Iterative Optimization in Medical Imaging

Charles Byrne
(Charles_Byrne@uml.edu)
http://faculty.uml.edu/cbyrne/cbyrne.html
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854, USA

March 12, 2008

1 Overview

Medical imaging is a broad field and we shall focus here primarily on only two areas: x-ray transmission tomography (CAT-scans); and emission tomography (PET and SPECT scans). We use transmission tomography to illustrate the non-iterative filtered-backprojection (FBP) approach to imaging, and emission tomography, mainly single-photon emission computed tomography (SPECT), to illustrate iterative methods. The topics we shall consider in this article are as follows:

- 1. transmission tomography and non-iterative filtered backprojection;
- 2. discretization and iteration- ART and MART;
- 3. emission tomography and iterative likelihood maximization;
- 4. the EMML algorithm for emission tomography;
- 5. acceleration through block-iterative methods;
- 6. iterative entropy maximization;
- 7. the split feasibility problem and the CQ algorithm;
- 8. intensity-modulated radiation therapy (IMRT).

2 Transmission Tomography

In transmission tomography, radiation, usually x-ray, is transmitted along many lines through the object of interest and the initial and final intensities are measured. The intensity drop associated with a given line indicates the amount of attenuation the ray encountered as it passed along the line. It is this distribution of attenuating matter

within the patient, described by a function of two or three spatial variables, that is the object of interest. Unexpected absence of attenuation can indicate a break in a bone, for example. The data are usually modeled as line integrals of that function. The Radon transform is the function that associates with each line its line integral.

2.1 The Exponential-Decay Model

As an x-ray beam passes through the body, it encounters various types of matter, such as soft tissue, bone, ligaments, air, each weakening the beam to a greater or lesser extent. If the intensity of the beam upon entry is I_{in} and I_{out} is its lower intensity after passing through the body, then

$$I_{out} = I_{in} e^{-\int_L f}, \tag{1}$$

where $f = f(x, y) \geq 0$ is the *attenuation function* describing the two-dimensional distribution of matter within the slice of the body being scanned and $\int_L f$ is the integral of the function f over the line L along which the x-ray beam has passed.

To see why this is the case, imagine the line L parameterized by the variable s and consider the intensity function $I(s)$ as a function of s. For small $\Delta s > 0$, the drop in intensity from the start to the end of the interval $[s, s + \Delta s]$ is approximately proportional to the intensity $I(s)$, to the attenuation $f(s)$ and to Δs, the length of the interval; that is,

$$I(s) - I(s + \Delta s) \approx f(s) I(s) \Delta s. \tag{2}$$

Dividing by Δs and letting Δs approach zero, we get

$$I'(s) = -f(s) I(s). \tag{3}$$

The solution to this differential equation is

$$I(s) = I(0) \exp(-\int_{u=0}^{u=s} f(u) du). \tag{4}$$

From knowledge of I_{in} and I_{out}, we can determine $\int_L f$. If we know $\int_L f$ for every line in the x, y-plane we can reconstruct the attenuation function f.

2.2 In Practice

In the real world we know line integrals only approximately and only for finitely many lines. The goal in x-ray transmission tomography is to estimate the attenuation function $f(x, y)$ in the slice, from finitely many noisy measurements of the line integrals. We usually have prior information about the values that $f(x, y)$ can take on. We also expect to find sharp boundaries separating regions where the function $f(x, y)$ varies only slightly. Therefore, we need algorithms capable of providing such images. As we shall see, the line-integral data can be viewed as values of the Fourier transform of the attenuation function.

2.3 Reconstruction from Line Integrals

We turn now to the underlying problem of reconstructing such functions from line-integral data. Our goal is to reconstruct the function $f(x, y)$ from line-integral data. Let θ be a fixed angle in the interval $[0, \pi)$. Form the t, s-axis system with the positive t-axis making the angle θ with the positive x-axis, as shown in Figure 1. Each point (x, y) in the original coordinate system has coordinates (t, s) in the second system, where the t and s are given by

$$t = x \cos\theta + y \sin\theta, \tag{5}$$

and

$$s = -x \sin\theta + y \cos\theta. \tag{6}$$

If we have the new coordinates (t, s) of a point, the old coordinates are (x, y) given by

$$x = t \cos\theta - s \sin\theta, \tag{7}$$

and

$$y = t \sin\theta + s \cos\theta. \tag{8}$$

We can then write the function f as a function of the variables t and s. For each fixed value of t, we compute the integral

$$\int_L f(x, y)ds = \int f(t \cos\theta - s \sin\theta, t \sin\theta + s \cos\theta)ds \tag{9}$$

along the single line L corresponding to the fixed values of θ and t. We repeat this process for every value of t and then change the angle θ and repeat again. In this way we obtain the integrals of f over every line L in the plane. We denote by $r_f(\theta, t)$ the integral

$$r_f(\theta, t) = \int_L f(x, y)ds. \tag{10}$$

The function $r_f(\theta, t)$ is called the *Radon transform* of f.

For fixed θ the function $r_f(\theta, t)$ is a function of the single real variable t; let $R_f(\theta, \omega)$ be its Fourier transform. Then

$$R_f(\theta, \omega) = \int r_f(\theta, t)e^{i\omega t}dt \tag{11}$$

$$= \int\int f(t \cos\theta - s \sin\theta, t \sin\theta + s \cos\theta)e^{i\omega t}dsdt \tag{12}$$

$$= \int\int f(x, y)e^{i\omega(x \cos\theta + y \sin\theta)}dxdy = F(\omega \cos\theta, \omega \sin\theta), \tag{13}$$

where $F(\omega\cos\theta, \omega\sin\theta)$ is the two-dimensional Fourier transform of the function $f(x, y)$, evaluated at the point $(\omega\cos\theta, \omega\sin\theta)$; this relationship is called the *Central Slice Theorem*.

For fixed θ, as we change the value of ω, we obtain the values of the function F along the points of the line making the angle θ with the horizontal axis. As θ varies in $[0, \pi)$, we get all the values of the function F. Once we have F, we can obtain f using the formula for the two-dimensional inverse Fourier transform. We conclude that we are able to determine f from its line integrals.

The Fourier-transform inversion formula for two-dimensional functions tells us that the function $f(x, y)$ can be obtained as

$$f(x, y) = \frac{1}{4\pi^2} \int \int F(u, v) e^{-i(xu+yv)} du dv. \tag{14}$$

The *filtered backprojection* methods commonly used in the clinic [28, 29] are derived from different ways of calculating the double integral in Equation (14).

