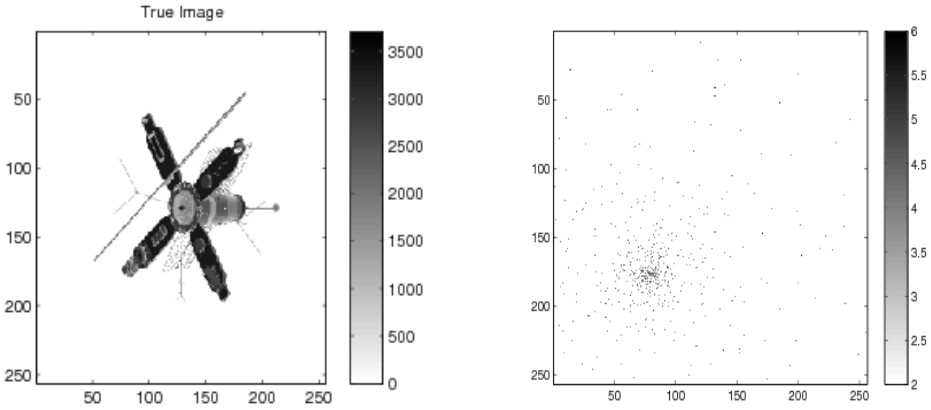

Figure 5.1: *True image of the satellite on the left and the true image of the star field, plotted on a log scale, with entries less than 100 set to 0, on the right.*

The noisy data was generated in Matlab using the `poissrnd` and `randn` functions. The values of γ and σ were chosen to be $\gamma = 10$ and $\sigma = 5$. The corresponding blurred, noisy signals are displayed in Figure 5.2. Again note that the star field data is plotted on a log scale.

In solving the inverse problem, (5.1) is approximated by

$$\mathbf{z} + \sigma^2 = \text{Poiss}(\mathbf{A}\mathbf{u}_e + \gamma + \sigma^2). \tag{5.3}$$

This is done by making the approximation $N(\sigma^2, \sigma^2) \approx \text{Poiss}(\sigma^2)$ and using independence properties. The approximation is made because the solution of the resulting maximum likelihood problem is much easier to compute.

The signal-to-noise ratio (SNR) for (5.1) is defined to be

$$SNR = \sqrt{\frac{\|\mathbf{A}\mathbf{u}_e + \gamma\|^2}{E(\|\mathbf{z} - (\mathbf{A}\mathbf{u}_e + \gamma)\|^2)}}. \tag{5.4}$$

Note that under the square root the numerator contains the noise-free signal and the denominator contains the variance of the data $\mathbf{z}$, also known as the noise power. For (5.1), $E(\|\mathbf{z} - (\mathbf{A}\mathbf{u}_e + \gamma)\|^2) = \sum_{j=1}^{N}([\mathbf{A}\mathbf{u}_e]_j + \gamma_i + \sigma_i^2)$. In order to test the methods on multiple data sets, the intensity of the true

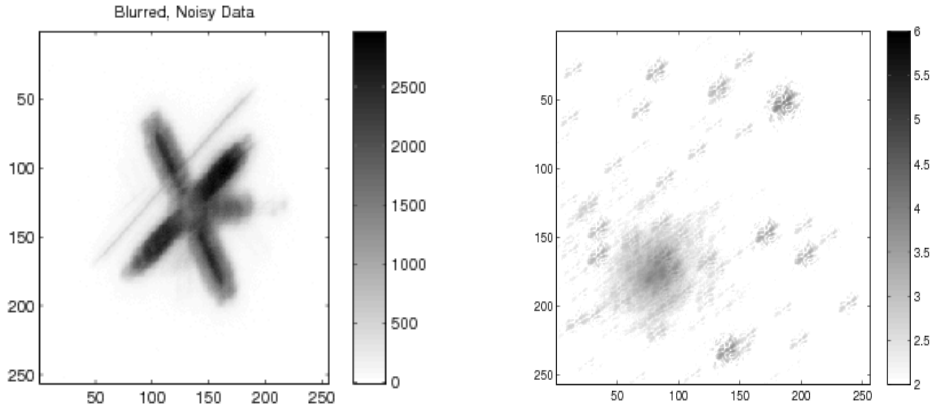

Figure 5.2: *Blurred, noisy images of the satellite on the left, and the star field, plotted on a log scale, with entries less than 100 set to 0, on the right, both with SNR = 30.*

image is varied to obtain noise at four different levels of SNR.

5.1.2 Regularization Operators

Recall from Chapter 3 that three choices for $\mathbf{C}$ are considered: $\mathbf{C} = \mathbf{I}$, $\mathbf{C} = \mathbf{L}$, where $\mathbf{L}$ is a discretization of the negative laplacian, and $\mathbf{C} = \Theta$, where

$$\Theta = \mathbf{D}_x^T \Lambda \mathbf{D}_x + \mathbf{D}_y^T \Lambda \mathbf{D}_y. \tag{5.5}$$

$\mathbf{D}_x$ and $\mathbf{D}_y$ are discretizations of the x- and y- partial derivatives respectively and Λ is a diagonal matrix with $[\Lambda]_{ii} = 1$ if the ith pixel lies in a region that is known to be smooth and $[\Lambda]_{ii} < 1$ if the ith pixel is known to be adjacent to an edge. $\mathbf{L}$ should be used in problems where the true image is known to be smooth and Θ should only be used when the true image is known to be smooth except in regions which contain sharp jumps in intensity (edges). In Chapter 3, the construction of Λ is not addressed. If the location of the edges are prior knowledge, then the construction of Λ is simple. In fact, an edge-preserving regularization matrix can be constructed which is a discretization of an anisotropic diffusion operator [21]. In the examples presented here, the locations of the edges are not assumed to be known a priori.

The construction of Λ when the locations of the edges in $\mathbf{u}_e$ are unknown involves first computing a smooth approximation $\mathbf{u}_{\text{approx}}$ of $\mathbf{u}_e$. This is done by solving (4.1) with $\mathbf{C} = \mathbf{L}$ and α chosen using one of the three selection methods that have been presented. Let

$$\mathbf{v} = [\mathbf{D}_x \mathbf{u}_{\text{approx}}]^2 + [\mathbf{D}_y \mathbf{u}_{\text{approx}}]^2, \tag{5.6}$$

and define the vector $\mathbf{v}_\varepsilon$ by

$$[\mathbf{v}_\varepsilon]_i = \begin{cases} v_i & v_i \geq \varepsilon \|\mathbf{v}\|_\infty, \\ 0 & \text{otherwise}, \end{cases} \tag{5.7}$$

where $\|\cdot\|_\infty$ denotes the ℓ-∞ norm and $0 < \varepsilon < 1$. For the experiments performed here, we chose $\varepsilon = 0.01$. Now Λ is defined by

$$\Lambda = \text{diag}\left(\max\left\{\frac{1}{1+\rho\mathbf{v}_\varepsilon}, \frac{1}{10}\right\}\right), \tag{5.8}$$

where ρ is a scaling parameter. For the experiments performed here, $\rho = 1$. Note that under the assumption that $[\mathbf{v}_\varepsilon]_i$ will be large if the ith pixel is near an edge, Λ has the effect of decreasing the regularization parameter by an order of magnitude which results in a decrease in smoothing at the ith pixel. However if $[\mathbf{v}_\varepsilon]_i = 0$ then the regularization parameter remains approximately the same and a reconstruction that is not smooth at that pixel will be penalized. For different applications, the values of ε, ρ and $\frac{1}{10}$ in (5.8) might need to be adjusted.

