Image Reconstruction in Emission Computed Tomography: Iterative Methods

April 17, 2007

1 The Problem

1.1 Overview

The overall goal of this module is to show students the connection between physical problems in medical imaging and theoretical problems in mathematics. In particular, students will learn how such a physical problem can be translated into the language of mathematics, how one then uses standard mathematical techniques to solve the problem, and finally how the mathematical solution is re-translated back into a physical interpretation. As with all problems involving mathematical modelling of physical phenomena, it is important to evaluate the methods used and results obtained, and so students will also try to assess the work they have conducted.

The specific type of medical imaging we will investigate is called single-photon emission computed tomography (SPECT) [1, 2, 3, 4]. The purpose of SPECT imaging is to develop pictures, in the form of digital images, of the interior of the human body. These images measure how well the organs of the body are functioning. If there is some type of physiological dysfunction, it will be visually apparent to the physician looking at the image. In this way, the image can be used by the doctor as a diagnostic tool, without having to physically cut open the patient and do exploratory surgery. As a potential patient, anyone should clearly see how useful SPECT imaging could be!

Images are made by first collecting data on the organ function of the patient using a SPECT imaging system. The data then need to be processed using a mathematical algorithm implemented on a computer in order to form an image of the patient. This data processing step is called *image reconstruction*, which requires us to develop a mathematical model of the data collection process. The standard model that we will use is to equate the data collection process with a linear system of equations. We can then draw upon a variety of solution methodologies from the field of linear algebra in order to solve the linear system. After solving the linear system, we then re-interpret its solution as the

reconstructed image, and can then assess the quality of the model developed, the methods used to solve the problem, as well as the solution itself. By the end of the module, students should be able to see the connection between physical problems and mathematics, develop their own models and implement them, and to analyze and critically assess the results.

This module is intended for a general audience with a background in single- and multi-variable calculus, and linear algebra. Although no computer formal programming skills are required, students should be comfortable using MAPLE and MATLAB, as these programs are used in the latter sections. The only mathematical concepts that students might not have been exposed to would be the iterative solution methodologies themselves [5, 6]. These topics will be extensively reviewed within the module, however students should have taken a course in "Numerical Methods" or "Numerical Linear Algebra" in order to better learn these important concepts. Also, after having completed this module, students will be able to extend and apply the concepts and skills learned to the *Image Reconstruction in Emission Tomography I. Non-iterative Methods* module. In that module, students will study *non-iterative* methods for solving linear systems and how they can be applied to the same image reconstruction problem. The flexibility of this module lies in the fact that it is essentially self-contained, except for the required mathematical knowledge. Ideally, it will be used as an integrated component in a section of a course that focuses on the modelling physical phenomena using linear systems of equations. Also, it would be best used by incorporating both classroom lectures, group work and computer lab sessions into the presentation of the material.

Prior to beginning the module, students should do some background reading on emission computed tomography. The two most accessible references on SPECT imaging at the student level would be English and Brown [1] and Guy and ffytche [2]. As there is considerable overlap in the mathematics between SPECT and x-ray CT, students can also examine Section 11.13 of Anton and Rorres [7] to be able to see the link between linear algebra and CT imaging. In addition, the CT tutorial authored by Professor Elizabeth Jessup of the Department of Computer Science at the University of Colorado at Boulder would be helpful to examine. It can be accessed at:

http://www.cs.colorado.edu/ftp/pub/HPSC/README.html

and downloaded by selecting the "Tomography Tutorial" at the bottom of the page. Other relevant material at various levels of difficulty may be found in the **Bibliography** section of this module.

1.2 Introduction

Medical imaging, which is extremely commonplace in today's world, can trace its origins back to 1895, when Wilhelm Röntgen discovered x-rays [2, 8]. In a short time, x-ray *radiography* was developed, and today is still in use as a primary diagnostic tool. More than likely, you have had x-ray images taken of

your body, possibly in the form of dental images or to detect a fracture of a bone in your arm. In the latter case, a physician (or more likely a technician) places a film sensitive to x-rays on one side of your lower arm and then irradiates the other side using an x-ray transmission device. Some x-rays will pass through your forearm and expose the film, while some will be completely stopped (or "attenuated") and not expose it. Attenuation of x-rays is dependent on the density of the material an x-ray attempts to pass through, as well as the distance through the body that an x-ray must travel. Being dense, bone tends to stop most x-rays, while muscle typically allows x-rays to travel through unabated. The variable attenuating properties of different types of tissue allow for the film to be exposed in such a way as to show the contrast between the different structures in your arm. In this way, by examining the x-ray radiographs, the physician may detect any fracture in your radius or ulna bones. It is important to note though, that the radiograph is a two-dimensional (2D) "planar" projection image of your three-dimensional (3D) arm. Thus, the structures in your arm overlap and the 3D depth information is lost.

The problem of loss of depth information was finally resolved in 1972, when Sir Godfrey Hounsfield introduced the first commercially useful x-ray computer assisted tomographic (CAT or CT) scanner to the world [2, 8]. The word "tomography" has origins in the Greek root *Tomos*, which means "slice." Typical x-ray CT scanners use a detector (rather than film) which digitally records the number of x-rays that impinge on it. The x-ray detector/transmitter pair, which are located 180° opposed from each other, rotates around the patient in a 360° orbit. Several 2D planar projection images through the patient are taken and the data are reconstructed using the "filtered back-projection" (FBP) image reconstruction algorithm. The FBP algorithm has its origins in the 1917 classic paper by Austrian mathematician Johan Radon [9], which discussed line integral projections through an object. Although this mathematical framework for conducting tomographic imaging has been around for several decades, its usefulness was only fully realized with the advent of the digital computer. The result of applying FBP reconstruction to the data is a series of 2D cross-sectional images (or trans-axial slices), which can be stacked on top of each other to form a 3D representation of the interior of the patient. Such images can then be sequentially viewed by a physician in order to detect structural differences or defects.

The key point behind both x-ray radiography and x-ray CT, is that both imaging modalities yield images that describe the interior *structure* of the patient. In fact, one could create an x-ray CT image of a living person, and then repeat the process once the person has died and not be able to discern the difference between the two images. The disadvantage to only having morphological information is that doctors may not be able to detect diseased or cancerous tissue. For example, one might clearly discern the presence of the heart in a patient, but not be able to tell if the ventricles are working properly.

SPECT imaging, which uses γ-rays instead of x-rays, provides *physiological*, rather than morphological, information. Its origins can be traced back to the development of the atomic bomb in the 1940's, when nuclear fission reactors

produced artificial radio-isotopes. In 1948, Ansell and Rotblat acquired the first thyroid image, while in 1952, Hal Anger developed a γ-ray detector using a sodium-iodide (*NaI*) scintillation crystal. However it wasn't until 1963 when Kuhl and Edwards presented the first *functional* tomographic images to the world. Instead of x-rays being transmitted through the patient as previously described, γ-rays are *emitted* from within the patient. These emitted γ-rays are then recorded digitally using a detector that rotates around the patient in a 360° orbit. Data measurements, in the form of 2D planar projection images, are acquired at various angular positions around the patient. We then reconstruct the 3D image volume from the data measurements using a mathematical algorithm. The 2D cross-sectional images can then be viewed by the doctor and used to diagnose the health state of the patient. More details on how the reconstructed image is formed will be given in subsequent sections.

1.3 Statement

The key problem that we will address is "How do we obtain the reconstructed image from the projection data measurements?" To simplify matters, we shall only consider a single 2D trans-axial slice of the patient (a slice perpendicular to the long axis of the body), instead of the entire 3D volume, and the 1D projection data measurements that correspond to it. The first step is to develop a mathematical model for the data acquisition process, which in our case will be a linear system of equations. Next, we need to select a mathematical solution methodology from those available to us from the theory of linear algebra, given the physical constraints of the problem. Some examples of iterative techniques we might use include the Jacobi and Gauss-Seidel methods. A solution method is then applied to the measured projection data and the 2D reconstructed image is obtained. Finally, we need to assess the quality of the solution, and any factors in the modelling process or solution methodology that contributed to any negative effects affecting the solution. Therefore, it will be important to understand the theoretical aspects of the solution method used, as well as to use knowledge of the actual distribution of radioactivity in the patient for comparison purposes.