2.3.1 Ramp Filter then Backproject

Expressing the double integral in polar coordinates (ω, θ), with $\omega \geq 0$, $u = \omega\cos\theta$, and $v = \omega\sin\theta$, we get

$$f(x, y) = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{\infty} F(u, v) e^{-i(xu+yv)} \omega d\omega d\theta,$$

or

$$f(x, y) = \frac{1}{4\pi^2} \int_0^{\pi} \int_{-\infty}^{\infty} F(u, v) e^{-i(xu+yv)} |\omega| d\omega d\theta.$$

Now write

$$F(u, v) = F(\omega\cos\theta, \omega\sin\theta) = R_f(\theta, \omega),$$

where $R_f(\theta, \omega)$ is the FT with respect to t of $r_f(\theta, t)$ so that

$$\int_{-\infty}^{\infty} F(u, v) e^{-i(xu+yv)} |\omega| d\omega = \int_{-\infty}^{\infty} R_f(\theta, \omega) |\omega| e^{-i\omega t} d\omega.$$

The function $h_f(\theta, t)$ defined for $t = x\cos\theta + y\sin\theta$ by

$$h_f(\theta, x\cos\theta + y\sin\theta) = \frac{1}{2\pi} \int_{-\infty}^{\infty} R_f(\theta, \omega) |\omega| e^{-i\omega t} d\omega \tag{15}$$

is the result of a linear filtering of $r_f(\theta, t)$ using a *ramp filter* with transfer function $G(\omega) = |\omega|$. Then,

$$f(x, y) = \frac{1}{2\pi} \int_0^{\pi} h_f(\theta, x\cos\theta + y\sin\theta) d\theta \tag{16}$$

gives $f(x, y)$ as the result of a *backprojection operator*; for every fixed value of (θ, t) add $h_f(\theta, t)$ to the current value at the point (x, y) for all (x, y) lying on the straight

line determined by θ and t by $t = x\cos\theta + y\sin\theta$. The final value at a fixed point (x, y) is then the average of all the values $h_f(\theta, t)$ for those (θ, t) for which (x, y) is on the line $t = x\cos\theta + y\sin\theta$. It is therefore said that $f(x, y)$ can be obtained by *filtered backprojection* (FBP) of the line-integral data.

Knowing that $f(x, y)$ is related to the complete set of line integrals by filtered backprojection suggests that, when only finitely many line integrals are available, a similar ramp filtering and backprojection can be used to estimate $f(x, y)$; in the clinic this is the most widely used method for the reconstruction of tomographic images.

2.3.2 Backproject then Ramp Filter

There is a second way to recover $f(x, y)$ using backprojection and filtering, this time in the reverse order; that is, we backproject the Radon transform and then ramp filter the resulting function of two variables. We begin again with the relation

$$f(x,y) = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{\infty} F(u,v)e^{-i(xu+yv)}\omega d\omega d\theta,$$

which we write as

$$f(x,y) = \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{\infty} \frac{F(u,v)}{\sqrt{u^2+v^2}}\sqrt{u^2+v^2}e^{-i(xu+yv)}\omega d\omega d\theta$$

$$= \frac{1}{4\pi^2}\int_0^{2\pi}\int_0^{\infty} G(u,v)\sqrt{u^2+v^2}e^{-i(xu+yv)}\omega d\omega d\theta, \tag{17}$$

using

$$G(u,v) = \frac{F(u,v)}{\sqrt{u^2+v^2}}$$

for $(u, v) \neq (0, 0)$. Equation (17) expresses $f(x, y)$ as the result of ramp filtering $g(x, y)$, the inverse Fourier transform of $G(u, v)$. We show now that $g(x, y)$ is the backprojection of the function $r_f(\omega, t)$; that is, we show that

$$g(x,y) = \frac{1}{2\pi}\int_0^{\pi} r_f(\theta, x\cos\theta + y\sin\theta)d\theta.$$

From the central slice theorem we know that $g(x, y)$ can be written as

$$g(x,y) = \frac{1}{2\pi}\int_0^{\pi} h_g(\theta, x\cos\theta + y\sin\theta)d\theta,$$

where

$$h_g(\theta, x\cos\theta + y\sin\theta) = \frac{1}{2\pi}\int_{-\infty}^{\infty} R_g(\theta,\omega)|\omega|e^{-i\omega(x\cos\theta+y\sin\theta)}d\omega.$$

Since

$$R_g(\theta,\omega) = G(\omega\cos\theta, \omega\sin\theta),$$

we have

$$g(x,y) = \frac{1}{4\pi^2}\int_0^\pi \int_{-\infty}^{\infty} G(\omega\cos\theta, \omega\sin\theta)|\omega| e^{-i\omega(x\cos\theta + y\sin\theta)} d\omega d\theta$$

$$= \frac{1}{4\pi^2}\int_0^\pi \int_{-\infty}^{\infty} F(\omega\cos\theta, \omega\sin\theta) e^{-i\omega(x\cos\theta + y\sin\theta)} d\omega d\theta$$

$$= \frac{1}{4\pi^2}\int_0^\pi \int_{-\infty}^{\infty} R_f(\theta,\omega) e^{-i\omega(x\cos\theta + y\sin\theta)} d\omega d\theta$$

$$= \frac{1}{2\pi}\int_0^\pi r_f(\theta, x\cos\theta + y\sin\theta) d\theta,$$

as required.

3 The Discrete Model

The estimated attenuation function will ultimately be reduced to a finite array of numbers prior to display. This discretization can be performed at the end, or can be made part of the problem model from the start. In the latter case, the attenuation function is assumed to be constant over small pixels or voxels; these constants are the object of interest now. The problem has been reduced to solving a large system of linear equations, possibly subject to non-negativity or other constraints.

If the physical nature of the radiation is described using a statistical model, then the pixel values can be viewed as parameters to be estimated. The well-known maximum likelihood parameter estimation method can then be employed to obtain these pixel values. This involves a large-scale optimization of the likelihood function.

3.1 Discretization

For $j = 1, ..., J$, let x_j be the unknown constant value of the attenuation function within the jth pixel or voxel (see Figure 2). For $i = 1, ..., I$, let L_i be the set of pixel indices j for which the j-th pixel intersects the i-th line segment, let $|L_i|$ be the cardinality of the set L_i, and $b_i > 0$ the measured approximation of the line integral of f along L_i. Let $A_{ij} = 1$ for j in L_i, and $A_{ij} = 0$ otherwise. The problem is then to solve $Ax = b$, perhaps with the added constraint that the vector x be non-negative.

3.2 ART and MART

The ART and MART algorithms are due to Gordon, Bender, and Herman [19] (see also [20]). With $i = k(\mathrm{mod}\, I) + 1$, the iterative step of the ART algorithm is

$$x_j^{k+1} = x_j^k + \frac{1}{|L_i|}(b_i - (Ax^k)_i), \tag{18}$$

for j in L_i, and

$$x_j^{k+1} = x_j^k, \tag{19}$$

if j is not in L_i. In each step of ART, we take the error, $b_i - (Ax^k)_i$, associated with the current x^k and the i-th equation, and distribute it equally over each of the pixels that intersects L_i. For the ART we do not require that the A_{ij}, b_i or the x_j be non-negative.

3.2.1 Solving $Ax = b$ with the ART

The ART can be used to solve any system of linear equations $Ax = b$. The iterative step of ART is

$$x_j^{k+1} = x_j^k + \frac{1}{||a^i||^2}(b_i - (Ax^k)_i)a^i, \tag{20}$$

where a^i denotes the ith column of $A^\dagger$. When $Ax = b$ has solutions, the sequence $\{x^k\}$ converges to the solution closest to x^0, as shown in Figure 3. When $Ax = b$ has no solutions, the sequence $\{x^k\}$ does not converge; however, subsequences converge to distinct vectors forming a *limit cycle*, as shown in Figure 4.