Constructing Λ in this manner suggests an approach in which the reconstructions are iteratively refined by using (5.5) with (5.6)-(5.8) and taking $\mathbf{u}_{\text{approx}}$ to be the result of the previous iteration. The value of α can be selected using one of the three regularization parameter selection methods presented in Chapter 4. To summarize:

Iteratively Updated Diffusion Regularization

Step 0: Set $\Lambda = \mathbf{I}$.

Step 1: Select α using either of the GCV, UPRE, or DP methods and compute the solution $\mathbf{u}_\alpha$ of (4.1) with regularization function (5.5).

Step 2: Set $\mathbf{u}_{\text{approx}} = \mathbf{u}_\alpha$ and update Λ using (5.6)-(5.8), then return to Step 1.

5.1.3 Regularization Parameter Selection Methods

The three regularization parameter selection methods presented in Chapter 4 each require the solution of a minimization problem. For GCV, the GCV function given in (4.34) must be minimized; for UPRE the function is given by (4.45). For DP, recall that an appropriate choice for α is one for which

$$\|\mathbf{Z}_\alpha^{-1/2}(\mathbf{A}\mathbf{u}_\alpha - (\mathbf{z}-\gamma))\|^2 = N.$$

Such an α might not exist, so α is selected by solving

$$\alpha_{\text{DP}} = \text{argmin}_{\alpha>0}\{(\|\mathbf{Z}_\alpha^{-1/2}(\mathbf{A}\mathbf{u}_\alpha - (\mathbf{z}-\gamma)\|^2 - N)^2\}. \tag{5.9}$$

In all three cases the minimization problem has the constraint that $\alpha > 0$. Matlab's `fminbnd` function was used to solve the minimization problems. `fminbnd` requires that upper and lower bounds for the minimizer be included in the input. A lower bound of 0 and an upper bound of 0.01 were used in each case. The `TolX` parameter was set to 10^{-8} for the satellite example and 10^{-10} for the star field example. It should also be noted that when DP is implemented, the minimum value of the right-hand side of (5.9) achieved by `fminbnd` was on the order of 10^{-8} in all cases except one in which the mininum value was on the order of 10^{-6}.

The evaluation of the function to be minimized in each of the three methods requires the computation of $\mathbf{u}_\alpha$, where $\mathbf{u}_\alpha$ is the solution of (4.1) which is itself a minimization problem. This problem is solved by applying the GPRN algorithm which is presented in Chapter 2. The maximum number of iterations

was set to 50, and the tolerances on the step size and the size of the norm of the projected gradient were both set to 10^{-6}. In each iteration of the GPRN algorithm, a maximum of 5 gradient projection iterations and 40 conjugate gradient iterations were allowed. The stopping tolerances for both the gradient projection step and the reduced Newton step were both set to 0.1.

Experiments were performed on the star field and satellite data using $\mathbf{C} = \mathbf{I}$ and with SNR = 5, 10, 30, and 100. Using $\mathbf{C} = \mathbf{L}$ and $\mathbf{C} = \Theta$ experiments were performed on the satellite data (the star field data was excluded because the true image is known to not be smooth) with SNR = 10 and 30. In each case the relative error, given by

$$\frac{\|\mathbf{u}_\alpha - \mathbf{u}_e\|}{\|\mathbf{u}_e\|},$$

is calculated for a range of values of α and the values that are recommended by the various methods.

The results of the satellite test case with $\mathbf{C} = \mathbf{I}_n$ are displayed in Figure 5.3. All three selection methods gave good recommendations, with GCV and UPRE recommending a value for α that is slightly worse than that recommended by DP. Note that as α tends to 0, $\mathbf{u}_\alpha$ tends towards the solution of the unregularized maximum likelihood problem which is not close to the true image because that problem is ill-posed. When α is too large the penalty term dominates the minimization problem and the contribution from the data model is underemphasized, yielding a reconstruction which is also far from the true image.

The results from the star field test case are displayed in Figure 5.4. Again the three methods yielded good recommendations for α with the caveat that the methods worked better when the SNR was 30 or 100 than when the SNR was 5 or 10. In this example UPRE gave the best recommendation. The plots indicate that the optimal value of α is very small and hence regularization is not needed. One explanation for this is that the nonnegativity constraints and the point-source nature of the true image have a stabilizing effect on the inverse problem [20].

The results obtained from using $\mathbf{C} = \mathbf{L}$ with the satellite data are displayed in the top row in Figure 5.5. The three methods each gave a good recommendation for α, with UPRE giving the best recommenda-

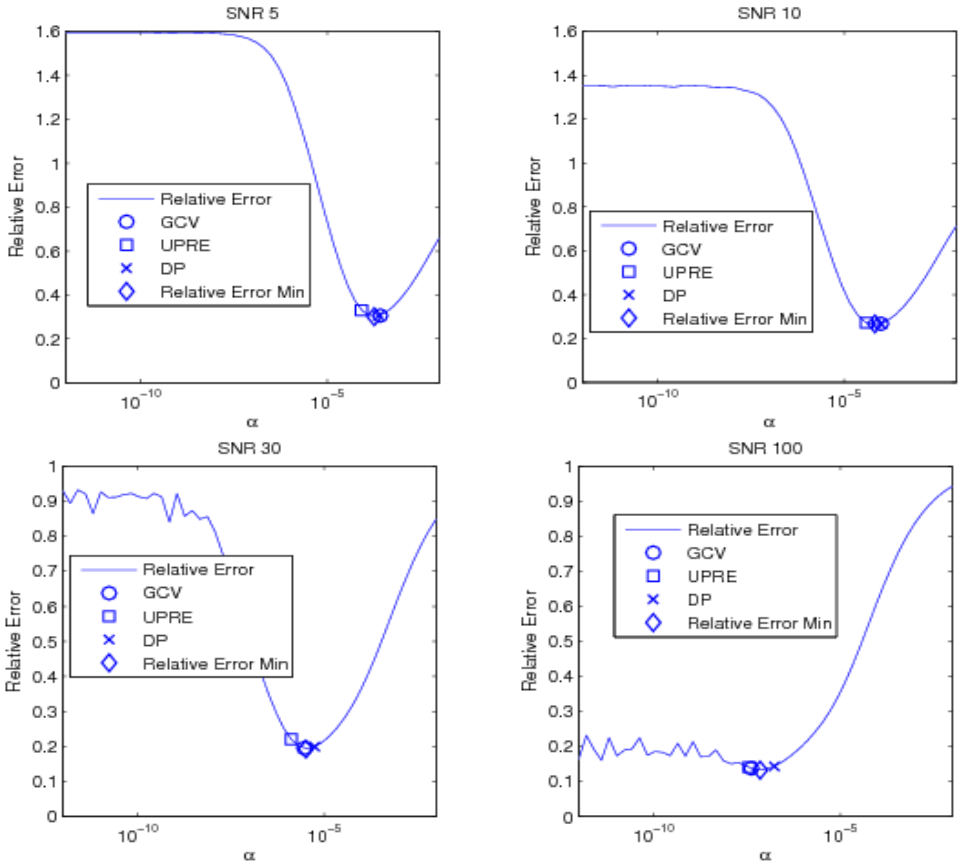

Figure 5.3: *Satellite Test Case with* $\mathbf{C} = \mathbf{I}_n$. *Plot of relative error together with the values of* α *chosen by the regularization parameter selection methods GCV, UPRE, and DP.*