2 Model, Method and Implementation

2.1 Background Information

A schematic of a typical SPECT imaging system can be seen in Figure 1. The acquisition begins by administering a radio-tracer to the patient, which is often done by injection. This radio-tracer typically consists of a radioactive isotope that has been chemically attached to a tracer compound. The selection of the tracer is specific, in that it is a known substance that will be metabolized by the cells comprising the target organ. For example, if we want to measure brain function, we could use fluorodeoxyglucose (FDG), which is ^{18}F (radioactive

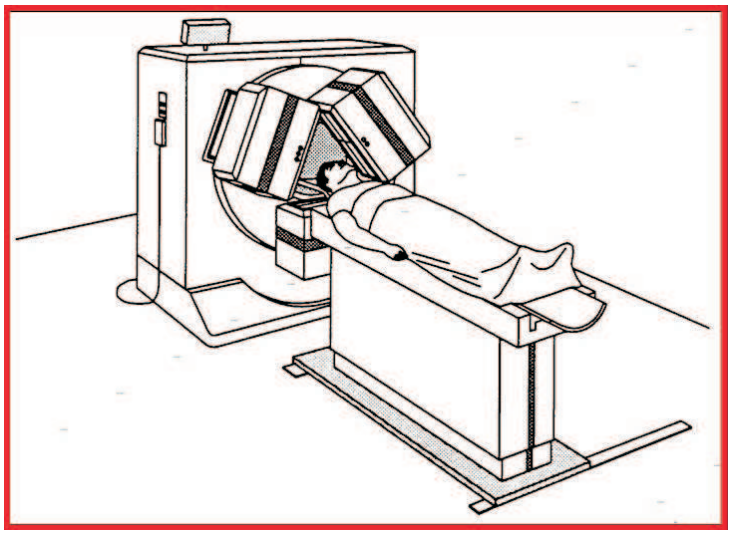

Figure 1: A schematic of a SPECT system with three detector heads. The figure is provided courtesy of Dr. F. Beekman of the Image Sciences Institute, University Medical Center, Utrecht, NE.

Fluorine) attached to glucose[1]. FDG is used in brain perfusion imaging since the brain cells process glucose, as well as the fact that glucose is one of the few substances that can permeate the blood-brain barrier. Alternatively, we might wish to measure cardiac function, and so ^{201}Tl (radioactive Thallium) could be injected into the patient's bloodstream, which would then perfuse to the heart. The cardiac tissue would then metabolize the isotope, which would allow us to measure cardiac dysfunction, including the presence of scar tissue[2].

After being administered to the patient, the radio-tracer travels to the target organ and is subsequently metabolized by the organ's cells. The degree of cell function is proportional to the amount of radio-tracer that is metabolized. Healthy tissue will metabolize the radio-tracer at a normal rate consistent with healthy cell function, while tumorous and/or diseased (lesionous) tissue will metabolize at a higher or lower rate than normal. As the process of metabolization occurs, the radio-isotope begins to decay, and γ-rays are emitted from the area of the target organ. The γ-rays are simply another form of high energy electromagnetic wave, much like x-rays, but at a different wavelength. They can pass through the body's tissue, and are randomly emitted in all directions from within the patient. As they exit the body, some will impinge on the γ-ray detector, which is used to measure the photon flux emanating from the patient (which we will generically call the "object")[3] .

[1]Refer to Sorensen and Phelps [4], pp. 502-509.
[2]Refer to English and Brown [1], pp. 85-89.
[3]Refer to Sorensen and Phelps [4], pp. 429-430 or Guy and fftyche [2], pp. 177-186 for

Figure 2: A schematic detailing the emission and detection process of a typical γ-ray.

Figure 2 provides a good schematic of the emission and detection process for a given γ-ray photon. A single cross-sectional slice of both the detector and object are shown. Since patients roughly have an elliptical cross-sectional shape, the object is simply modelled as an ellipse. In order to be detected, an emitted photon must travel within the field-of-view (FOV) of one of the measurement locations on the detector. The FOV for a given detector location is defined by the septa (or walls) of the collimator, which is a device attached to the front of the detector. Basically, it can be thought of as a lead plate with a pattern of holes drilled into it, much like a honeycomb. Only those photons travelling along a line almost perpendicular to the detector face will pass through one of the holes in the collimator and strike the detector, and thus will be measured. If a photon travels within the FOV, it will hit the NaI scintillation crystal behind the collimator. As the photon interacts with the NaI crystal, the γ-ray

descriptive photos and schematics of a typical SPECT acquisition system.

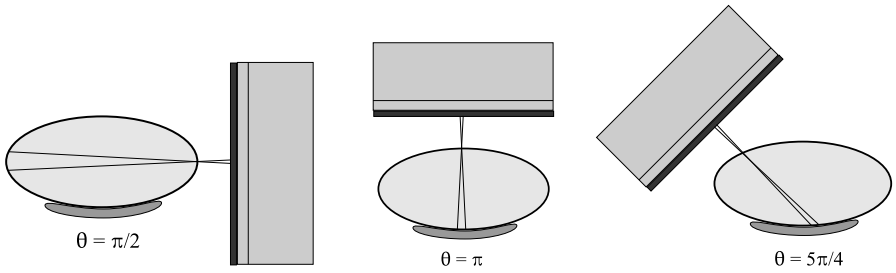

Figure 3: The detector is rotated, so that data can be measured from several angular positions around the patient. Shown are acquisition angles $\theta = \frac{\pi}{2}$, π, and $\frac{5\pi}{4}$.

is converted into optical photons (visible light). The array of photo-multiplier tubes (PMT), together with the position logic circuits, then view the optical photons, and estimate the energy of the γ-ray and the position at which it impinged on the detector face. This datum is then digitally recorded using a computer. Each photon that impinges on the detector is measured in a similar fashion and the sum total of all γ-ray contributions are recorded. After acquiring data at one angular position, the detector is then rotated around the patient to a different position, so that data may be recorded at a different viewing angle. Typically, the detector is rotated over 360° in equally-spaced increments, so that an adequate sampling of the object may be obtained. This can be seen in Figure 3.

The FOV constraint isn't the only factor that influences the detection of a γ-ray. In SPECT imaging, photon attenuation, detector blurring and photon scatter also play an important role. Photon attenuation occurs when an exiting photon is stopped by the body's tissue and thus is unable to contribute to a given measurement. This results in a lower number of γ-rays being recorded than should be. Detector blurring affects the measurement process by spreading the contribution of a given γ-ray to several measurement locations on the detector. This stems from the fact that some γ-rays penetrate the septa and that the detector's electronics are imperfect, and so only *estimate* the position at which the γ-ray impinges on the detector. Finally, photons can be diverted (scattered) from their original exit path by interacting with the body's tissue and can be erroneously recorded in an incorrect measurement location. Accounting for these degrading factors in the modelling process is important for both quantitative and qualitative accuracy.

2.2 Model Details

Now that we know how an emitted γ-ray is measured, we must find a way to translate this physical process into a mathematical model. Remember that our goal is to obtain a 2D image of the distribution of radio-tracer within the patient,

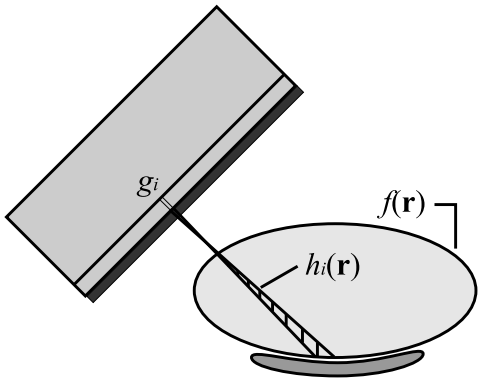

Figure 4: A schematic denoting the mathematical representations of the constituent pieces of the imaging process. The area where the sensitivity function is non-zero is shaded.

which shows the amount of radioactivity in the organs of interest. It is common when developing mathematical models to make simplifying assumptions in order to make the translation step easier, and so we will take this same approach.

First, we note that the patient (or object) is continuous in nature, while the data are measured at discrete locations on the detector. So, let the function $f(\mathbf{r})$ represent the concentration of radio-tracer at location $\mathbf{r}$ within the patient, where $\mathbf{r} = (x,y)$ is the 2D spatial coordinate vector. We will regard this as the model for the continuous 2D object. Also, let g_i represent the i^{th} detector measurement that is recorded. Furthermore, we will let $h_i(\mathbf{r})$ define the continuous 2D sensitivity function for the i^{th} measurement. The sensitivity function basically represents the response of the imaging system to the object. All three of these quantities can be seen in Figure 4. Note that for data measurement g_i, the subscript i incorporates both the spatial location on the detector face and the angular position of the detector in relation to the patient. The sensitivity function $h_i(\mathbf{r})$ is non-zero only in the shaded FOV defined by the i^{th} measurement location and is defined mathematically as

$$h_i(\mathbf{r}) = \begin{cases} 1 & \mathbf{r} \in \text{FOV} \\ 0 & \mathbf{r} \notin \text{FOV} \, . \end{cases} \tag{1}$$

One should note that this is a very simplistic definition for a sensitivity function, in that it does not account for such degrading factors as photon attenuation, detector blurring and scatter. But it is useful as a first approximation to the true response function for a SPECT system.