Suppose, now, that we have $A_{ij} \geq 0$, $b_i > 0$ and we know that the desired image we wish to reconstruct must be nonnegative. We can begin with $x^0 > 0$, but as we compute the ART steps, we may lose nonnegativity. One way to avoid this loss is to correct the current x^k multiplicatively, rather than additively, as in ART. This leads to the *multiplicative* ART (MART).

The MART, in this case, has the iterative step

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big), \tag{21}$$

for those j in L_i, and

$$x_j^{k+1} = x_j^k, \tag{22}$$

otherwise. Therefore, we can write the iterative step as

$$x_j^{k+1} = x_j^k \Big(\frac{b_i}{(Ax^k)_i} \Big)^{A_{ij}}. \tag{23}$$

3.2.2 Solving $y = Px$ using MART

The MART can be used to solve systems of the form $y = Px$, where y has positive entries, P has non-negative entries, and we seek a non-negative solution. The iterative step of MART is

$$x_j^{k+1} = x_j^k \Big(\frac{y_i}{(Px^k)_i} \Big)^{P_{i,j}/m_i}, \tag{24}$$

where $m_i = \max\{P_{i,j} \,|\, j = 1, ..., J\}$. If $y = Px$ has non-negative solutions, then MART converges to such a solution; if not, MART exhibits subsequential convergence to a limit cycle, similar to ART.

4 Using Prior Knowledge

In the tomography problems, as in many other instances of remote sensing, the image to be reconstructed represents a function of two or three continuous variables, and we have finitely many measurements pertaining to this function. Therefore, these problems are always under-determined and there are infinitely many solutions consistent with the measured data. How are we to obtain an accurate reconstruction under these circumstances? A partial answer is that we need to have some idea of what a good reconstruction should look like, and, if possible, make that prior knowledge part of the reconstruction method. In this section we illustrate this point, using weighted least-squares reconstruction.

4.1 An Example

If we take J, the number of pixels, to be larger than I, the number of measurements, then we have an under-determined problem, with multiple solutions. In such cases, prior knowledge can be used effectively to produce a reasonable reconstruction. Figure 5 illustrates this point.

The original image on the upper right of Figure 5 is a discrete rectangular array of intensity values simulating a slice of a head. The data was obtained by taking the two-dimensional discrete Fourier transform of the original image, and then discarding, that is, setting to zero, all these spatial frequency values, except for those in a smaller rectangular region around the origin. The problem then is under-determined. A minimum-norm solution would seem to be a reasonable reconstruction method.

4.2 The Minimum-Norm Reconstruction

The minimum-norm solution is shown on the lower right. It is calculated simply by performing an inverse discrete Fourier transform on the array of modified discrete Fourier transform values. The original image has relatively large values where the skull is located, but the minimum-norm reconstruction does not want such high values; the norm involves the sum of squares of intensities, and high values contribute disproportionately to the norm. Consequently, the minimum-norm reconstruction chooses instead to conform to the measured data by spreading what should be the skull intensities throughout the interior of the skull. The minimum-norm reconstruction does tell us something about the original; it tells us about the existence of the skull itself, which, of course, is indeed a prominent feature of the original. However, in all likelihood, we would already know about the skull; it would be the interior that we want to know about.

4.3 Using a Prior Estimate as a Weight

Using our knowledge of the presence of a skull, which we might have obtained from the minimum-norm reconstruction itself, we construct the prior estimate shown in the upper left. Now we use the same data as before, and calculate a minimum-weighted-norm reconstruction, using as the weight vector the reciprocals of the values of the prior image. This minimum-weighted-norm reconstruction is shown on the lower left;

it is clearly almost the same as the original image. The calculation of the minimum-weighted norm solution can be done iteratively using the ART algorithm [33].

4.4 Why it Works

When we weight the skull area with the inverse of the prior image, we allow the reconstruction to place higher values there without having much of an affect on the overall weighted norm. In addition, the reciprocal weighting in the interior makes spreading intensity into that region costly, so the interior remains relatively clear, allowing us to see what is really present there.

When we try to reconstruct an image from limited data, it is easy to assume that the information we seek has been lost, particularly when a reasonable reconstruction method fails to reveal what we want to know. As this example, and many others, show, the information we seek is often still in the data, but needs to be brought out in a more subtle way.

4.5 The Burrito Test

Years ago, I went with a colleague to lunch at a Mexican restaurant. He had never been to one before and was not a fan of "exotic" cuisine, in general. I suggested he order a combination plate, so he might find something he liked. When he bit into his first burrito, he looked at me and asked "How do they know when they have made a bad one?" This is a question we should remember to ask when we reconstruct an image from limited data; this is the *burrito test*. It is usually possible to fit limited data to almost any finite-parameter model, if there are enough parameters. Mere consistency with the measured data will not be enough to give us a good reconstruction, in general. We should have some way to find out if we have made a bad one.

In the case of the minimum-weighted-norm reconstruction above, an obvious question is how our choice of prior affects the final answer. Although there is no foolproof way to know when the prior is good and when it is bad, there is something we can check. If the prior is inappropriate, the reconstruction will be consistent with the data and have the minimum weighted norm among all images consistent with the data, but that minimum weighted norm can still be very large. The size of this minimum value is an indicator of how suitable the prior estimate is. For example, if we make the skull a bit too large, the minimum weighted norm itself will probably not be large, but if we make the skull too small, this norm will probably be much larger. This is an ill-conditioning effect (see [33]).

4.6 The Phase Problem

It is often the case that the prior weighting function is appropriate when the data is noise-free, but not otherwise. For safety, one should never take a prior that is precisely zero outside some small region, but one that takes on a small but positive value there. In [4], this sensitivity to errors in the data was used to solve the *phase problem*. In that problem one has only finitely many values of the magnitude of the complex-valued

function; the phase is lost. When grossly incorrect phases are assumed, the minimum-norm solution consistent with the support and this data will have an unrealistically large norm. Only when we have succeeded in approximating the true phases will the norm of the reconstruction be reasonable. By monitoring the norm of the reconstructions, we are led to an appropriate choice of phases and to an acceptable reconstruction.

5 Emission Tomography

In emission tomography [36], a carefully designed chemical tagged with a radioisotope is introduced into the body of the patient. The chemical is selected to accumulate in a specific organ or region of the body, such as the brain, or the heart wall. On the basis of emissions from the radioisotope that are detected outside the body, the distribution of the chemical within the body is estimated. Unexpected absence of the radionuclide from a given region, or a higher than expected concentration, can indicate a medical problem.

5.1 SPECT and PET

There are two basic types of emission tomography: single photon emission computed tomography (SPECT); and positron emission tomography (PET). In SPECT the radioisotope emits a single photon, while in PET a positron is emitted, which shortly meets an electron and the resulting annihilation produces two gamma-ray photons traveling in essentially opposite directions.

In both SPECT and PET the data can be approximated as integrals along lines through the body and image reconstruction performed using FBP. However, more sophisticated models that more accurately describe the physics of the situation are preferred. The photons that travel through the body toward the external detectors are sometimes absorbed by the body itself and not detected. The probability of being detected depends on the attenuation presented by the body. This attenuation, while not the object of interest now, is an important part of the physical model and needs to be included in the reconstruction method. The randomness inherent in emission can also be included, leading once again to probabilistic models and a maximum likelihood approach to reconstruction.