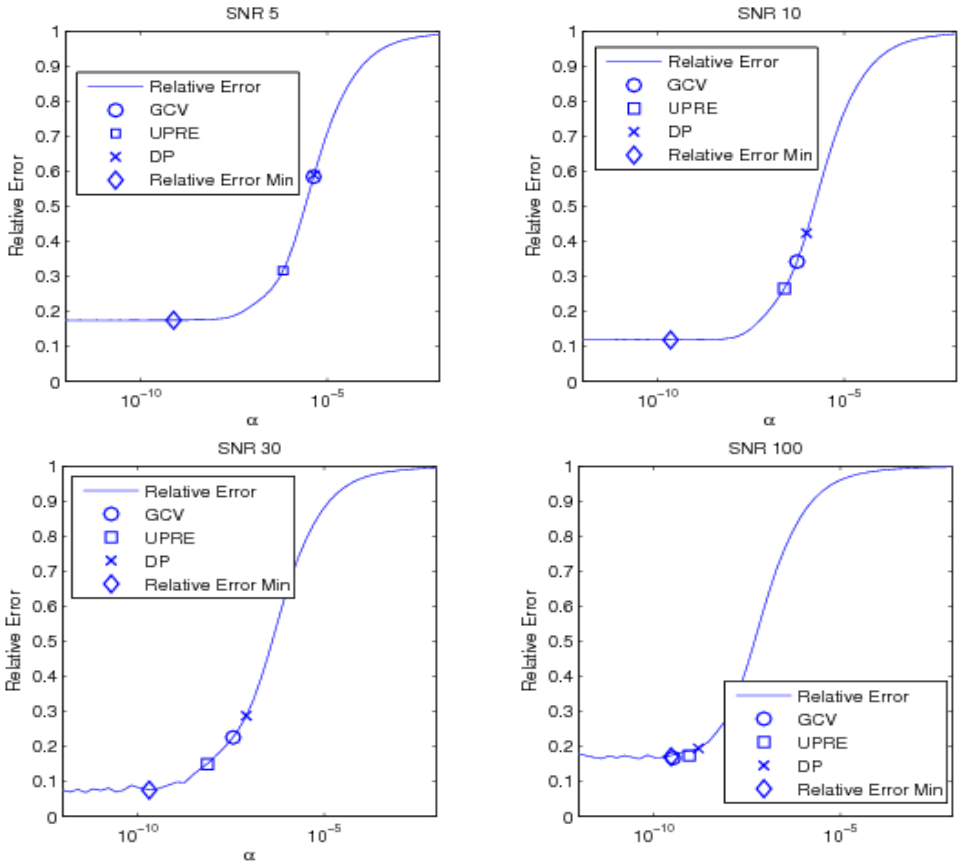

Figure 5.4: *Star Field Test Case. Plot of relative error together with the values of* α *chosen by the regularization parameter selection methods GCV, UPRE, and DP.*

tion in both cases of SNR. The bottom row of Figure 5.5 contains the results that were obtained from the satellite test case with $\mathbf{C} = \Theta$. Here $\mathbf{u}_{\text{approx}}$ was constructed using $\mathbf{C} = \mathbf{L}$ and the value for α that was selected by the UPRE method. Again the three methods each gave good recommendations for α, with the UPRE method giving the best recommendations.

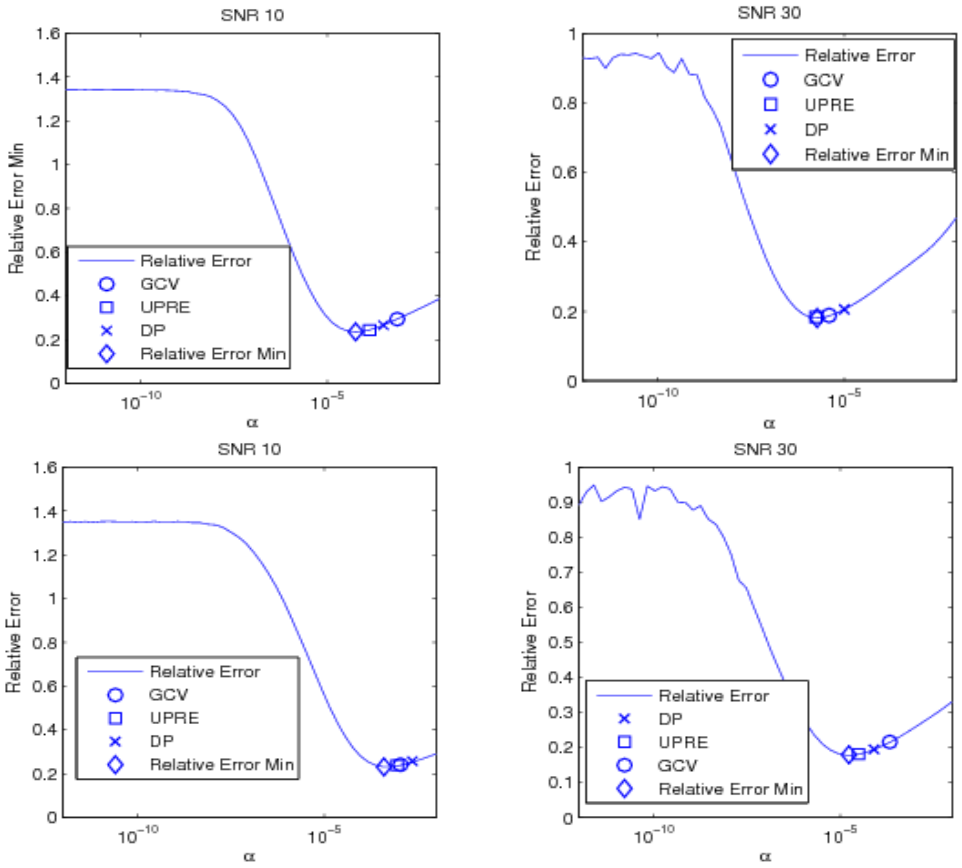

Figure 5.5: *In the top row, Satellite Test Case, with* $\mathbf{C} = \mathbf{L}$. *In the bottom row, Satellite Test Case, with* $\mathbf{C} = \Theta$. *Plot of relative error together with the values of* α *chosen by the regularization parameter selection methods GCV, UPRE, DP.*

The reconstructions that were obtained from the blurred, noisy satellite data with a SNR of 30 and using each of the three regularization matrices are displayed in Figure 5.6. The reconstructions were each computed with the UPRE value of α. Included in Figure 5.6 is the reconstruction that was obtained from the blurred, noisy star field data with a SNR of 30 and the value of α that was selected by UPRE.

In addition to testing the parameter selection methods, the iteratively updated diffusion regularization method was examined using the satellite test data with SNR 30. At each iteration, the value of α was selected using UPRE. Figure 5.7 contains, on the left, the plot of the reconstruction that was obtained after four iterations and on the left a plot of the reconstruction that was obtained after 6 iterations. Note that the bottom left plot in Figure 5.6 contains the result after 2 iterations. There is a noticeable improvement going from 2 to 4 iterations. However going from 4 to 6 iterations does not yield a reconstruction that is visibly much improved.