The measurement g_i will consist of the *sum* of all radioactivity in the object $f(\mathbf{r})$ within the area defined by the sensitivity function $h_i(\mathbf{r})$, as this is the only region in the object that can contribute to measurement g_i. Mathematically, this means that we need to multiply the object and the sensitivity function (which

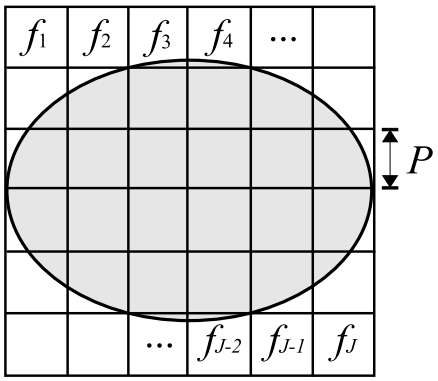

Figure 5: The continuous object $f(\mathbf{r})$ is discretized into a set of J pixels $\{f_j;\ j = 1,2,\ldots,J\}$. Each square pixel is of dimension P.

selects out that portion of the object that can contribute to measurement g_i), and then "add up" (or integrate) the result. Thus, the model for the data acquisition process is given by the double integral

$$g_i = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\mathbf{r})h_i(\mathbf{r})d\mathbf{r}\ . \tag{2}$$

At this point, it would be helpful to further simplify our model, as finding a simple mathematical definition for the FOV would be difficult. One way to accomplish this would be to discretize both the continuous object $f(\mathbf{r})$ and sensitivity functions $h_i(\mathbf{r})$ into an ordered set of *pixels* (picture elements) of square dimension P, which we denote by $\{f_j;\ j = 1,2,\ldots,J\}$ and $\{h_{i,j};\ j = 1,2,\ldots,J\}$, respectively. The object is easily discretized into pixels, by dividing it into a square grid. Each box within the grid represents a pixel and is indexed by j. For a given j, the value of f_j would be the average amount of radiotracer within that particular pixel. We show the discretization of the object in Figure 5. The discretization of the sensitivity functions $h_i(\mathbf{r})$ will be discussed momentarily.

Replacing the continuous quantities $f(\mathbf{r})$ and $h_i(\mathbf{r})$ with the discrete quantities f_j and $h_{i,j}$, respectively, in Eqn. (2) and replacing the integrals with a summation, the model for the data acquisition process reduces to

$$g_i = \sum_{j=1}^{J} h_{i,j}f_j\ . \tag{3}$$

Again, f_j is the average amount of radioactivity in pixel j, and $h_{i,j}$ represents the contribution of object pixel j to detector location i. If there are I total detector measurements for J total object pixels, we can arrange the data set

$\{g_i,\ i = 1,2,\ldots,I\}$ into vector format, yielding the following matrix equation as the model for the data acquisition process

$$\begin{pmatrix} g_1 \\ g_2 \\ \vdots \\ g_I \end{pmatrix} = \begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,J} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ h_{I,1} & h_{I,2} & \cdots & h_{I,J} \end{pmatrix} \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_J \end{pmatrix}, \tag{4}$$

or more succinctly, $\mathbf{g} = \mathbf{Hf}$. Thus, the image reconstruction problem simply reduces to solving this linear system of equations. In Eqn. (4), the vector $\mathbf{f}$ is the discrete approximation to the function $f(\mathbf{r})$ which represents the tracer concentration within the object, the vector $\mathbf{g}$ is the known measured data, and the matrix $\mathbf{H}$ is a model for the SPECT imaging system itself. We usually call $\mathbf{H}$ a *projector matrix*, as it it creates the projection data $\mathbf{g}$ from the object $\mathbf{f}$.

At this point, the imaging system must be modelled, meaning we must find a way to discretize the sensitivity functions $h_i(\mathbf{r})$ into $h_{i,j}$, thereby creating the projector matrix $\mathbf{H}$. Anton and Rorres [7] give several suggestions for computing $h_{i,j}$ for x-ray CT imaging, and we can use the same principles for SPECT. For simplicity, we shall use the length of the line segment originating at the center of detector location i that passes through pixel j, as our value for $h_{i,j}$. This can be seen in Figure 6. If pixel j doesn't contribute to measurement i, then $h_{i,j} = 0$.

The computation of $h_{i,j}$ when the detector is oriented at angular positions $\theta \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$ is elementary, as each pixel that contributes to a given measurement location will have a line segment length of P. This is due to the fact that the detector is orthogonal to the grid as can be seen in Figure 7.

To better illustrate this computation, let's begin with detector position $\theta = 0$ and consider measurement g_4. The 2D coordinates of the center of this measurement location are $x_4 = \frac{P}{2}$ and $y_4 = -R$, where the radius-of-rotation R is the distance from the detector face to the center of rotation. The equation of the line originating at the point $(x_4, y_4) = (\frac{P}{2}, -R)$ and passing through the object is $x = \frac{P}{2}$. The x-coordinates for each point of intersection with the grid lines are all $x = \frac{P}{2}$, while the y-coordinates are y = -3P, -2P, -P, 0, P, 2P and 3P, respectively as we travel up the ray. After determining the 2D coordinates of each point that intersects the grid, we can then applying the distance formula between successive points (x_α, y_α) and (x_β, y_β)

$$d = \sqrt{(x_\alpha - x_\beta)^2 + (y_\alpha - y_\beta)^2}\ , \tag{5}$$

in order to compute the length of each line segment. For pixel f_4, the intersection points are given by $(x_\alpha, y_\alpha) = (\frac{P}{2}, 3P)$ and $(x_\beta, y_\beta) = (\frac{P}{2}, 2P)$. Using Eqn. (5), we compute $h_{4,4}$ to be

$$\begin{aligned} h_{4,4} &= \sqrt{(\tfrac{P}{2} - \tfrac{P}{2})^2 + (3P - 2P)^2} \\ &= \sqrt{P^2} \\ &= P \end{aligned} \quad . \tag{6}$$

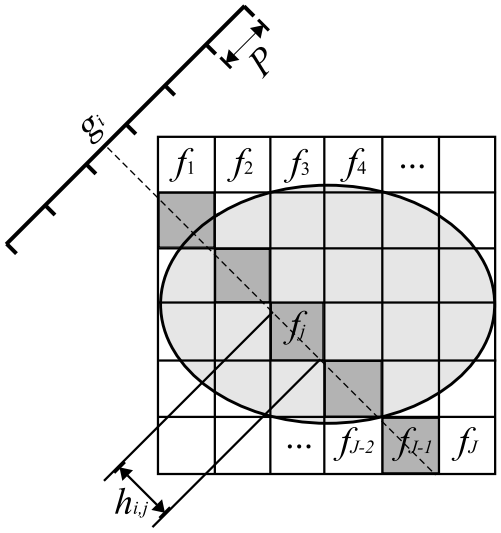

Figure 6: The discretization of the sensitivity functions $h_i(\mathbf{r})$ into $\{h_{i,j},\ j = 1,2,\ldots,J\}$. The value for a given $h_{i,j}$ is calculated as the length of the line segment originating at the center of detector location i that passes through pixel j. Note that the pixels that contribute to measurement g_i are more darkly shaded.

A similar computation shows that each line segment must have length P. Thus, the equation that models measurement g_4 is given by

$$\begin{aligned} g_4 &= h_{4,4}f_4 + h_{4,10}f_{10} + h_{4,16}f_{16} + h_{4,22}f_{22} + h_{4,28}f_{28} + h_{4,34}f_{34} \\ &= Pf_4 + Pf_{10} + Pf_{16} + Pf_{22} + Pf_{28} + Pf_{34}\ . \end{aligned} \tag{7}$$

It should be noted that no other pixels contribute to measurement g_4, and thus for those pixels, $h_{i,j} = 0$.

When the detector is not orthogonal to the grid, how do we compute the line segment lengths? This may seem like a complicated process, but in fact, we only need a little algebra and trigonometry. The answer is to trace the ray from the point of origin on the detector through the object using the equation of the line, find the intersection points along the grid, and then use the distance formula (Eqn. (5)) to compute the length of each line segment.

To further illustrate this, let's consider Figure 8, which shows the detector at arbitrary angular position θ and measurement g_i. The 2D coordinates of the center of this measurement point, which we denote by (x_i, y_i) are found by counter-clockwise rotation of the coordinates of the measurement point for g_4, denoted by (x_4, y_4), by θ radians using the rotation equations

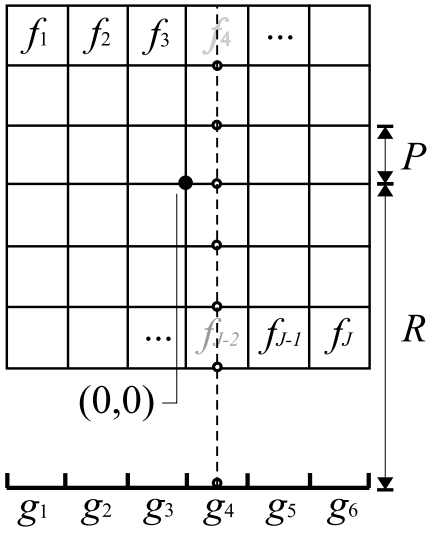

Figure 7: Determination of the lengths of the line segments for measurement g_4 using ray-tracing.

$$\begin{aligned} x_i &= x_4 \cos\theta - y_4 \sin\theta \\ y_i &= x_4 \sin\theta + y_4 \cos\theta \ . \end{aligned} \tag{8}$$

The slope of the line originating at point (x_i, y_i) and passing through the object is $m = -\tan(\frac{\pi}{2} - \theta)$ and thus the equation of the line is given by

$$\begin{aligned} y - y_i &= m(x - x_i) \\ &= -\tan(\tfrac{\pi}{2} - \theta)(x - x_i) \ . \end{aligned} \tag{9}$$