Although in both transmission and emission medical tomography the dosage to the patient is restricted, thereby decreasing the signal-to-noise ratio, the amount of data is still considerable and the need to produce the reconstructed image in a few minutes paramount. Much work has gone into methods for accelerating the iterative reconstruction algorithms.

5.2 The Poisson Model in SPECT

The discrete model for emission tomography is the following:

- for $j = 1, ..., J$, $x_j \geq 0$ is the unknown expected number of photons emitted from the jth pixel during the scan;

- for $i = 1, ..., I$, $y_i > 0$ is the number of photons detected at the ith detector;
- $P_{ij} \geq 0$ is the probability that a photon emitted at j will be detected at i, which we will assume is known;
- $s_j = \sum_{i=1}^{I} P_{ij}$ is the sensitivity to j, that is, the probability that a photon emitted at j will be detected;
- the y_i are realizations of independent Poisson random variables with expected values $(Px)_i = \sum_{j=1}^{J} P_{ij} x_j$.

5.3 Likelihood Maximization in SPECT

We view the unknown values $x_j \geq 0$ as parameters to be estimated. To within a constant, the log of the likelihood function is then

$$LL(x) = \sum_{i=1}^{I} y_i \log(Px)_i - (Px)_i. \tag{25}$$

The EMML algorithm ([32, 35, 27]) for maximizing $LL(x)$ over $x \geq 0$ has the iterative step

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \Big(\frac{y_i}{(Px^k)_i} \Big). \tag{26}$$

5.4 Problems with the EMML Algorithm

Although the EMML algorithm allows for more accurate description of the physical situation, there are several disadvantages that must be removed before the EM algorithm can be a useful clinical tool:

- Calculating $(Px^k)_i = \sum_{j=1}^{J} P_{ij} x_j^k$, for each i, at each step of the iteration is expensive, since I and J can be in the tens of thousands;
- The sequence $\{x^k\}$ usually converges quite slowly to the maximizer of $LL(x)$;
- The maximum-likelihood (ML) solution will be a non-negative solution of $y = Px$, in the consistent case, that is, if such solutions exist, so may overfit noisy data;
- The ML solution may not be a good choice, in the inconsistent case, either.

5.5 Controlling Noise

It can be shown that, when the system $y = Px$ has no non-negative solutions, the maximum-likelihood solution will have at most $I - 1$ non-zero pixel values, so, if J is greater than I, the ML solution may be useless. To control noise and obtain a useful image, one usually uses *regularization*, which means maximizing the sum of the likelihood function and another penalty function that is larger when the image is smooth [18, 26, 5].

5.6 Acceleration

In the early 1990's it was noticed that if, when performing one step of the EM iteration, one summed only over some of the detector indices, instead of over all of them, one could usually obtain a useful reconstruction more quickly [21, 22, 23].

5.6.1 The OSEM

Suppose that we take a partition $B_1 \cup B_2 \cup \cdots \cup B_N$ of the set $\{i = 1, ..., I\}$, let $s_{n,j} = \sum_{i \in B_n} P_{ij}$, and at the kth step of the iteration, we use only B_n, for $n = k(\mathrm{mod}\, N)+1$, and compute

$$x_j^{k+1} = x_j^k s_{n,j}^{-1} \sum_{i \in B_n} P_{ij} \Big(\frac{y_i}{(Px^k)_i} \Big). \tag{27}$$

This is the *ordered-subset* EM (OSEM) algorithm [22, 23]. Although its mathematical foundations are a bit shaky, it has proven to be a useful clinical tool.

5.6.2 Limit Cycles

Without *strong under-relaxation* [17, 3], algorithms such as ART, MART and the OSEM that use only some of the data at each step of the iteration cannot converge to a single vector in the *inconsistent case*. For MART and OSEM this means there is no non-negative solution of $y = Px$, while for ART it simply means $Ax = b$ has no solution. Instead, these algorithms exhibit subsequential convergence to a limit cycle of (usually) N distinct vectors. Proving that a limit cycle exists seems to be a difficult problem; existence has been shown for ART [34], but not for these other methods. One problem with OSEM is that it sometimes produces a limit cycle, even when there is a non-negative solution of $y = Px$. This makes the OSEM images noisier than they need to be, when the data is noisy.

5.6.3 Rescaled Block-Iterative Reconstruction

The *rescaled block-iterative* EMML (RBI-EMML) [6, 9] is similar to the OSEM, but converges to a non-negative solution of $y = Px$, whenever such solutions exist, for every starting vector $x^0 > 0$ and every choice of blocks. Let x^0 be an arbitrary positive vector. For $k = 0, 1, ...$, let $n = k(\mathrm{mod}\, N) + 1$. Then let

$$x_j^{k+1} = (1 - m_n^{-1} s_j^{-1} s_{n,j}) x_j^k + m_n^{-1} s_j^{-1} x_j^k \sum_{i \in B_n} (P_{ij} \frac{y_i}{(Px^k)_i}), \tag{28}$$

with

$$m_n = \max\{s_{n,j} s_j^{-1} \,|j = 1, ..., J\}.$$

The RBI-EMML is also related to the RAMLA method of [3].

6 Entropy Maximization

When there are multiple non-negative solutions of $y = Px$, it makes sense to select the solution closest to a prior estimate of x, according to some measure of distance. The *cross-entropy* or Kullback-Leibler distance [24] is frequently used. As we shall see, this distance is also closely related to the EMML algorithm.

6.1 The Kullback-Leibler Distance

The Kullback-Leibler distance between positive numbers α and β is

$$KL(\alpha, \beta) = \alpha \log \frac{\alpha}{\beta} + \beta - \alpha.$$

We also define $KL(\alpha, 0) = +\infty$ and $KL(0, \beta) = \beta$. Extending to non-negative vectors $a = (a_1, ..., a_J)^T$ and $b = (b_1, ..., b_J)^T$, we have

$$KL(a, b) = \sum_{j=1}^{J} KL(a_j, b_j) = \sum_{j=1}^{J} \Big(a_j \log \frac{a_j}{b_j} + b_j - a_j \Big).$$

With $a_+ = \sum_{j=1}^{J} a_j$, and $b_+ > 0$, we have

$$KL(a, b) = KL(a_+, b_+) + KL(a, \frac{a_+}{b_+} b). \tag{29}$$

6.2 The EMML and Simultaneous MART

The EMML algorithm has the iterative step

$$x_j^{k+1} = x_j^k s_j^{-1} \sum_{i=1}^{I} P_{ij} \Big(\frac{y_i}{(Px^k)_i} \Big). \tag{30}$$

It is interesting to compare this iteration with that of the *simultaneous* MART (SMART)[16, 31, 14]:

$$x_j^{k+1} = x_j^k \exp \Big[s_j^{-1} \sum_{i=1}^{I} P_{ij} \log \Big(\frac{y_i}{(Px^k)_i} \Big) \Big]. \tag{31}$$

We have the following result [32, 27, 35, 5]:

Theorem 1. *The SMART sequence $\{x^k\}$ converges to the non-negative minimizer of $KL(Px, y)$ for which $KL(x, x^0)$ is minimized, for any choice of $x^0 > 0$. The EMML sequence $\{x^k\}$ converges to a non-negative minimizer of $KL(y, Px)$, for any choice of $x^0 > 0$.*

It is an open question to which minimizer the EMML sequence converges. In the consistent case, the limit is a non-negative solution of $y = Px$. If there are multiple non-negative solutions of $y = Px$, the limit will depend on $x^0 > 0$, but we do not know how it depends on x^0.