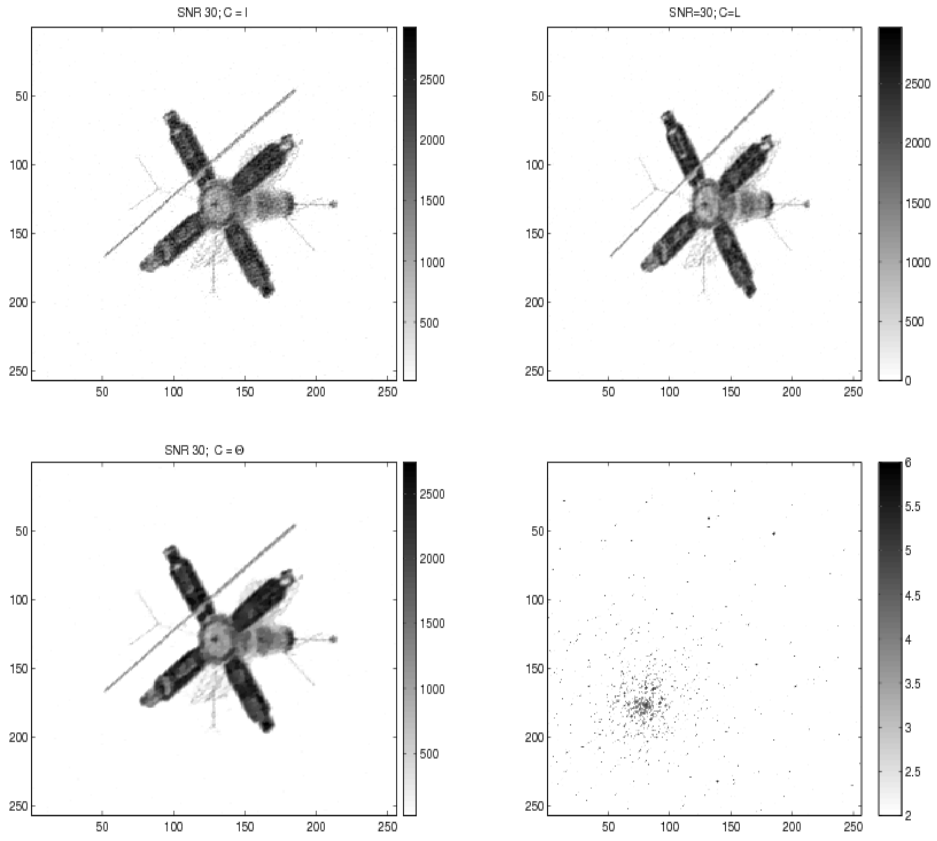

Figure 5.6: *Reconstructions of the satellite: on the upper-left with Tikhonov regularization ($\mathbf{C} = \mathbf{I}$), on the upper-right with Laplacian regularization ($\mathbf{C} = \mathbf{L}$), and on the lower-left with $\mathbf{C} = \Theta$. Reconstruction of the star field with Tikhonov regularization is given (on a log scale and with entries less than 100 set to 0) on the lower-right.*

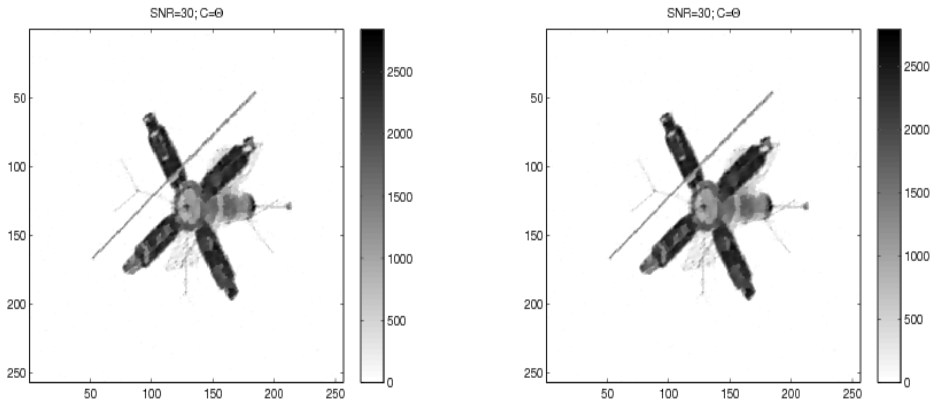

Figure 5.7: *The results from implementing the iteratively updated diffusion regularization function. On the left is the result after 4 iterations and on the right is the result after 6.*

5.2 PET Example

5.2.1 Statement of the Problem

Recall that in the PET imaging example the problem of interest is to estimate the true emission density $\mathbf{u}_e$ given observations $\mathbf{z}$ for which the data noise model is

$$\mathbf{z} = \text{Poiss}(\mathbf{A}\mathbf{u}_e + \gamma), \tag{5.10}$$

where γ gives the expected erroneous counts and $\mathbf{A}$ is the forward-model matrix. Here the underlying emission density array is $n \times n$ and for each of n_ϕ angles there are detectors at n_s different directed distances from center of the computational domain oriented at angle and so $\mathbf{A}$ is $M \times N$, where $M = n_\phi n_s$ and $N = n^2$.

$\mathbf{A}$ has the form

$$\mathbf{A} = \mathbf{G}\mathbf{A}^{\text{radon}}, \tag{5.11}$$

where

$$\mathbf{G} = \text{diag}\left(e^{-\int_{L_1}\mu(x)dx}, e^{-\int_{L_2}\mu(x)dx}, \ldots, e^{-\int_{L_M}\mu(x)dx}\right),$$

L_i indicates the ith line of response (LOR), $\mu(x)$ is the attenuation function, and $\mathbf{A}^{\text{radon}}$ is the matrix that arises from the discretization of the radon transform. Note that $[\mathbf{A}^{\text{radon}}]_{ij}$ is the length of the i LOR with with the j computational grid element. Thus $\mathbf{A}$ is sparse and so multiplication by $\mathbf{A}$ is not prohibitively expensive.

5.2.2 Regularization Parameter Selection Results

The synthetically generated true emission density $\mathbf{u}_e$ that was used to generate the transformed, noisy data is similar to that used in [32] and is shown on the left in Figure 5.8. As in the astronomical imaging example, Matlab's `poissrnd` function was employed to generate the noisy data using statistical model (5.10). The noisy sinogram is shown on the right in Figure 5.8, γ was assumed to be 1, and the attenuation function was assumed to be zero at all pixels. The unknown emission density array is 128×128, and the data was collected using 128 detectors at 128 different angles. Therefore $M = N = 128^2$. As in the astronomical imaging example, the SNR was varied in order to test the methods on multiple data sets. For data arising from model (5.10), the SNR is given by

$$SNR = \sqrt{\frac{\|\mathbf{A}\mathbf{u}_e + \gamma\|^2}{\sum_{i=1}^{M}([\mathbf{A}\mathbf{u}_e]_i + \gamma_i)}}.$$

Plots of the relative solution error obtained for a range of values of α and the values recommended by the selection methods are displayed in Figure 5.9. The MAP estimate of $\mathbf{u}_e$ is calculated from (4.1) with $\mathbf{C} = \mathbf{L}$. In both cases of SNR, the UPRE and GCV methods gave slightly better recommendations than the DP method. The reconstructions obtained using the DP method are displayed in the top row in Figure 5.10, and the bottom row contains reconstructions computed with the UPRE recommendations.

In addition to implementing the parameter selection methods with $\mathbf{C} = \mathbf{L}$ in (4.1), the methods were also employed to choose a value for α when $\mathbf{C} = \Theta$, the regularization matrix whose form is given in (5.5) and which allows for edges present in the true image to be preserved in the reconstruction. Θ

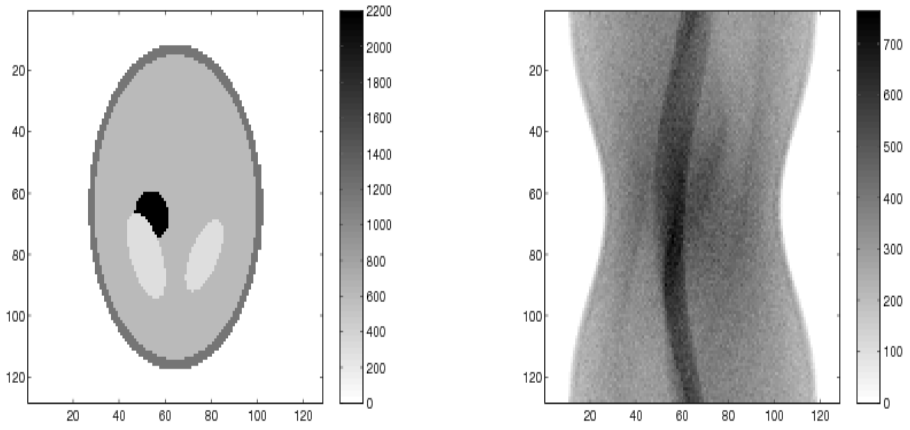

Figure 5.8: *The true emission density ub_e is plotted on the left and the data* **b** *is plotted on the right. The signal-to-noise ratio of* **z** *is 20.*