Next, we trace along the ray from the point of origin (x_i, y_i) to determine the first point of intersection with the grid, which occurs in pixel $J-1$. Note that the y-coordinate of this point given by $y = -3P$. So, we simply solve Eqn. (9) for x and substitute $y = -3P$ to obtain the x-coordinate

$$x = \frac{-3\,P - y_i}{-\tan(\frac{\pi}{2} - \theta)} + x_i \ . \tag{10}$$

Every other point of intersection along the ray is obtained in a similar fashion. We always know either the x- or y-coordinate of each point, as they occur on the grid. By solving Eqn. (9) for the unknown coordinate, we can find all of the 2D points of intersection. Once all of the points are know for a given measurement, the distance formula (Eqn. (5)) can be applied to find the line segment lengths and thereby compute the values of $\{h_{i,j};\ j = 1,2,\ldots,J\}$. The equation that governs measurement g_i is therefore

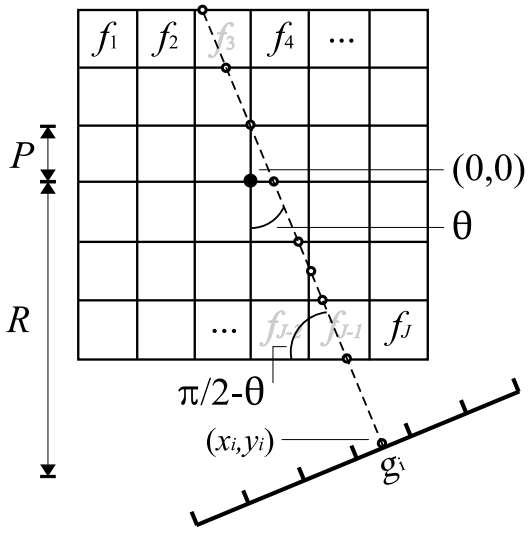

Figure 8: Determination of the lengths of the line segments for measurement g_i using ray-tracing.

$$g_i = h_{i,3}f_3 + h_{i,9}f_9 + h_{i,16}f_{16} + h_{i,22}f_{22} + h_{i,28}f_{28} + h_{i,29}f_{29} + h_{i,35}f_{35} \ . \quad (11)$$

Again, we emphasize that no other pixels contribute to measurement g_i, and thus for those pixels, $h_{i,j} = 0$.

2.2.1 Example

Consider the imaging scenario given in Figure 9. Assume the object $\mathbf{f}$ has been discretized on a 3×3 square grid into $\{f_j; j = 1, 2, \ldots, 9\}$ with $P = 1.0$ and data will be acquired at three projection angles equally spaced from 0 to 2π with $R = 3\sqrt{2}$, giving us $\{g_i; i = 1, 2, \ldots, 9\}$. Our goal is to derive the corresponding projector matrix $\mathbf{H}$ for this tomographic system.

First, we note that the data are acquired at three projection angles, $\theta \in \{0, \frac{2\pi}{3}, \frac{4\pi}{3}\}$. Because measurements g_1, g_2 and g_3 are acquired at $\theta = 0$, we further note that the corresponding line segments are of length P, as the detector is orthogonal to the grid for that acquisition angle. Thus, the equations that govern these measurements are given by

$$\begin{array}{rcl} g_1 & = & 1f_1 + 1f_4 + 1f_7 \\ g_2 & = & 1f_2 + 1f_5 + 1f_8 \\ g_3 & = & 1f_3 + 1f_6 + 1f_9 \end{array} \ . \quad (12)$$

The detector is then rotated counter-clockwise to acquisition angle $\theta = \frac{2\pi}{3}$. To find the spatial coordinates on the detector for measurements g_4, g_5 and g_6,

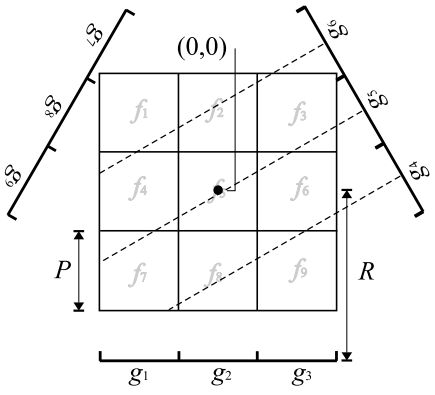

Figure 9: Example of a SPECT imaging scenario with a 3×3 discretized object and corresponding projection data measurements.

we must rotate the spatial coordinates for detector measurements g_1, g_2 and g_3, respectively, by $\theta = \frac{2\pi}{3}$ radians using Eqn. (8). The spatial coordinates for g_1, g_2 and g_3 are given, respectively, as (-1,-3$\sqrt{2}$), (0,-3$\sqrt{2}$) and (1,-3$\sqrt{2}$). When properly rotated, we find the spatial coordinates for g_4, g_5 and g_6 to be, respectively, (4.1742,1.2553), (3.6742,2.1213) and (3.1742,2.9873). The slopes of the lines originating at each of these detector locations is $m = \cot(\pi - \frac{2\pi}{3}) = \frac{1}{\sqrt{3}}$. Thus, the equations of the lines for g_4, g_5 and g_6 are given, respectively, as

$$\begin{array}{rcl} y - 1.2553 & = & \frac{1}{\sqrt{3}}(x - 4.1742) \\ y - 2.1213 & = & \frac{1}{\sqrt{3}}(x - 3.6742) \\ y - 2.9873 & = & \frac{1}{\sqrt{3}}(x - 3.1742) \end{array} . \qquad (13)$$

Each point of intersection occurs on the grid and so possesses either x-coordinates of $\pm\frac{1}{2}$, $\pm\frac{3}{2}$ or y-coordinates of $\pm\frac{1}{2}$, $\pm\frac{3}{2}$. Using the equation of the line for measurement g_4, we compute the points of intersection and the lengths of the line segments between them, by moving from the point of origin on the detector along the ray through the object. The first intersection point occurs at $x = 1.5$, the second at $y = -0.5$, and so on. We list all of the intersection points in the table below as they sequentially occur, so that the lengths of the line segments can easily be computed by using Eqn. (5) on the sequential points. Thus, we obtain

x	y	$h_{i,j}$
1.5	−0.2887	−
1.1339	−0.5	0.4227
0.5	−0.866	0.732
−0.5	−1.4434	1.1547
−0.5981	−1.5	0.1133

.

The equation that governs measurement g_4 is thus

$$g_4 = 0.4227f_6 + 0.1133f_7 + 1.1547f_8 + 0.732f_9 \quad . \tag{14}$$

Continuing in the same matter, we find the points of intersection for measurement g_5 and the lengths of the line segments between them to be

$$\begin{array}{rrr} x & y & h_{i,j} \\ 1.5 & 0.866 & - \\ 0.866 & 0.5 & 0.7321 \\ 0.5 & 0.2887 & 0.4226 \\ -0.5 & -0.2887 & 1.1547 \\ -0.866 & -0.5 & 0.4226 \\ -1.5 & -0.866 & 0.7321 \end{array} ,$$

and so the equation that governs measurement g_5 is

$$g_5 = 0.732f_3 + 0.4226f_4 + 1.1547f_5 + 0.4226f_6 + 0.7321f_7 \quad . \tag{15}$$

Finally, we compute the intersection points and line segment lengths for g_6 as

$$\begin{array}{rrr} x & y & h_{i,j} \\ 0.5981 & 1.5 & - \\ 0.5 & 1.4433 & 0.1133 \\ -0.5 & 0.866 & 1.1547 \\ -1.1339 & 0.5 & 0.732 \\ -1.5 & 0.2886 & 0.4227 \end{array} ,$$

and the resulting governing equation for g_6 to be

$$g_6 = 0.732f_1 + 1.1547f_2 + 0.1133f_3 + 0.4227f_4 \quad . \tag{16}$$

We can proceed in this same manner for computing the governing equations for the detector oriented at acquisition angle $\theta = \frac{4\pi}{3}$. Alternatively, we can use the symmetry between detector positions $\theta = \frac{2\pi}{3}$ and $\theta = \frac{4\pi}{3}$ to more easily derive these governing equations. We simply reflect the detector and projection rays relative to the line $x = 0$. Taking this symmetry into account, we find the governing equations to be

$$\begin{array}{rcl} g_7 & = & 0.1133f_1 + 1.1547f_2 + 0.732f_3 + 0.4227f_6 \\ g_8 & = & 0.732f_1 + 0.4226f_4 + 1.1547f_5 + 0.4226f_6 + 0.7321f_9 \\ g_9 & = & 0.4227f_4 + 0.732f_7 + 1.1547f_8 + 0.1133f_9 \end{array} \quad . \tag{17}$$

Using Eqns. (12) and (14)-(17), we construct the projector matrix $\mathbf{H}$ for this tomographic system as

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0.4227 & 0.1133 & 1.1547 & 0.732 \\ 0 & 0 & 0.732 & 0.4226 & 1.1547 & 0.4226 & 0.7321 & 0 & 0 \\ 0.732 & 1.1547 & 0.1133 & 0.4227 & 0 & 0 & 0 & 0 & 0 \\ 0.1133 & 1.1547 & 0.732 & 0 & 0 & 0.4227 & 0 & 0 & 0 \\ 0.732 & 0 & 0 & 0.4226 & 1.1547 & 0.4226 & 0 & 0 & 0.732 \\ 0 & 0 & 0 & 0.4227 & 0 & 0 & 0.732 & 1.1547 & 0.1133 \end{pmatrix}. \tag{18}$$

Although this process is a bit computationally intensive, it is not difficult to understand conceptually. However, modelling a system "by hand" is prohibitive when the number of pixels J and/or the number of measurements I is large.