6.3 SMART and Shannon Entropy

When $y = Px$ has non-negative solutions, the SMART and MART algorithms produce sequences that converge to the unique non-negative solution that minimizes $KL(x, x^0)$, for any $x^0 > 0$. If x^0 is the vector whose entries are all one, then minimizing $KL(x, x^0)$ is equivalent to maximizing the *Shannon entropy*

$$SE(x) = \sum_{j=1}^{J} x_j(\log x_j) - x_j.$$

So the SMART and MART can be used to maximize entropy.

7 The Split Feasibility Problem and the CQ Algorithm

Finally, we consider the split feasibility problem (SFP), the iterative CQ algorithm for solving the SFP [7, 8], and the application of the CQ algorithm to radiation therapy [12, 13].

7.1 The Split Feasibility Problem

Let A be a real I by J matrix, and C and Q non-empty closed, convex sets in R^J and R^I, respectively. The *split feasibility problem* (SFP) is to find a vector x in C, such that Ax is in Q. When the SFP has no solution, it is sensible to seek a minimizer of the function

$$f(x) = \frac{1}{2}||P_Q Ax - Ax||_2^2, \tag{32}$$

over x in C; P_Q denotes the orthogonal projection onto Q.

7.2 The CQ Algorithm

For arbitrary x^0 and $k = 0, 1, ...$, and γ in the interval $(0, 2/\rho(A^T A))$, where $\rho(A^T A)$ denotes the largest eigenvalue of the matrix $A^T A$, let

$$x^{k+1} = P_C(x^k - \gamma A^T(I - P_Q)Ax^k). \tag{33}$$

This is the CQ algorithm [7, 8]. The CQ algorithm converges to a solution of the SFP, whenever solutions exist. When there are no solutions of the SFP, the CQ algorithm converges to a minimizer, over x in C, of the function

$$f(x) = \frac{1}{2}||P_Q Ax - Ax||_2^2, \tag{34}$$

whenever such minimizers exist.

7.3 Estimating $\rho(A^TA)$

The CQ algorithm employs the relaxation parameter γ in the interval $(0, 2/\rho(A^TA))$, where $\rho(A^TA)$ is the largest eigenvalue of the matrix A^TA. Choosing the best relaxation parameter in any algorithm is a nontrivial procedure. Generally speaking, we want to select γ near to $1/\rho(A^TA)$. A simple estimate for $\rho(A^TA)$ that is particularly useful when A is sparse is the following: if A is normalized so that each row has length one, then the spectral radius of A^TA does not exceed the maximum number of nonzero elements in any column of A. A similar upper bound on $\rho(A^TA)$ was obtained for non-normalized, ϵ-sparse A [7].

7.4 Intensity-Modulated Radiation Therapy

Recently, Censor, Elfving, Kopf and Bortfeld [12] have extended the CQ algorithm to the case in which the sets C and Q are the intersections of finitely many other convex sets. The new algorithm employs the orthogonal projections onto these other convex sets.

In [13] Censor, Bortfeld, Martin, and Trofimov use this new algorithm to determine intensity-modulation protocols for radiation therapy. The issue here is to determine the intensities of the radiation sources external to the patient, subject to constraints on how spatially varying the machinery permits these intensities to be, on the maximum dosage directed to healthy areas, and on the minimum dosage directly to the targets.

8 Particular Cases of the CQ Algorithm

It is easy to find important examples of the SFP: if $C = R^J$ and $Q = \{b\}$ then solving the SFP amounts to solving the linear system of equations $Ax = b$; if C is a proper subset of R^J, such as the nonnegative cone, then we seek solutions of $Ax = b$ that lie within C, if there are any. Generally, we cannot solve the SFP in closed form and iterative methods are needed.

8.1 Landweber and Projected Landweber

A number of well known iterative algorithms, such as the Landweber [25] and projected Landweber methods (see [2]), are particular cases of the CQ algorithm.

8.1.1 The Landweber Algorithm

With x^0 arbitrary and $k = 0, 1, ...$, the Landweber algorithm for finding a (possibly least-squares) solution of $Ax = b$ has the iterative step

$$x^{k+1} = x^k + \gamma A^T(b - Ax^k). \tag{35}$$

8.1.2 The Projected Landweber Algorithm

For a general nonempty closed convex C, x^0 arbitrary, and $k = 0, 1, ...$, the projected Landweber algorithm for finding a (possibly constrained least-squares) solution of $Ax = b$ in C has the iterative step

$$x^{k+1} = P_C(x^k + \gamma A^T(b - Ax^k)). \tag{36}$$

8.1.3 The Simultaneous ART (SART)

Another example of the CQ algorithm is the *simultaneous algebraic reconstruction technique* (SART) of Anderson and Kak [1] for solving $Ax = b$, for nonnegative matrix A . Let A be an M by N matrix with nonnegative entries. Let $A_{m+} > 0$ be the sum of the entries in the mth row of A and $A_{+n} > 0$ be the sum of the entries in the nth column of A. Consider the (possibly inconsistent) system $Ax = b$. For x^0 arbitrary and $k = 0, 1, ...$, let

$$x_n^{k+1} = x_n^k + \frac{1}{A_{+n}} \sum\nolimits_{m=1}^{M} A_{mn}(b_m - (Ax^k)_m)/A_{m+}. \tag{37}$$

This is the SART algorithm. With a change of variables, the SART becomes a particular case of the Landweber iteration.

We make the following changes of variables:

$$B_{mn} = A_{mn}/(A_{m+})^{1/2}(A_{+n})^{1/2}, \tag{38}$$

$$z_n = x_n(A_{+n})^{1/2}, \tag{39}$$

and

$$c_m = b_m/(A_{m+})^{1/2}. \tag{40}$$

Then the SART iterative step can be written as

$$z^{k+1} = z^k + B^T(c - Bz^k). \tag{41}$$

This is a particular case of the Landweber algorithm, with $\gamma = 1$. The convergence of SART follows, once we know that the largest eigenvalue of $B^T B$ is less than two; in fact, it is one [7].

9 Proximal Minimization

The CQ algorithm is a particular case of an iterative algorithm based on Moreau's notion of proximity operator.

9.1 Proximity Operators

The Moreau envelope of a convex function f is the function

$$m_f(z) = \inf_x \{f(x) + \frac{1}{2}||x - z||_2^2\}, \tag{42}$$

which is also the infimal convolution of the functions $f(x)$ and $\frac{1}{2}||x||_2^2$. It can be shown that the infimum is uniquely attained at the point denoted $x = \text{prox}_f z$ (see Rockafellar [30]). The function $m_f(z)$ is differentiable and $\nabla m_f(z) = z - \text{prox}_f z$. The point $x = \text{prox}_f z$ is characterized by the property $z - x \in \partial f(x)$. Consequently, x is a global minimizer of f if and only if $x = \text{prox}_f x$.