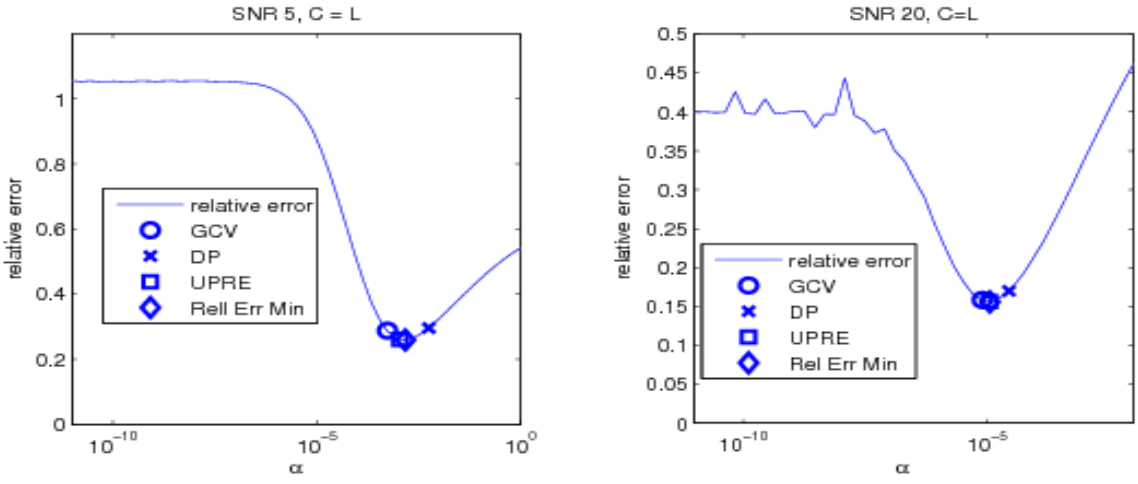

Figure 5.9: *Plots of α versus relative error are shown. The plot on the left is from data with a SNR of 5 and the plot on the right is from data with a SNR of 20.*

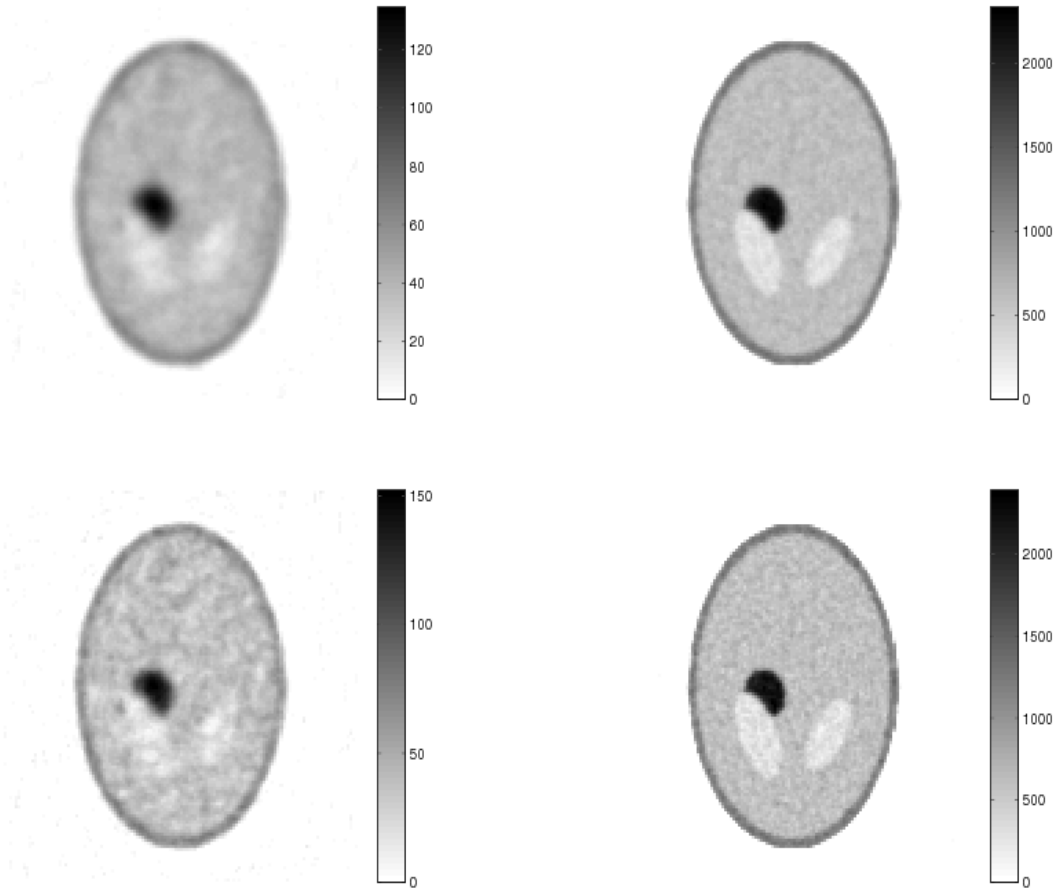

Figure 5.10: *Plots of the reconstructions obtained from the two data sets, with the reconstructions obtained from data with SNR=5 on the left and SNR=20 on the right. The top row contains the reconstructions that were computed with the DP recommendation and the bottom row contains reconstructions that were computed with the UPRE recommendation.*

was constructed in the same manner as in the astronomical imaging example, with the exception that the scaling parameter τ present in (5.8) was set to 500. The DP method was used to select the value of α that was used to compute $\mathbf{u}_{\text{approx}}$. Figure 5.11 contains plots of the relative solution error that was computed over a range of values of α as well as the solution error corresponding to the parameter selection method recommendations. As in the case when $\mathbf{C} = \mathbf{L}$, the UPRE and GCV methods both gave slightly better recommendations in terms of the relative solution error than the DP method did for both instances of SNR. The reconstructions computed using the DP recommendation are displayed in Figure 5.12.

The computation of the estimates of the true emission density required the solution of the nonnegatively constrained minimization problem (4.1). The minimizer was computed using the GPRN algorithm outlined in Chapter 2. A maximum of 50 iterations of the algorithm were allowed. The tolerances on the step size and the size of the norm of the projected gradient were both set to 10^{-5}. In each iteration of the GPRN algorithm, a maximum of 5 gradient projection iterations and 30 conjugate gradient iterations were allowed. The stopping tolerances for both the gradient projection step and the reduced Newton step were both set to 0.1.