In such cases, we can take advantage of the computer to perform such calculations. We will see that the MATLAB function `project2d.m` will easily compute projections of any 2D object for a given number of projection angles, equally spaced in the interval $[0, 2\pi]$. Instead of rotating the detector and computing the lengths of the line segments of intersection, `project2d` rotates the object (clockwise), interpolates the pixel values back to a rectangular grid, and then computes the projections by summing along straight rays whose line segments are all unit length.

2.3 Solution Methodology and Implementation

Thus far, we have presented a physical problem in SPECT and then discussed how such a problem is modelled mathematically. We found that a SPECT data acquisition can be approximated by a linear system of equations. Now, we focus on the mathematics of solving such a linear system, which we know to be equivalent to reconstructing a SPECT image from the data.

2.3.1 Determinancy of the Linear System

Before we begin to select a possible solution methodology, it would be useful to think about the number of solutions that might exist for our linear system of equations. From our study of linear algebra, we know that a given system could have 1) a unique solution, 2) no solution, or 3) an infinite number of solutions. Linear systems for which no solutions exist are called *inconsistent*, while a system that has at least one solution (and possibly an infinite number) is called *consistent*.

Recall that for our linear system given by Eqn. (4), $\mathbf{H}$ has dimension $I \times J$, $\mathbf{f}$ is $J \times 1$, and $\mathbf{g}$ is $I \times 1$, where I is the number of equations (data measurements) and J is the number of unknowns (image pixels). From a dimensionality standpoint, there are three possible determinancies for our linear system:

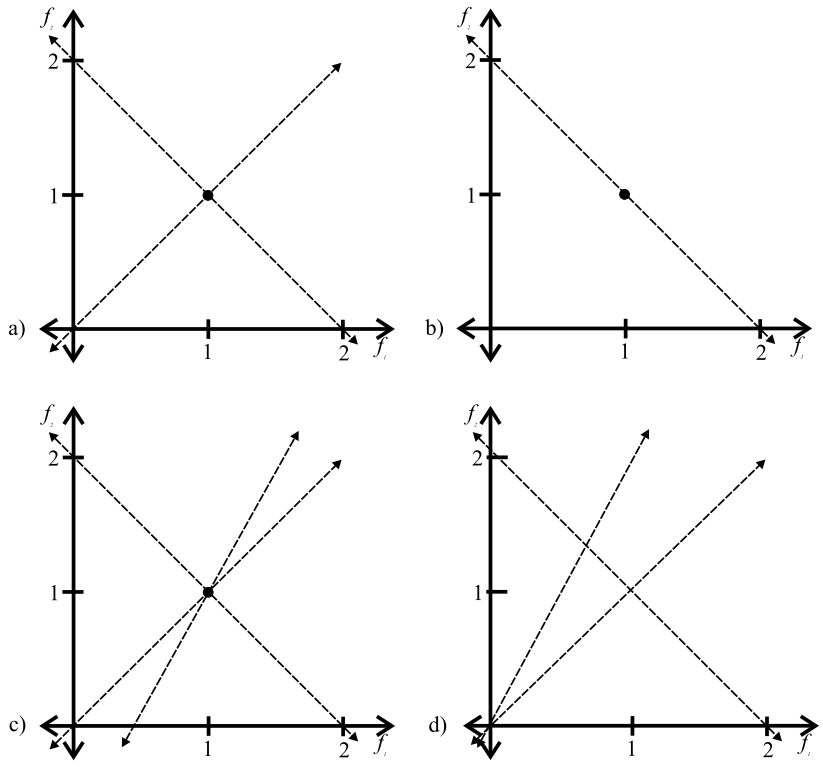

Figure 10: Examples of linear systems with different determinancy: a) even-determined (unique solution); b) under-determined (infinite number of solutions); c) over-determined (unique solution); d) over-determined (no solution).

- When $I = J$, the system is called *even-determined*. The number of equations is equal to the number of unknowns and so there are the same number of data measurements and object pixels. In this case, $\mathbf{H}$ is a square matrix.
- When $I < J$, the system is called *under-determined*. The number of equations is less than the number of unknowns and so there are fewer data measurements than object pixels. In this case, $\mathbf{H}$ is a rectangular matrix with more columns than rows.
- When $I > J$, the system is called *over-determined*. The number of equations is greater than the number of unknowns and so there are more data measurements than object pixels. In this case, $\mathbf{H}$ is a rectangular matrix with more rows than columns.

We now give examples of each type of linear system, along with their respective geometrical interpretations in Figure 10. An example of an even-determined system might be

$$\begin{array}{rcrcl} -f_1 & + & f_2 & = & 0 \\ f_1 & + & f_2 & = & 2\,. \end{array} \tag{19}$$

In this case, the number of equations is the same as the number of unknowns ($I = J = 2$) and the unique solution to the linear system is $\mathbf{f} = (1,1)^t$. An under-determined system such as

$$f_1 \quad + \quad f_2 \quad = \quad 2\,, \tag{20}$$

has fewer equations than unknowns. In this example, $I = 1 < J = 2$. One should note that there is no unique solution to this linear system. Although $\mathbf{f} = (1,1)^t$ is a solution, any point that lies along the line given by Eqn. (20) is a solution as well. Thus, there are an infinite number of solutions. When a linear system is over-determined, there are more equations than unknowns. A unique solution may exist, such as for the linear system given by

$$\begin{array}{rcrcr} -f_1 & + & f_2 & = & 0 \\ f_1 & + & f_2 & = & 2 \\ -2f_1 & + & f_2 & = & -1 \end{array}\,, \tag{21}$$

where $I = 3 > J = 2$. Here the unique solution to Eqn. (21) is $\mathbf{f} = (1,1)^t$. However, more than likely, we could also have an over-determined system with no solution, such as

$$\begin{array}{rcrcl} -f_1 & + & f_2 & = & 0 \\ f_1 & + & f_2 & = & 2 \\ -2f_1 & + & f_2 & = & 0\,. \end{array} \tag{22}$$

No unique solution exists for this system, because the equations are inconsistent (i.e., the lines don't intersect at a single point). However, we can obtain a type of solution to this system that in some manner "best fits" all of the equations.

The solution methodologies we shall consider for this module will be referred to as *linear iterative* inversion methods, since they iteratively generates a sequence of vectors $\{\hat{\mathbf{f}}^{(k)}\}_{k=0}^{\infty}$ that ideally converge to a solution $\hat{\mathbf{f}}$

$$\lim_{k\to\infty} \hat{\mathbf{f}}^{(k)} = \hat{\mathbf{f}}\,. \tag{23}$$

We can express such iteration schemes generically using the general linear iterative algorithm (GLIA)

$$\hat{\mathbf{f}}^{(k+1)} = \hat{\mathbf{f}}^{(k)} + \alpha \mathbf{O}(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(k)})\,, \tag{24}$$

where k is the iteration number. Operationally, one re-projects the k^{th} estimate of the object denoted by $\hat{\mathbf{f}}^{(k)}$, forms the difference between the measured projections and the re-projections, "reconstructs" these difference projections with the matrix operator $\mathbf{O}$ weighted by the scalar α, and then finally adds this "correction image" to the k^{th} estimate to form the $(k+1)^{st}$ estimate of the

object, denoted by $\hat{\mathbf{f}}^{(k+1)}$. The scalar α is a weight that controls how much the correction image contributes to the $(k+1)^{st}$ estimate of the object.

Equation (24) is also known as a *fixed-point* iteration scheme. If convergence to a fixed point $\hat{\mathbf{f}}$ is obtained, this implies that $\mathbf{g} = \mathbf{H}\hat{\mathbf{f}}$ and so the correction image becomes the zero vector.

In some cases, our estimate $\hat{\mathbf{f}}$ will in fact be the exact solution to the linear system, and so $\hat{\mathbf{f}} = \mathbf{f}$. When no exact solution exists, we will simply call $\hat{\mathbf{f}}$ our approximate solution to the linear system. Equation (24) expresses a standard methodology used to iteratively solve linear systems of equations from numerical linear algebra. While this methodology is theoretically attractive, there may be aspects of the problem that preclude a particular method from converging to a solution $\hat{\mathbf{f}}$. We shall now see that convergence is directly related to the projector matrix $\mathbf{H}$, the "reconstruction" matrix $\mathbf{O}$ and the parameter α.

To understand how convergence can be predicted, we begin by converting our linear system $\mathbf{g} = \mathbf{H}\mathbf{f}$ into the form $\mathbf{f} = \mathbf{T}\mathbf{f} + \mathbf{c}$. Writing this as an iterative relation, we get

$$\hat{\mathbf{f}}^{(k+1)} = \mathbf{T}\hat{\mathbf{f}}^{(k)} + \mathbf{c} \; . \tag{25}$$

An important theorem regarding the convergence of a sequence of vectors $\{\hat{\mathbf{f}}^{(k)}\}_{k=0}^{\infty}$ generated via Eqn. (25) can now be invoked [5].