The *conjugate function* associated with f is the function $f^*(x^*) = \sup_x(\langle x^*, x\rangle - f(x))$. In similar fashion, we can define $m_{f^*} z$ and $\text{prox}_{f^*} z$. Both m_f and m_{f^*} are convex and differentiable.

9.2 Moreau's Theorem

Theorem 2. *Let f be a closed, proper, convex function with conjugate f^*. Then*

$$\begin{aligned} m_f z + m_{f^*} z &= \frac{1}{2}||z||^2; \\ \text{prox}_f z + \text{prox}_{f^*} z &= z; \\ \text{prox}_{f^*} z &\in \partial f(\text{prox}_f z); \\ \text{prox}_{f^*} z &= \nabla m_f(z), \text{ and} \\ \text{prox}_f z &= \nabla m_{f^*}(z). \end{aligned} \tag{43}$$

9.2.1 An Example

For example, consider the indicator function of the convex set C, $f(x) = \iota_C(x)$ that is zero if x is in the closed convex set C and $+\infty$ otherwise. Then $m_f z$ is the minimum of $\frac{1}{2}||x - z||_2^2$ over all x in C, and $\text{prox}_f z = P_C z$, the orthogonal projection of z onto the set C. The operators $\text{prox}_f : z \to \text{prox}_f z$ are proximity operators. These operators generalize the projections onto convex sets, and, like those operators, are firmly non-expansive (see Combettes and Wajs [15]).

The support function of the convex set C is $\sigma_C(x) = \sup_{u \in C}\langle x, u\rangle$. It is easy to see that $\sigma_C = \iota_C^*$. For $f^*(z) = \sigma_C(z)$, we can find $m_{f^*} z$ using Moreau's Theorem:

$$\text{prox}_{\sigma_C} z = z - \text{prox}_{\iota_C} z = z - P_C z. \tag{44}$$

9.3 Using Moreau's Theorem

The minimizers of m_f and the minimizers of f are the same. From Moreau's Theorem we know that

$$\nabla m_f(z) = \text{prox}_{f^*} z = z - \text{prox}_f z, \tag{45}$$

so $\nabla m_f z = 0$ is equivalent to $z = \text{prox}_f z$.

Because the minimizers of m_f are also minimizers of f, we can find global minimizers of f using gradient descent iterative methods on m_f.

Let x^0 be arbitrary. Then let

$$x^{k+1} = x^k - \gamma_k \nabla m_f(x^k). \tag{46}$$

We know from Moreau's Theorem that

$$\nabla m_f z = \text{prox}_{f^*} z = z - \text{prox}_f z, \tag{47}$$

so that Equation (46) can be written as

$$\begin{aligned} x^{k+1} &= x^k - \gamma_k (x^k - \text{prox}_f x^k) \\ &= (1 - \gamma_k) x^k + \gamma_k \text{prox}_f x^k. \end{aligned} \tag{48}$$

It follows from the definition of $\partial f(x^{k+1})$ that $f(x^k) \geq f(x^{k+1})$ for the iteration in Equation (48).

10 The CQ Algorithm as Forward-Backward Splitting

In [15] Combettes and Wajs consider the problem of minimizing the function $F(x) = f_1(x) + f_2(x)$, where $f_2(x)$ is differentiable and its gradient is λ-Lipschitz continuous. The function F is minimized at the point x if and only if

$$0 \in \partial F(x) = \partial f_1(x) + \nabla f_2(x), \tag{49}$$

so we have

$$-\gamma \nabla f_2(x) \in \gamma \partial f_1(x), \tag{50}$$

for any $\gamma > 0$. Therefore

$$x - \gamma \nabla f_2(x) - x \in \gamma \partial f_1(x). \tag{51}$$

From Equation (51) we conclude that

$$x = \text{prox}_{\gamma f_1}(x - \gamma \nabla f_2(x)). \tag{52}$$

This suggests an algorithm, called the *forward-backward splitting* for minimizing the function $F(x)$.

10.1 Forward-Backward Splitting

Beginning with an arbitrary x^0, and having calculated x^k, we let

$$x^{k+1} = \text{prox}_{\gamma f_1}(x^k - \gamma \nabla f_2(x^k)), \tag{53}$$

with γ chosen to lie in the interval $(0, 2/\lambda)$. The operator $I - \gamma \nabla f_2$ is then averaged. Since the operator $\text{prox}_{\gamma f_1}$ is firmly non-expansive, the sequence $\{x^k\}$ converges to a minimizer of the function $F(x)$, whenever minimizers exist. It is also possible to allow γ to vary with the k.

10.2 The CQ Algorithm as Forward-Backward Splitting

Recall that the split-feasibility problem (SFP) is to find x in C with Ax in Q. The CQ algorithm minimizes the function

$$f(x) = ||P_Q Ax - Ax||_2^2, \tag{54}$$

over $x \in C$, whenever such minimizers exist, and so solves the SFP whenever it has solutions. The CQ algorithm therefore minimizes the function

$$F(x) = \iota_C(x) + f(x), \tag{55}$$

where ι_C is the indicator function of the set C. With $f_1(x) = \iota_C(x)$ and $f_2(x) = f(x)$, the function $F(x)$ has the form considered by Combettes and Wajs, and the CQ algorithm becomes a special case of their forward-backward splitting method.

11 Acknowledgments and Final Comments

This work was supported, in part, by the National Institutes of Health (NIH), under grant CA23452. The contents of this article are solely the responsibility of the author and do not necessarily reflect the official views of NIH.

All of the topics discussed here are treated in more detail in the books [10, 11]. My articles are available as pdf files on my web site.

References

[1] Anderson, A. and Kak, A. (1984) "Simultaneous algebraic reconstruction technique (SART): a superior implementation of the ART algorithm" , *Ultrasonic Imaging*, **6** 81–94.

[2] Bertero, M., and Boccacci, P. (1998) *Introduction to Inverse Problems in Imaging* Bristol, UK: Institute of Physics Publishing.

[3] Browne, J. and A. DePierro, A. (1996) "A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography."*IEEE Trans. Med. Imag.* **15**, pp. 687–699.

[4] Byrne, C. and Fiddy, M. (1987) "Estimation of continuous object distributions from Fourier magnitude measurements."*JOSA A* **4**, pp. 412–417.

[5] Byrne, C. (1993) "Iterative image reconstruction algorithms based on cross-entropy minimization."*IEEE Transactions on Image Processing* **IP-2**, pp. 96–103.

[6] Byrne, C. (1996) "Block-iterative methods for image reconstruction from projections."*IEEE Transactions on Image Processing* **IP-5**, pp. 792–794.

[7] Byrne, C. (2002) "Iterative oblique projection onto convex sets and the split feasibility problem."*Inverse Problems* **18**, pp. 441–453.

[8] Byrne, C. (2004) "A unified treatment of some iterative algorithms in signal processing and image reconstruction."*Inverse Problems* **20**, pp. 103–120.

[9] Byrne, C. (2005) "Choosing parameters in block-iterative or ordered-subset reconstruction algorithms" *IEEE Transactions on Image Processing*, **14 (3)**, pp. 321–327.