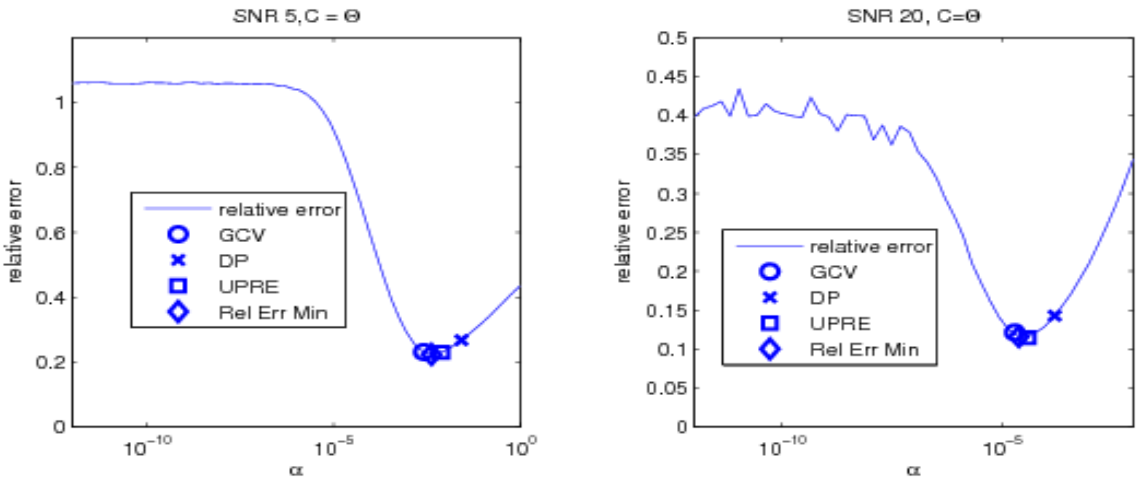

Figure 5.11: *Plots of α versus relative error are displayed.*

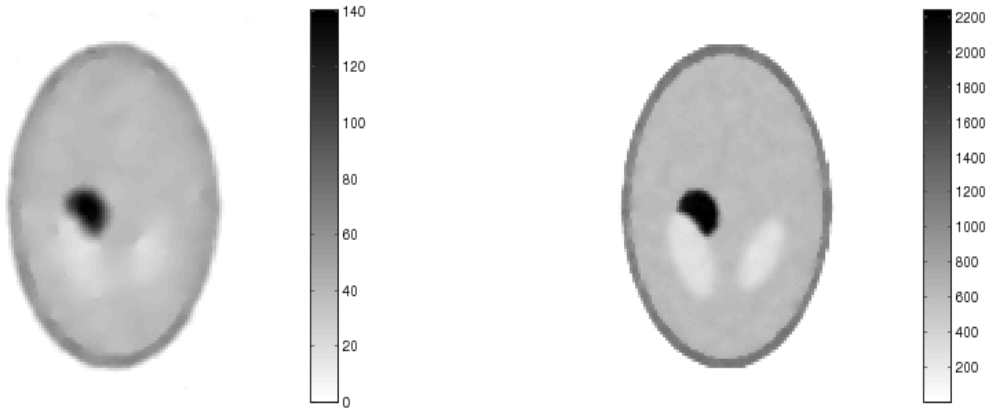

Figure 5.12: *Reconstructions computed using* $\mathbf{C} = \Theta$ *and the DP recommendation for* α *are shown. The plot on the left is the reconstruction computed from data with a SNR of 5 and on the right is the reconstruction that was obtained from data with a SNR of 20.*

5.2.3 Total Variation Regularization

In (4.1), the use of $\mathbf{C} = \Theta$ yielded a quadratic penalty term that allowed for edge preservation. Total variation regularization is another regularization technique that allows for the presence of sharp edges in the resulting reconstruction. As discussed in Chapter 3, the fact that the algorithm used to solve the nonnegatively constrained problem (4.1) requires that the function to be minimized be differentiable, and so an approximation of the total variation function must be used due to the fact that the total variation function is not differentiable at zero. The approximation that is used is given by

$$J(\mathbf{u}) = \frac{1}{2}\sum_{i=1}^{M}\sqrt{[\mathbf{D}_x\mathbf{u}]_i^2 + [\mathbf{D}_y\mathbf{u}]_i^2 + \beta}, \tag{5.12}$$

where $\beta > 0$. In our experiments $\beta = 1$. Computing $\mathbf{u}_\alpha$ using total variation regularization entails solving

$$\mathbf{u}_\alpha = \text{argmin}_{\mathbf{u}>\mathbf{0}}\left\{T_\alpha(\mathbf{u}) \stackrel{\text{def}}{=} T_0(\mathbf{u};\mathbf{z}) + \alpha J(\mathbf{u})\right\}. \tag{5.13}$$

The solution of (5.13) was computed using an algorithm very similar to the GPRN algorithm outlined in Chapter 2. The key difference is in the second stage. Instead of using conjugate gradient iterations

to solve the reduced Newton system

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

the system

$$\nabla^2_{\text{LD}} T_\alpha(\mathbf{u}_k)\mathbf{p} = -\nabla_{\text{red}} T_\alpha(\mathbf{u}_k), \tag{5.14}$$

where $\nabla^2_{\text{LD}} T(\mathbf{u}_k) = \nabla^2_{\text{red}} T_0(\mathbf{u}_k\mathbf{z}) + \alpha\mathbf{L}_1(\mathbf{u}_k)$, and $\mathbf{L}_1$ is defined in (3.23) is solved instead. This step will be referred to as the reduced lagged-diffusivity step due to the fact that when $T_0(\mathbf{u})$ is the regular least squares function, the lagged-diffusivity fixed point iteration of [34] requires the solution of the unreduced system

$$\nabla^2_{\text{LD}} T_\alpha(\mathbf{u}_k)\mathbf{p} = -\nabla T_\alpha(\mathbf{u}_k)$$

at each iteration. Hence the algorithm used to solve (5.13) will be referred as GPLD.

Figure 5.13 contains the plots analogous to those in Figure 5.9 which resulted from computing $\mathbf{u}_e$ from (5.13). Note that all three parameter selection methods yielded good recommendations for the parameter value. Figure 5.14 contains plots of the reconstructions that were obtained using the UPRE method to select the value of α for the data with a SNR of 5 and the DP method for the data with a SNR of 20.

The evaluation of both the GCV and UPRE functions requires the computation of $\mathbf{A}_\alpha$. Recall that the form of $\mathbf{A}_\alpha$ was motivated by the fact that $\mathbf{u}_\alpha$ is the solution of equation (4.26) which has minimum norm. In the case of total variation regularization, the solution of minimum norm of equation (4.26) has the form

$$(\mathbf{D}_\alpha(\mathbf{A}^T\mathbf{Z}_\alpha^{-1}\mathbf{A} + \alpha\mathbf{L}_1(\mathbf{D}_\alpha\mathbf{u}))\mathbf{D}_\alpha)^\dagger\mathbf{D}_\alpha\mathbf{A}^T\mathbf{Z}_\alpha^{-1}(\mathbf{z}-\gamma), \tag{5.15}$$

which motivates the following expression for $\mathbf{A}_\alpha$:

$$\mathbf{A}_\alpha = (\mathbf{D}_\alpha(\mathbf{A}^T\mathbf{Z}_\alpha^{-1}\mathbf{A} + \alpha\mathbf{L}_1(\mathbf{D}_\alpha\mathbf{u}))\mathbf{D}_\alpha)^\dagger\mathbf{D}_\alpha\mathbf{A}^T\mathbf{Z}_\alpha^{-1/2}. \tag{5.16}$$

However it was found that better results were obtained by modifying (5.17) in the following manner:

$$\mathbf{A}_\alpha = (\mathbf{D}_\alpha(\mathbf{A}^T\mathbf{Z}_\alpha^{-1}\mathbf{A} + \alpha(\mathbf{L}_1(\mathbf{u}) + 2\mathbf{L}_2(\mathbf{u}))\mathbf{D}_\alpha)^\dagger \mathbf{D}_\alpha \mathbf{A}^T \mathbf{Z}_\alpha^{-1/2}. \tag{5.17}$$