Theorem 1 *For any J-dimensional real vector $\hat{\mathbf{f}}^{(0)}$, the sequence $\{\hat{\mathbf{f}}^{(k)}\}_{k=0}^{\infty}$ defined by*

$$\hat{\mathbf{f}}^{(k+1)} = \mathbf{T}\hat{\mathbf{f}}^{(k)} + \mathbf{c} \; , \tag{26}$$

converges to the unique solution of $\mathbf{f} = \mathbf{T}\mathbf{f} + \mathbf{c}$ if and only if $\rho(\mathbf{T}) < 1$.

The proof of this theorem is given in [5]. The vector $\hat{\mathbf{f}}^{(0)}$ represents the initial guess, while the expression $\rho(\mathbf{T})$ refers to the *spectral radius* of the matrix $\mathbf{T}$, which is defined to be

$$\rho(\mathbf{T}) = \max |\lambda| \; , \tag{27}$$

where λ is an eigenvalue of $\mathbf{T}$. Thus, to predict convergence of a GLIA, we first need to rewrite Eqn. (24) in the form of Eqn. (25)

$$\begin{aligned} \hat{\mathbf{f}}^{(k+1)} &= \hat{\mathbf{f}}^{(k)} + \alpha\mathbf{O}(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(k)}) \\ &= \hat{\mathbf{f}}^{(k)} + \alpha\mathbf{O}\mathbf{g} - \alpha\mathbf{O}\mathbf{H}\hat{\mathbf{f}}^{(k)} \\ &= (\mathbf{I} - \alpha\mathbf{O}\mathbf{H})\hat{\mathbf{f}}^{(k)} + \alpha\mathbf{O}\mathbf{g} \\ &= \mathbf{T}\hat{\mathbf{f}}^{(k)} + \mathbf{c} \end{aligned} \; . \tag{28}$$

It is then easy to identify the composition of the matrix $\mathbf{T}$ and vector $\mathbf{c}$

$$\begin{aligned} \mathbf{T} &= (\mathbf{I} - \alpha\mathbf{O}\mathbf{H}) \\ \mathbf{c} &= \alpha\mathbf{O}\mathbf{g} \end{aligned} \; , \tag{29}$$

and to then check to see if $\rho(\mathbf{T}) < 1$. If $\mathbf{T}$ has this property, then our GLIA will converge. From Eqn. (29), we can clearly see how the selection of the "reconstruction" matrix $\mathbf{O}$ and weighting parameter α play a significant role in the convergence of our GLIA.

Although Eqn. (24) can be used to implement our iterative scheme, we can use an equivalent formulation given by

$$\begin{array}{rcl} \hat{\mathbf{f}}^{(k+1)} & = & \sum_{k'=0}^{k} (\mathbf{I} - \alpha\mathbf{OH})^{k'} \alpha\mathbf{Og} + (\mathbf{I} - \alpha\mathbf{OH})^{(k+1)} \hat{\mathbf{f}}^{(0)} \\ & = & \mathbf{B}^{(k)}\mathbf{g} + \mathbf{Q}^{(k)}\hat{\mathbf{f}}^{(0)} \ , \end{array} \tag{30}$$

where

$$\begin{array}{rcl} \mathbf{B}^{(k)} & = & \sum_{k'=0}^{k} (\mathbf{I} - \alpha\mathbf{OH})^{k'} \alpha\mathbf{O} \\ \mathbf{Q}^{(k)} & = & (\mathbf{I} - \alpha\mathbf{OH})^{(k+1)} \ . \end{array} \tag{31}$$

The proof of this equivalence can be shown by induction. We note that the reconstruction is a function of the data and the initial estimate, unless $\mathbf{Q}^{(k)}$ is zero. The advantage to this formulation is that after computing the matrices $\mathbf{B}^{(k)}$ and $\mathbf{Q}^{(k)}$, we only need to use matrix-vector multiplication in order to compute $\hat{\mathbf{f}}^{(k+1)}$.

Now that we know how to predict the convergence of a GLIA, we need to choose which methods we wish to study. In this module, we will study two standard methods typically presented in a numerical linear algebra course, as well as two non-standard GLIAs:

- Jacobi method
- Gauss-Seidel method
- Landweber method
- sSimultaneous Iterative Reconstruction Technique

The details of the Jacobi and Gauss-Seidel methods can be found in the text by Burden and Faires [5], while the other methods are best described in the paper by Xu *et.al.* [6].

2.3.2 Jacobi Method

The Jacobi (JAC) method is an iterative technique that can only be used for even-determined linear systems of equations, and so $I = J$. The idea is to rewrite each equation in the linear system by solving for the respective variable for that particular equation. So, given the $J \times J$ linear system

$$\begin{array}{ccccccccccc} h_{1,1}f_1 & + & h_{1,2}f_2 & + & h_{1,3}f_3 & + & \ldots & + & h_{1,J}f_J & = & g_1 \\ h_{2,1}f_1 & + & h_{2,2}f_2 & + & h_{2,3}f_3 & + & \ldots & + & h_{2,J}f_J & = & g_2 \\ \vdots & & \vdots & & \vdots & & & & \vdots & & \vdots \\ h_{J,1}f_1 & + & h_{J,2}f_2 & + & h_{J,3}f_3 & + & \ldots & + & h_{J,J}f_J & = & g_J \end{array} \ , \tag{32}$$

we solve the jth equation for f_j

$$\begin{array}{rcl} f_1 & = & \frac{-h_{1,2}f_2 - h_{1,3}f_3 - \ldots - h_{1,J}f_J + g_1}{h_{1,1}} \\ f_2 & = & \frac{-h_{2,1}f_1 - h_{2,3}f_3 - \ldots - h_{2,J}f_J + g_2}{h_{2,2}} \\ \vdots & & \vdots \\ f_J & = & \frac{-h_{J,1}f_1 - h_{J,2}f_2 - \ldots - h_{J,J-1}f_{J-1} + g_J}{h_{J,J}} \end{array} \quad . \tag{33}$$

Note that this requires that $h_{j,j} \neq 0$ for each $j = 1, 2, \ldots, J$. If this is not the case, the equations must be re-ordered so that each pivot is non-zero. This can now be turned into an iterative relation by writing each f_j as a function of iteration k

$$\begin{array}{rcl} f_1^{(k+1)} & = & \frac{-h_{1,2}f_2^{(k)} - h_{1,3}f_3^{(k)} - \ldots - h_{1,J}f_J^{(k)} + g_1}{h_{1,1}} \\ f_2^{(k+1)} & = & \frac{-h_{2,1}f_1^{(k)} - h_{2,3}f_3^{(k)} - \ldots - h_{2,J}f_J^{(k)} + g_2}{h_{2,2}} \\ \vdots & & \vdots \\ f_J^{(k+1)} & = & \frac{-h_{J,1}f_1^{(k)} - h_{J,2}f_2^{(k)} - \ldots - h_{J,J-1}f_{J-1}^{(k)} + g_J}{h_{J,J}} \end{array} \quad . \tag{34}$$

But Eqn. (34) can be written much more succinctly using matrices. To accomplish this, we decompose the projector matrix $\mathbf{H}$ into an upper-triangular part denoted by $\mathbf{U}$

$$\mathbf{U} = \begin{pmatrix} 0 & -h_{1,2} & \cdots & -h_{1,J} \\ 0 & 0 & \cdots & -h_{2,J} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} , \tag{35}$$

a lower-triangular part denoted by $\mathbf{L}$

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ -h_{2,1} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -h_{J,1} & -h_{J,2} & \cdots & 0 \end{pmatrix} , \tag{36}$$

and a diagonal part denoted by $\mathbf{D}$

$$\mathbf{D} = \begin{pmatrix} h_{1,1} & 0 & \cdots & 0 \\ 0 & h_{2,2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_{J,J} \end{pmatrix} . \tag{37}$$

Noting that $\mathbf{H} = \mathbf{D}$ - $\mathbf{L}$ - $\mathbf{U}$, we rewrite our linear system as

$$\begin{aligned} \mathbf{Hf} &= \mathbf{g} \\ (\mathbf{D}-(\mathbf{L}+\mathbf{U}))\mathbf{f} &= \mathbf{g} \\ \mathbf{Df} &= (\mathbf{L}+\mathbf{U})\mathbf{f}+\mathbf{g} \\ \mathbf{f} &= \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{f}+\mathbf{D}^{-1}\mathbf{g} \end{aligned} \quad , \tag{38}$$

which can then be turned into an iterative relation by including dependence of the f_j on iteration k

$$\mathbf{f}^{(k+1)} = \mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})\mathbf{f}^{(k)}+\mathbf{D}^{-1}\mathbf{g} \quad . \tag{39}$$

It will be left to the student to show that Eqn. (39) can be expressed in GLIA form as

$$\hat{\mathbf{f}}^{(k+1)} = \hat{\mathbf{f}}^{(k)} + \mathbf{D}^{-1}(\mathbf{g}-\mathbf{H}\hat{\mathbf{f}}^{(k)}) \; . \tag{40}$$

(Hint: Add and subtract $\hat{\mathbf{f}}^{(k)}$ on the right-hand side.) Thus, for the Jacobi method, we find the "reconstruction" matrix to be $\mathbf{O} = \frac{1}{\alpha}\mathbf{D}^{-1}$. It should be noted that the scalar weight α plays no role in the iteration scheme for the JAC method as shown by Eqn. (40).