[10] Byrne, C. (2005) *Signal Processing: A Mathematical Approach*, AK Peters, Publ., Wellesley, MA.

[11] Byrne, C. (2007) *Applied Iterative Methods*, AK Peters, Publ., Wellesley, MA.

[12] Y. Censor, T. Elfving, N. Kopf, and T. Bortfeld. "The Multiple-sets Split Feasibility Problem and its Application for Inverse Problems." *Inverse Problems* 21 (2005), 2071-2084.

[13] Y. Censor, T. Bortfeld, B. Martin, and A. Trofimov, A. "A Unified Approach for Inversion Problems in Intensity-modulated Radiation Therapy." *Physics in Medicine and Biology* 51 (2006), 2353-2365.

[14] Censor, Y. and Segman, J. (1987) "On block-iterative maximization."*J. of Information and Optimization Sciences* **8**, pp. 275–291.

[15] Combettes, P., and Wajs, V. (2005) Signal recovery by proximal forward-backward splitting, *Multiscale Modeling and Simulation*, **4(4)**, pp. 1168–1200.

[16] Darroch, J. and Ratcliff, D. (1972) "Generalized iterative scaling for log-linear models."*Annals of Mathematical Statistics* **43**, pp. 1470–1480.

[17] Eggermont, P.P.B., Herman, G.T., and Lent, A. (1981) "Iterative algorithms for large partitioned linear systems, with applications to image reconstruction."*Linear Algebra and its Applications* **40**, pp. 37–67.

[18] Geman, S., and Geman, D. (1984) "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images."*IEEE Transactions on Pattern Analysis and Machine Intelligence* **PAMI-6**, pp. 721–741.

[19] Gordon, R., Bender, R., and Herman, G.T. (1970) "Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and x-ray photography."*J. Theoret. Biol.* **29**, pp. 471–481.

[20] Herman, G. T. and Meyer, L. (1993) "Algebraic reconstruction techniques can be made computationally efficient."*IEEE Transactions on Medical Imaging* **12**, pp. 600–609.

[21] Holte, S., Schmidlin, P., Linden, A., Rosenqvist, G. and Eriksson, L. (1990) "Iterative image reconstruction for positron emission tomography: a study of convergence and quantitation problems."*IEEE Transactions on Nuclear Science* **37**, pp. 629–635.

[22] Hudson, M., Hutton, B., and Larkin, R. (1992) "Accelerated EM reconstruction using ordered subsets." *Journal of Nuclear Medicine*, **33**, p.960.

[23] Hudson, H.M. and Larkin, R.S. (1994) "Accelerated image reconstruction using ordered subsets of projection data."*IEEE Transactions on Medical Imaging* **13**, pp. 601–609.

[24] Kullback, S. and Leibler, R. (1951) "On information and sufficiency."*Annals of Mathematical Statistics* **22**, pp. 79–86.

[25] Landweber, L. (1951) "An iterative formula for Fredholm integral equations of the first kind."*Amer. J. of Math.* **73**, pp. 615–624.

[26] Lange, K., Bahn, M. and Little, R. (1987) "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography."*IEEE Trans. Med. Imag.* **MI-6(2)**, pp. 106–114.

[27] Lange, K. and Carson, R. (1984) "EM reconstruction algorithms for emission and transmission tomography."*Journal of Computer Assisted Tomography* **8**, pp. 306–316.

[28] Natterer, F. (1986) *Mathematics of Computed Tomography*. New York: John Wiley and Sons, Inc.

[29] Natterer, F., and Wübbeling, F. (2001) *Mathematical Methods in Image Reconstruction*. Philadelphia, PA: SIAM Publ.

[30] Rockafellar, R. (1970) *Convex Analysis*. Princeton, NJ: Princeton University Press.

[31] Schmidlin, P. (1972) "Iterative separation of sections in tomographic scintigrams."*Nucl. Med.* **15(1)**.

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

[33] Shieh, M., Byrne, C., Testorf, M., and Fiddy, M. (2006) "Iterative image reconstruction using prior knowledge." *Journal of the Optical Society of America, A*, **23(6)**, pp. 1292–1300.

[34] Tanabe, K. (1971) "Projection method for solving a singular system of linear equations and its applications."*Numer. Math.* **17**, pp. 203–214.

[35] Vardi, Y., Shepp, L.A. and Kaufman, L. (1985) "A statistical model for positron emission tomography."*Journal of the American Statistical Association* **80**, pp. 8–20.

[36] Wernick, M. and Aarsvold, J., editors (2004) *Emission Tomography: The Fundamentals of PET and SPECT*. San Diego: Elsevier Academic Press.

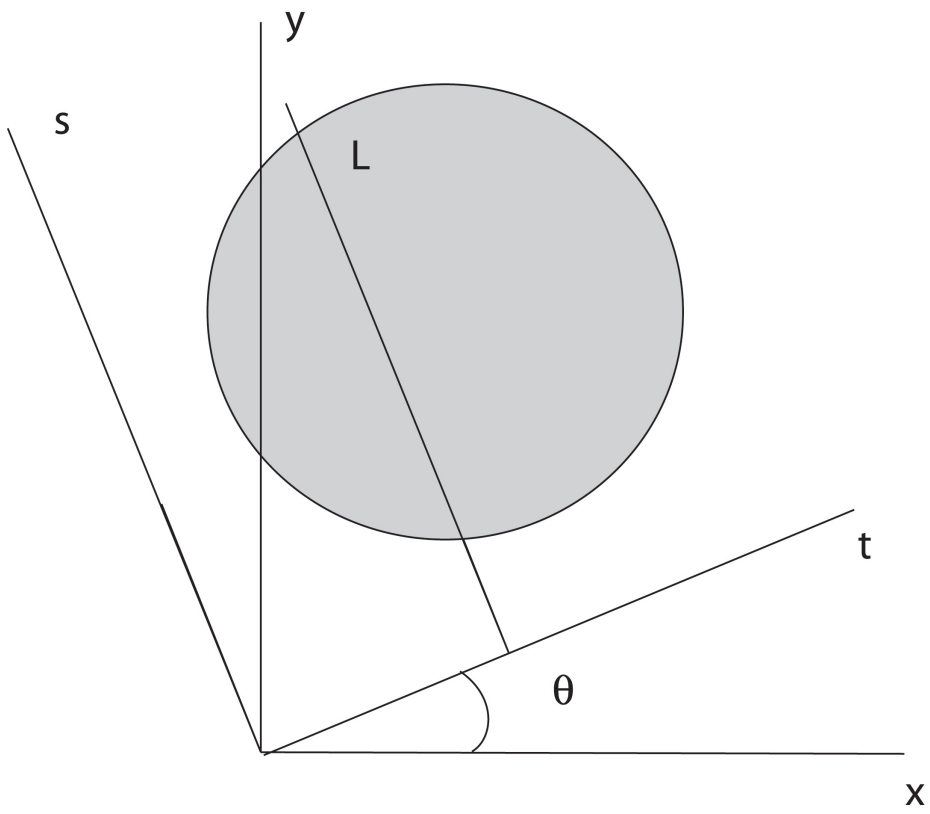

Figure 1: The Radon transform of f at (t, θ) is the line integral of f along line L.