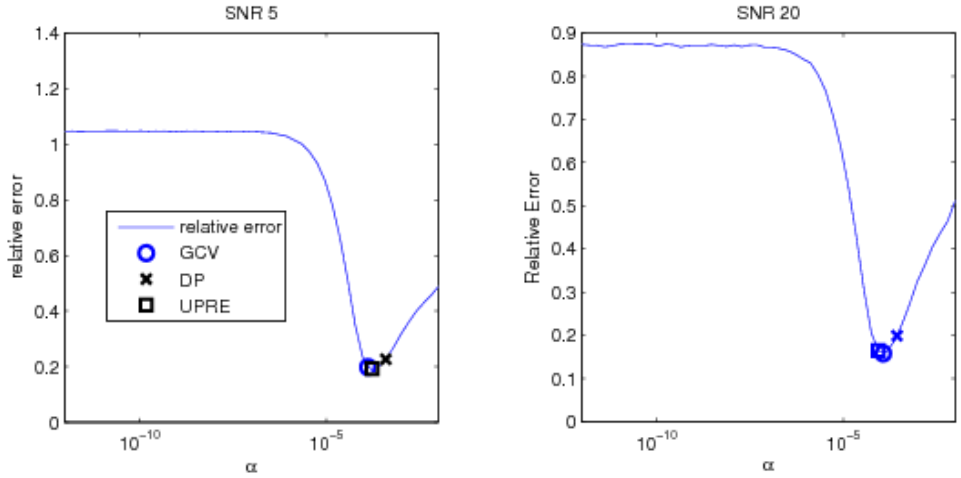

Figure 5.13: *Plots of α versus relative error are shown. The plot on the left is from data with a SNR of 5 and the plot on the right is from data with a SNR of 20.*

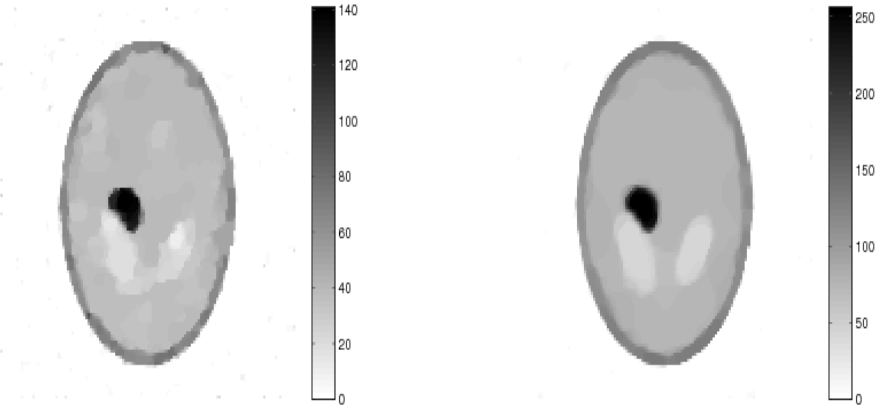

Figure 5.14: *Plots of the reconstructions obtained from the two data sets with the UPRE recommendation for the regularization parameter with SNR 5 (on the left) and the DP recommendation for the regularization parameter with SNR 20 (on the right).*

Numerical tests of the parameter selection methods applied to (5.13) were also performed on a synthetically generated emission density that is based on an anatomical model of a normal brain and was obtained from [1]. The true emission density is given on the left in Figure 5.15. The noisy sinogram

data, generated using statistical model (5.10) and MATLAB's `poissrnd` function, is shown on the right in Figure 5.15. We again assumed that γ is a constant vector of 1s at all pixels, and that the density vector μ is zero. Our computational grid is defined here by 128 detectors and angles, as well as a 129×129 uniform computational grid for the unknown emission density. Thus $M = 128^2$ and $N = 129^2$. Figure 5.16 contains, on the left, a plot of the relative error over a range of values of α along with the recommendations by the three methods. On the right is a plot of the reconstruction obtained using the DP recommendation. As in the previous cases, the three methods each gave a good recommendation for α.

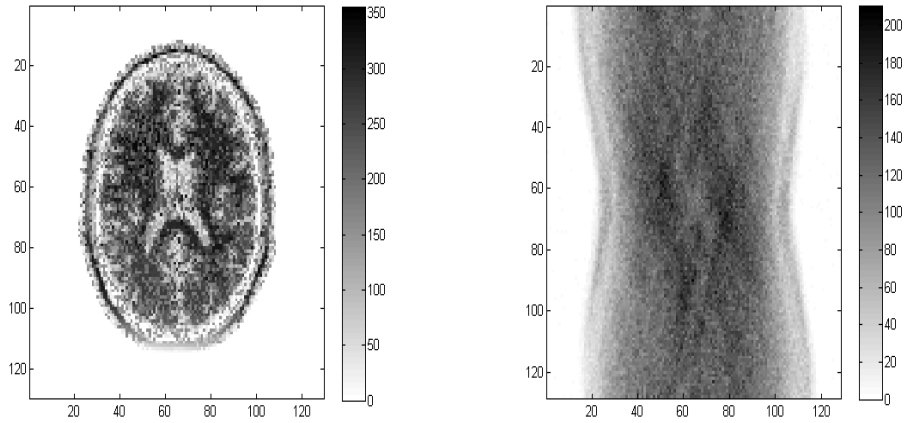

Figure 5.15: *The true emission density* $\mathbf{x}_e$ *is plotted on the left and the data* $\mathbf{b}$ *is plotted on the right.*

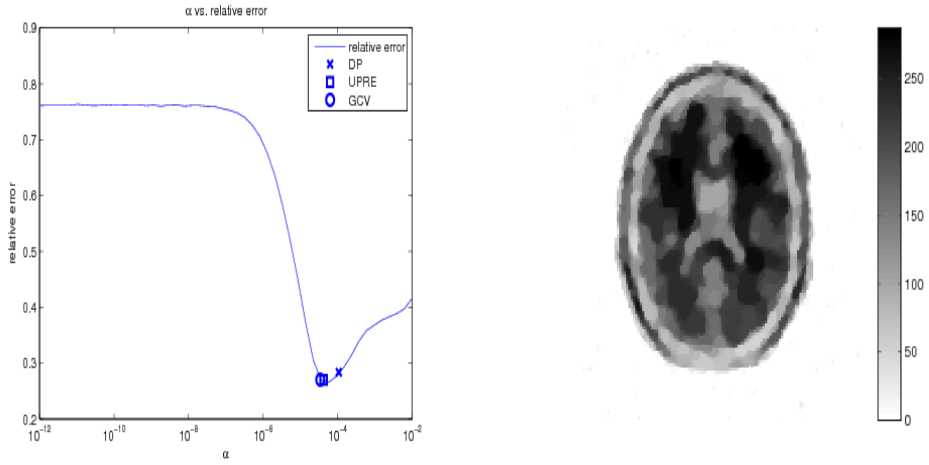

Figure 5.16: *On the left is a plot of* α *vs. the relative error for the brain image. On the right is the reconstruction obtained using the DP recommendation.*

The efficiency of the GPLD algorithm was evaluated by comparing its performance to that of the EM-TV algorithm presented in [31]. The algorithms were tested on data generated from the emission density shown on the left in Figure 5.8 with SNR 5. The CPU time for computing $\mathbf{u}_\alpha$ using the GPLD algorithm with 50 iterations and the DP recommendation for α, 0.00015, was 43.8 seconds. The relative solution error was 0.1860. The CPU time for computing $\mathbf{u}_\alpha$ using the EM-TV algorithm with 50 outer iterations and 50 iterations to compute the dual variable and with $\alpha = 0.05$ was 37.5 seconds. The relative solution error was 0.2253. Note that a smaller relative error could be obtained by using a better value of α, however I have so far been unable to incorporate the parameter selection methods into the Matlab code for the EM-TV algorithm. Also in the EM-TV algorithm, in the computation of the dual variable,if the iterates are stopped when the appropriate quantity becomes lower than some stopping tolerance. A better relative error could be obtained, however the CPU time could be increased significantly.

Chapter 6

Conclusions

In Chapter 1, Ill-posed Poisson estimation was introduced in the contexts of astronomical and PET imaging. Since the maximum likelihood estimate is insufficient in such problems, regularization techniques must be employed. Regularization introduced two important issues: what regularization function should be used and what value should the regularization parameter be set to. Also, computing the solution of the regularized problem required a method for solving nonnegatively constrained minimization problems.