2.3.3 Gauss-Seidel Method

The Gauss-Seidel (G-S) method is similar to that of Jacobi, as it can only be used for even-determined linear systems. The difference stems from the fact that as soon as the updated value $f_j^{(k+1)}$ is computed in equation j, it is immediately used in the subsequent equations to compute the new values for $\{f_{j+1}^{(k)}, f_{j+2}^{(k)}, \ldots, f_J^{(k)}\}$. This means that Eqn. (34) can be re-expressed as

$$\begin{aligned} f_1^{(k+1)} &= \frac{-h_{1,2}f_2^{(k)}-h_{1,3}f_3^{(k)}-\ldots-h_{1,J}f_J^{(k)}+g_1}{h_{1,1}} \\ f_2^{(k+1)} &= \frac{-h_{2,1}f_1^{(k+1)}-h_{2,3}f_3^{(k)}-\ldots-h_{2,J}f_J^{(k)}+g_2}{h_{2,2}} \\ &\vdots \qquad\qquad\qquad \vdots \\ f_J^{(k+1)} &= \frac{-h_{J,1}f_1^{(k+1)}-h_{J,2}f_2^{(k+1)}-\ldots-h_{J,J-1}f_{J-1}^{(k+1)}+g_J}{h_{J,J}} \end{aligned} \quad . \tag{41}$$

As with the JAC method, we require that $h_{j,j} \neq 0$ for each $j = 1, 2, \ldots, J$. If this is not the case, the equations must be re-ordered so that each pivot is non-zero. We also see that Eqn. (41) can be written more succinctly using matrices. Using the same decomposition of $\mathbf{H}$ as in the JAC method, we write

$$\begin{aligned} \mathbf{Hf} &= \mathbf{g} \\ ((\mathbf{D}-\mathbf{L})-\mathbf{U})\mathbf{f} &= \mathbf{g} \\ (\mathbf{D}-\mathbf{L})\mathbf{f} &= \mathbf{Uf}+\mathbf{g} \\ \mathbf{f} &= (\mathbf{D}-\mathbf{L})^{-1}\mathbf{Uf}+(\mathbf{D}-\mathbf{L})^{-1}\mathbf{g} \end{aligned} \quad , \tag{42}$$

which we turn into an iterative relation by including dependence of the f_j on iteration k

$$\mathbf{f}^{(k+1)} \quad = \quad (\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}\mathbf{f}^{(k)} + (\mathbf{D}-\mathbf{L})^{-1}\mathbf{g} \quad . \tag{43}$$

It will be left to the student to show that Eqn. (43) can be expressed in GLIA form as

$$\hat{\mathbf{f}}^{(k+1)} = \hat{\mathbf{f}}^{(k)} + (\mathbf{D}-\mathbf{L})^{-1}(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(k)}) \; . \tag{44}$$

(Hint: Add and subtract $\hat{\mathbf{f}}^{(k)}$ on the right-hand side.) Thus, for the G-S method, we find the "reconstruction" matrix to be $\mathbf{O} = \frac{1}{\alpha}(\mathbf{D}-\mathbf{L})^{-1}$. We note that the scalar weight α also plays no role in the iteration scheme for the G-S method as shown by Eqn. (44).

2.3.4 Landweber Method

The Landweber (LWB) method [11] does not have the restriction of even-determinancy possessed by both the JAC and G-S methods. Thus, it can be used for under-, even- and over-determined systems. It's derivation is complicated, and so we will attempt to give the student a basic idea of its origins.

The derivation of the algorithm is motivated by attempting to find an $\mathbf{f}$ that minimizes the objective function

$$\begin{aligned} O(\mathbf{f}) &= \|\mathbf{H}\mathbf{f} - \mathbf{g}\|_2^2 \\ &= (\mathbf{H}\mathbf{f} - \mathbf{g})^t(\mathbf{H}\mathbf{f} - \mathbf{g}) \\ &= \mathbf{f}^t\mathbf{H}^t\mathbf{H}\mathbf{f} - 2\mathbf{f}^t\mathbf{H}^t\mathbf{g} + \mathbf{g}^t\mathbf{g} \end{aligned} \quad . \tag{45}$$

From calculus, we know that the minimum of a function is found by finding the critical points, and using the second derivative, we can conclude if the point is indeed a minimum. Thus, taking the derivative of Eqn. (45) with respect to $\mathbf{f}$

$$\frac{\partial}{\partial \mathbf{f}} O(\mathbf{f}) = 2\mathbf{H}^t\mathbf{H}\mathbf{f} - 2\mathbf{H}^t\mathbf{g} \; , \tag{46}$$

and setting the result equal to zero, we obtain the *normal equations*

$$\mathbf{H}^t\mathbf{H}\mathbf{f} = \mathbf{H}^t\mathbf{g} \; . \tag{47}$$

A solution to Eqn. (47) can the be obtained iteratively [10] using the fixed-point algorithm given by

$$\hat{\mathbf{f}}^{(k+1)} = \hat{\mathbf{f}}^{(k)} + \alpha\mathbf{H}^t(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(k)}); \;\; \hat{\mathbf{f}}^{(0)} = \mathbf{0} \; , \tag{48}$$

which is known as the LWB method. Thus, for this method, we find the "reconstruction" matrix to be $\mathbf{O} = \mathbf{H}^t$.

2.3.5 Simultaneous Iterative Reconstruction Technique

Like the LWB method, the simultaneous iterative reconstruction technique (SIRT) does not have the restriction regarding determinancy possessed by both the JAC and G-S methods. Thus, it can also be used for under-, even- and over-determined systems. Its form is similar to that for the LWB method. The difference lies in that the "reconstruction" matrix $\mathbf{H}^t$ is normalized so that the columns sum to 1. This is accomplished by defining the matrix $\mathbf{S}$ as

$$(\mathbf{S})_{i,j} = \begin{cases} (\sum_{i=1}^{I} h_{i,j})^{-1} & i = j \\ 0 & i \neq j \ . \end{cases} \tag{49}$$

The iteration scheme then written as

$$\hat{\mathbf{f}}^{(k+1)} = \hat{\mathbf{f}}^{(k)} + \alpha \mathbf{S}\mathbf{H}^t(\mathbf{g} - \mathbf{H}\hat{\mathbf{f}}^{(k)}) \ , \tag{50}$$

which is known as the SIRT method. Thus, for this method, we find the "reconstruction" matrix to be $\mathbf{O} = \mathbf{S}\mathbf{H}^t$.

2.4 Assessment of the Model

Thus far, we've discussed the details of the SPECT imaging process and how it can be modelled mathematically. After making some simplifications, we showed that the imaging process could be modelled as a linear system of equations. We then presented four different solution methodologies from the theory of linear algebra, which allow us to solve the linear system, and are equivalent to reconstructing a SPECT image. Now that we know how to compute a solution, we need to evaluate the effectiveness of the model and, if possible, the accuracy of the solution.

The model expressed in Eqn. (4) is an mathematical approximation to the actual physical imaging process, with the matrix $\mathbf{H}$ representing the imaging system itself. It is impossible to compare the matrix $\mathbf{H}$ to a physical device. However, we do possess the measured projection data $\mathbf{g}$ that we obtained from our imaging system. And from the data, we computed a reconstructed image $\hat{\mathbf{f}}^{(k)}$. By applying our imaging model $\mathbf{H}$ to the reconstructed image, we can compute the data $\hat{\mathbf{g}}$ that our model imaging system would predict

$$\hat{\mathbf{g}} = \mathbf{H}\hat{\mathbf{f}}^{(k)} \ . \tag{51}$$

We can then compare the measured data to the predicted data using the root-mean square (RMS) error

$$RMS_{data} = \sqrt{\frac{1}{I}\sum_{i=1}^{I}(g_i - \hat{g}_i)^2} \ , \tag{52}$$

which is related to the Euclidean distance between the two I-dimensional vectors. This will allow us to evaluate the effectiveness of out model. If RMS_{data} = 0, then the data can be fit exactly by the model, meaning the real imaging

Figure 11: Simulated images of a uniform disk object (left) and MCAT activity map (right). The MCAT is provided courtesy of Dr. B.M.W. Tsui of the Department of Biomedical Engineering, University of North Carolina, Chapel Hill.

system and our model both produce the same data set. Large RMS errors imply mismatches between the given quantities being compared.

The model for the imaging system, given by the projector matrix $\mathbf{H}$ can, in a sense, be assessed by examining its eigenvalue spectrum. The matrix $\mathbf{T}$ in Eqn. (28) is a square matrix and so possesses an eigenvalue/vector decomposition. In fact, we require that $\rho(\mathbf{T}) < 1$ in order for our fixed-point iteration scheme to converge. Thus, one way to assess the effectiveness of our model $\mathbf{H}$ and our choice of reconstruction matrix $\mathbf{O}$ is to be sure this convergence criterion is satisfied.