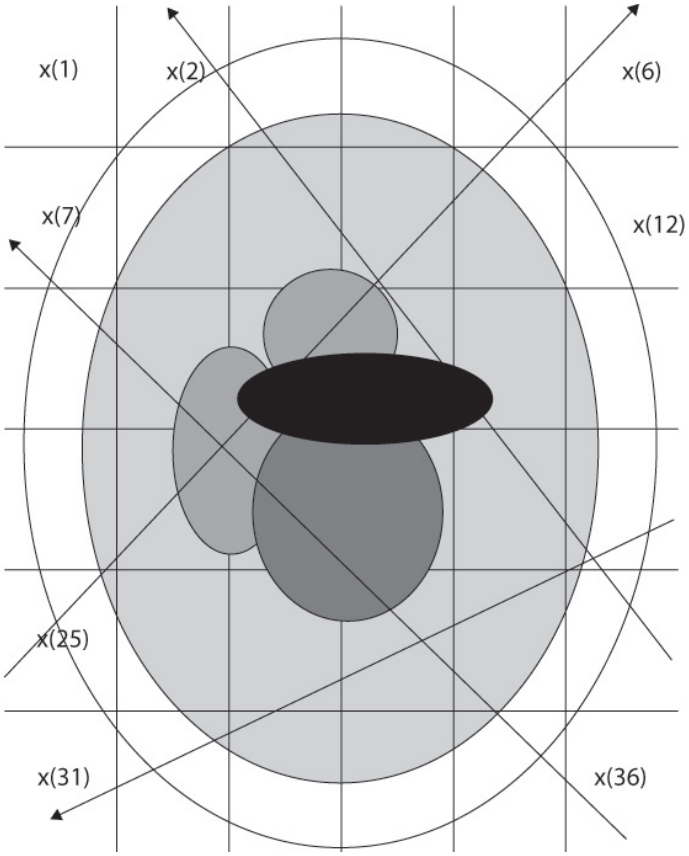

Figure 2: The Discrete Model.

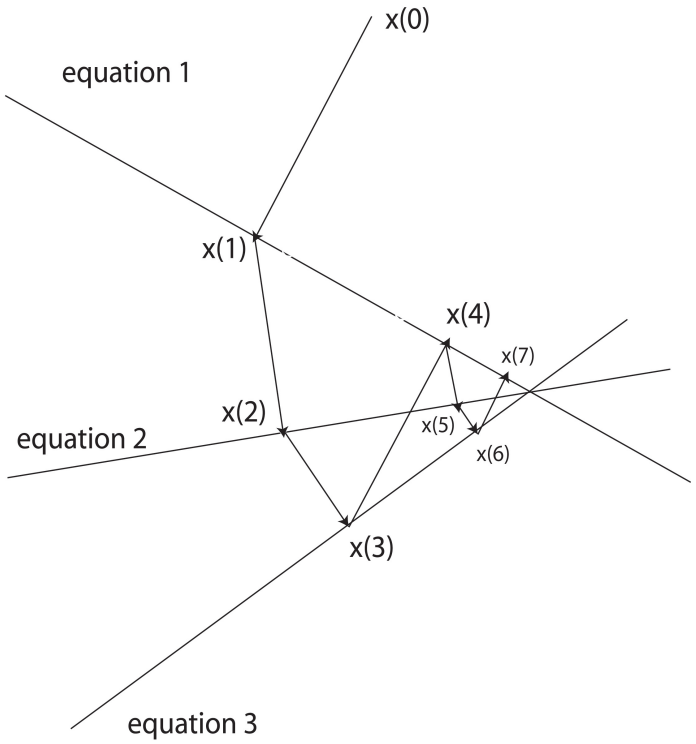

Figure 3: The ART in the consistent case.

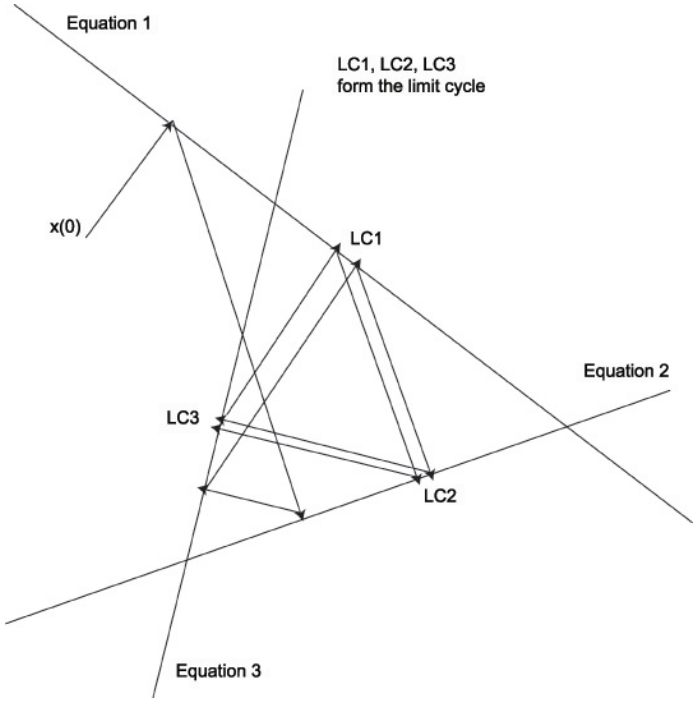

Figure 4: The ART in the inconsistent case.

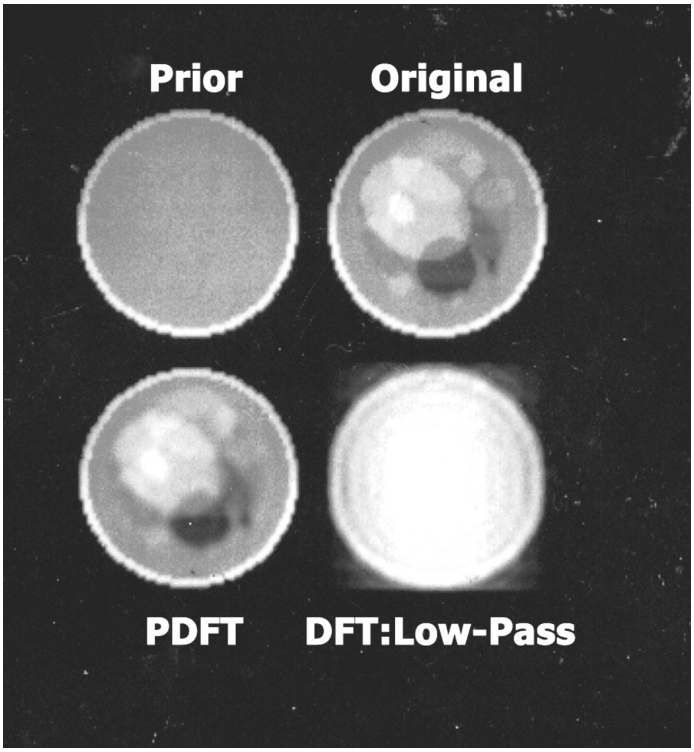

Figure 5: Minimum-norm reconstruction, with and without prior knowledge.