In Chapter 2, an algorithm for solving nonnegatively constrained minimization problems in which the objective function is twice-continuously differentiable was described. Convergence of the algorithm to a unique minimizer was proved for objective functions that are strictly convex, coercive, and whose gradient is Lipschitz continuous.

In Chapter 3, regularization was motivated by taking a Bayesian approach and formulating a MAP estimation problem. Chapter 3 also contains a discussion of various regularization functions and proofs that those functions yield an objective function that satisfies the criteria for convergence of the algorithm described in Chapter 2. Of interest were quadratic regularization terms in which the regularization matrix was either the identity, a discretization of the negative Laplacian, or a matrix that

yielded a regularization term that allowed for the preservation of edges. Total variation regularization was also examined.

Methods for selecting the regularization parameter was the topic of Chapter 4. A Taylor series argument and an application of the mean value theorem were used to derive a weighted sum of squares approximation of the negative log of the Poisson likelihood function. This approximation was used to extend parameter selection methods that have been developed for least squares problems to regularized negative-log Poisson reconstruction problems. The methods of interest were the discrepancy principle, generalized cross validation, and unbiased predictive risk estimator methods.

In Chapter 5 numerical tests of the methods were performed in the context of the astronomical and PET imaging examples. The selection methods were implemented with each of the regularization functions described in Chapter 3. The methods were shown to give good recommendations for the regularization parameter. Additionally the construction of an edge-preserving quadratic regularization function was detailed.

A complete computational framework for solving ill-posed Poisson imaging problems has been presented.

Bibliography

[1] *http://mouldy.bic.mni.mcgill.ca/brainweb/selection$_n$ormal.html*.

[2] *http://www.math.umt.edu/bardsley/co*.

[3] Sangtae Ahn and Jeffrey Fessler, *Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms*, IEEE Transactions on Medical Imaging **22** (2003), no. 5, 613–626.

[4] Johhathan M. Bardsley and John Goldes, *An iterative method for edge-preserving map estimation when data-noise is poisson*, SIAM Journal on Scientific Computing.

[5] Johnathan Bardsley and Aaron Luttman, *Total variation-penalized poisson likelihood estimation for ill-posed problems*, Advances in Computational Mathematics **31** (2009), no. 1, 35–59.

[6] Johnathan M. Bardsley, *An efficient computational method for total variation-penalized poisson likelihood estimation*, Inverse Problems and Imaging **2** (2008), no. 2, 167–185.

[7] ———, *Stopping rules fora nonnegatively constrained iterative method for ill-posed poisson imaging problems*, BIT Numerical Mathematics **49** (2008December), no. 4.

[8] Johnathan M. Bardsley and John Goldes, *Regularization parameter selection methods for ill-posed poisson maximum likelihood estimation*, Inverse Problems (2009).

[9] ———, *A computational framework for total variation-regularized positron emission tomography* (2010). submitted.

[10] ———, *Techniques for regularization parameter and hyper-parameter selection in positron emission tomography* (2010). submitted.

[11] Johnathan M. Bardsley and N'djekornum Dara Laboel, *An analysis of regularization by diffusion for ill-posed poisson likelihood estimation*, Inverse Problems in Science and Engineering **17** (2009June), no. 4, 537–550.

[12] Johnathan M. Bardsley and N'djekornom Dara Laobeul, *Regularized poisson likelihood estimation: Theoretical justification and a computational method*, Inverse Problems in Science and Engineering **16** (2008January), no. 2, 199–215.

[13] D. Calvetti and E. Somersalo, *Hypermodels in the bayesian imaging framework*, Inverse Problems (2008).

[14] Jeffrey Fessler, *Penalized weighted least squares image reconstruction for positron emission tomography*, IEEE Transactions on Medical Imaging (1994June), 290–300.

[15] Jeffrey Fessler and Alfred Hero, *Space-alternating generalized em algorithms*, IEEE Transactions on Signal Processing (1994October), 2664–2677.

[16] Jeffrey Fessler and Alfred O. Hero, *Penalized maximum-likelihood image reconstruction using space alternation generalized em algorithms*, IEEE Transactions on Image Processing (1995October), 1417–1429.

[17] Hongbin Guoa, Rosemary A. Renauta, Kewei Chenb, and Eric Reimanb, *Fdg-pet parametric imaging by total variation minization*, Computerized Medical Imaging and Graphics (2009), 295–303.

[18] T. Herbert and R. Leahy, *A generalized em algorithm for 3-d bayesian reconstruction from poisson data using gibbs priors*, IEEE Transactions on Medical Imaging (1989June), 194–202.

[19] T. Hsiao, A. Rangarajan, and G. Gindi, *Bayesian image reconstruction for transmission tomography using deterministic annealing*, Journal of Electronic Imaging **12** (2003), no. 7.

[20] J. Merikoski J. M. Bardsley and R. Vio, *The stabilizing properties of nonnegativity constraints in least-squares image reconstruction*, International Journal of Pure and Applied Mathematics **43** (2008), no. 1, 95–109.

[21] M. Vauhkonen J. P. Kaipio V. Kolehmainen and E. Somersalo, *Inverse problems with structural prior information*, Inverse Problems **15** (1999), 713–629.

[22] Elias Jonsson, Sung cheng Huang, and Tony Chan, *Total-variation regularization in positron emission tomography*, Technical Report 98-48, UCLA Group in Computational and Applied Mathematics, 1998.

[23] C.T. Kelley, *Iterative methods for optimization*, Society for Industrial and Applied Mathematics, 1999.

[24] S.-J. Lee, A. Rangarajan, and G. Gindi, *Bayesian image reconstruction in spect using higher order mechanical models as priors*, IEEE Transactions on Medical Imaging (1995), 669–680.

[25] J. J. Moré and G. Toraldo, *On the solution of large quadratic programming problems with bound constraints*, SIAM Journal on Optimization **1** (1991), no. 1, 93–113.

[26] V. A. Morozov, *On the solution of functional equations by the method of regularization*, Soviet Mathematics Doklady **7** (1966), 414–417.

[27] E. U. Mumcuoglu, R. Leahy, S. R. Cherry, and Z. Zhou, *Fast gradient-based methods for bayesian reconstruction of transmission and emission pet images*, IEEE Transactions On Medical Imaging (1994), 687–701.

[28] J. Nocedal and S. J. Wright, *Numerical optimization*, Spring-Verlag, 1999.

[29] John M. Ollinger and Jeffrey A. Fessler, *Positron-emission tomography*, IEEE Signal Processing Magazine (1997January).

[30] M. C. Roggeman and Byron Welsh, *Imaging through turbulence*, CRC Press, 1996.

[31] Alex Sawatzky, Christoph Brune, Jahn Müller, and Martin Burger, *Total variation processing of images with poisson statistics*, Caip '09: Proceedings of the 13th international conference on computer analysis of images and patterns, 2009, pp. 533–540.

[32] L.A. Shepp and Y. Vardi, *Maximum likelihood reconstruction in postron emission tomography*, IEEE Transactions on Medical Imaging **MI-1** (1982), 113–122.

[33] 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 **10** (1993), 1014–1023.

[34] Curtis R. Vogel, *Computational methods for inverse problems*, SIAM, Philidelphia, 2002.

[35] G. Wahba, *Practical approximate solutions to linear operator equations when the data are noisy*, SIAM Journal on Numerical Analysis **14** (1977), 651–667.

[36] D. F. Yu and J. A. Fessler, *Edge-preserving tomographic reconstruction with nonlocal regularization*, IEEE Transactions on Medical Imaging (2002), 159–173.

[37] E. Zeidler, *Applied functional analysis: Main principles and their applications*, Springer-Verlag, New York, 1995.