Evaluating the effectiveness of the solution is much harder for two main reasons. First, the actual object is a continuous function $f(\mathbf{r})$ and the reconstructed image is a discrete set of pixels. Second, we don't know what the object really is, for if we did, we wouldn't have gone to all of the trouble of making a reconstructed image! But researchers sometimes use *simulations* in order to test a theoretical imaging system, such as the one we have modelled. We show examples of such simulated objects in Figure 11. On the left is a disk with uniform activity level within its interior, while on the right is an image of a mathematical cardiac torso phantom (MCAT) with primary activity uptake in the walls of the heart. In such cases, the object $\mathbf{f}$ is known, and so it can be used for comparison purposes with the reconstruction $\hat{\mathbf{f}}^{(k)}$, again using the RMS error

$$RMS_{object} = \sqrt{\frac{1}{J}\sum_{j=1}^{J}(f_j - \hat{f}_j^{(k)})^2}\ . \tag{53}$$

A $RMS_{object} = 0$ means that the true object can exactly be recovered from the data. Again, large RMS errors imply mismatches between the given quantities being compared.

2.5 Conceptual Questions

1. How small/large can the radius-of-rotation R be? How does this relate to the size (or support) of the object?
2. What are the physical and/or mathematical constraints on $\mathbf{f}$, $\mathbf{g}$, and $\mathbf{H}$?
3. How will the discretization process affect the reconstructed image, both visually and quantitatively?
4. What factor(s) influenced the determinancy (over-, under-, or even-determined) of the linear system?
5. For each type of determinancy, discuss how many solutions could possibly exist (none, an infinite number, exactly one).
6. What solution methods could apply in each determinancy case? Which practical considerations must be accounted for?

3 Ancillary Materials

3.1 Glossary of Terms

Anger camera The specific name for the SPECT detector, named for Hal Anger.

collimator A device, similar to a honeycomb, attached to the front of the Anger camera in order to restrict the path that detectable photons can travel.

field of view (FOV) The path, defined by the septa of the collimator, that detectable photons can travel.

gamma ray A high energy particle of light (photon) used in SPECT imaging.

image A 2D discrete representation of an object, consisting of pixels.

image reconstruction The process by which measured projection data are converted into a 2D image.

medical imaging The field which deals with developing images of the interior of the body without invasive techniques being used.

optical photon A particle of light in the visible range of the electro-magnetic spectrum.

photo-multiplier tube A device which views optical photons produced by gamma rays interacting with the scintillation crystal.

pixel A 2D picture element.

projection data The data produced by a medical imaging system.

radius of rotation The distance between the center of the patient and the face of the Anger camera.

root mean square (RMS) error A mathematical measure of closeness, which is equivalent to the distance between two I-dimensional points.

scintillation crystal A rectangular volumetric crystal which, when impinged upon by gamma rays, produces optical photons.

septa The walls of the collimator

single-photon emission computed tomography (SPECT) The sub-field of medical imaging which measures the function of organs within the body with radio-isotopes that produce a single gamma ray particle.

Acknowledgments

The author would like to acknowledge many beneficial discussions on SPECT image reconstruction with Charles Byrne of the University of Massachusetts at Lowell, Michael King, Stephen Glick and colleagues of the University of Massachusetts Medical School, Harrison Barrett and colleagues of the University of Arizona, Craig Abbey of the University of California at Davis, Gene Gindi of SUNY Stony Brook, and Jeffrey Fessler of the University of Michigan. Many excellent suggestions for improving this module were offered by Elizabeth Jessup of the University of Colorado at Boulder. Finally, thanks go to all Consortium member, in particular Ignatios Vakalis of Capital University.

References

[1] R.J. English and S.E. Brown, *SPECT: Single-Photon Emission Computed Tomography: A Primer, 2nd edition*, New York: Society of Nuclear Medicine, 1990.

[2] C. Guy and D. ffytche, *An Introduction to the Principles of Medical Imaging*, London: Imperial College Press, 2000.

[3] H.H. Barrett and W. Swindell, *Radiological Imaging: The Theory of Image Formation, Detection and Processing, Revised edition*, New York: Academic Press, 1981.

[4] J.A. Sorensen and M.E. Phelps, *Physics in Nuclear Medicine, 2nd edition*, Philadelphia: W.B. Saunders, 1987.

[5] R.L. Burden and J.D. Faires, *Numerical Analysis, 8th edition*, Pacific Grove, CA: Brooks Cole, 2004.

[6] X-L Xu, J-S Liow and S.C. Strother, "Iterative algebraic reconstruction algorithms for emission computed tomography: A unified framework and its applications to positron emission tomography", *Med. Phys.*, vol. 20, no. 6, pp. 1675-1684, Nov/Dec 1993.

[7] H. Anton and C. Rorres, *Elementary Linear Algebra: Applications Version, 8th edition*, New York: John Wiley and Sons, 2000.

[8] S. Webb, *From the Watching of Shadows: The Origins of Radiological Imaging*, Bristol, England: IOP Publishing, 1990.

[9] J. Radon, "Über die bestimmung von funktionen durch ihre integralwerte langs gewisser mannigfaltigkeiten," *Ber. Saechs. Akad. Wiss. (Leipzig)*, vol. 69, pp. 262-278.

[10] G.T. Herman and A. Lent, "A computer implemenmtation of a Bayesian analysis of image reconstruction", *Inform. and Control*, vol. 31, pp. 364-384, 1976.

[11] L. Landweber, "An iteration formula for Fredholm integral equations of the first kind", *Am. J. Math.*, vol. 73, pp. 615-624, 1951.

[12] H. Murrell, "Computer-aided tomography," *The Mathematica Journal*, vol. 6, no. 2, pp. 60-65, 1996.

[13] M.F. Smith, C.E. Floyd, R.J. Jaszczak, and R.E. Coleman, "Reconstruction of SPECT images using generalized matrix inverses," *IEEE Trans. Med. Imag.*, vol. 11, no. 2, pp. 165-175, June 1992.

[14] W. Swindell and H.H. Barrett, "Computerized tomography: taking sectional x rays," *Physics Today*, pp. 32-41, Dec. 1977.

Appendix - Solutions

Conceptual Questions

1. The parameter R is the distance from the center-of-rotation to the face of the detector. Practically, the detector can only get as close as the body of the patient will allow. If we think of the patient as being circumscribed by a circle of radius r, then $r < R < \infty$. However, R can't realistically be too large, as the maximum distance is determined by the physical construction of the imaging system.

2. Since $\mathbf{f}$ represents the radio-tracer distribution within the patient, it must be the case that $f_j \geq 0$ for all $j = 1, \ldots, J$, as object pixels either have radio-tracer in them or they don't. The sensitivity factors $h_{i,j}$ represent the contribution of pixel j to detector measurement i. Either a pixel contributes *positively* to a measurement or it doesn't contribute at all. Thus, $h_{i,j} \geq 0$ for all $j = 1, \ldots, J$. Finally, $\mathbf{g}$ is the projection data, whose elements represent the sum of all radioactivity from pixels in $\mathbf{f}$ weighted by $h_{i,j}$, and so these must also satisfy $g_i \geq 0$ for all $i = 1, \ldots, I$ (the sum of non-negative numbers must be non-negative). In summary, $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{H}$ have *non-negativity* constraints.

3. Visually, a coarse discretization (J is small) of the object will make it appear blurry, while a finer one (J is large) will allow for detail to be clearly discerned. However, the larger J is, the larger the dimensionality of the linear system. Quantitatively, the discretization process affects the values of the image pixels. In pixels that partially contain the object, their values will reflect the percentage of the radio-tracer within that pixel. Additionally, if the radio-tracer distribution is non-uniform (such as the MCAT), a pixel that represents multiple structures (values) of the radio-tracer must reflect this in its value, but averaging the structural values. This shows up visually as a blurry image.

4. Determinancy is influenced by the discretization we choose, which is the number of object pixels (unknowns) J. Also, the number of measurements (equations) I that we choose also has influence. Both of these quantities define the dimensions of our linear system, which is in effect its determinancy.

5. Under-determined systems can either have an infinite number of solutions or no solution. It can't have a unique solution, as there are not enough constraints on the unknown variables f_j. Both even- and over-determined systems can have an infinite number of solutions, a unique solution or no solution.

6. Theoretically, we can only use LWB or SIRT to solve all three types of linear systems. The JAC and G-S methods can only be used in the even-determined case when the equations can be ordered so that the pivots are

all non-zero. However, round-off and truncation errors may make some of these methods impractical to use for a given linear system.

7. The data can be fit exactly when the linear system is consistent, i.e. when there is either an infinite number of solutions or an exact solution.

8. Yes, two different objects may produce the same data set. An easy example of this is given by the following two linear systems, which produce the same data set from different objects

$$\begin{pmatrix} 0 & 2 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{1}{2} \end{pmatrix} \quad . \qquad (54)$$
$$\begin{pmatrix} 0 & 2 \\ 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{1}{2} \end{pmatrix}$$

This is due to the fact that the components of both objects that make them different can be represented by the null vectors for the projector matrix.