CHAPTER 1
Statistical Image Reconstruction Methods for Transmission Tomography

Jeffrey A. Fessler
University of Michigan

Contents

1.1 Introduction

The problem of forming cross-sectional or *tomographic* images of the attenuation characteristics of objects arises in a variety of contexts, including medical x-ray computed tomography (CT) and nondestructive evaluation of objects in industrial inspection. In the context of emission imaging, such as positron emission tomography (PET) [1, 2], single photon emission computed tomography (SPECT) [3], and related methods used in the assay of containers of radioactive waste [4], it is useful to be able to form "attenuation maps," tomographic images of attenuation coefficients, from which one can compute attenuation correction factors for use in emission image reconstruction. One can measure the attenuating characteristics of an object by transmitting a collection of photons through the object along various paths or "rays" and observing the fraction that pass unabsorbed. From measurements collected over a large set of rays, one can reconstruct tomographic images of the object. Such image reconstruction is the subject of this chapter.

In all the above applications, the number of photons one can measure in a transmission scan is limited. In medical x-ray CT, source strength, patient motion, and absorbed dose considerations limit the total x-ray exposure. Implanted objects such as pacemakers also significantly reduce transmissivity and cause severe artifacts [5]. In industrial applications, source strength limitations, combined with the very large attenuation coefficients of metallic objects, often result in a small fraction of photons passing to the detector unabsorbed. In PET and SPECT imaging, the transmission scan only determines a "nuisance" parameter of secondary interest relative to the object's emission properties, so one would like to minimize the transmission scan duration. All the above considerations lead to "low-count" transmission scans. This chapter discusses algorithms for reconstructing attenuation images from low-count transmission scans. In this context, we define low-count to mean that the mean number of photons per ray is small enough that traditional filtered-backprojection (FBP) images, or even methods based on the Gaussian approximation to the distribution of the Poisson measurements (or logarithm thereof), are inadequate. We focus the presentation in the context of PET and SPECT transmission scans, but the methods are generally applicable to all low-count transmission studies. See [6] for an excellent survey of statistical approaches for the emission reconstruction problem.

Statistical methods for reconstructing attenuation images from transmission scans have increased in importance recently for several reasons. Factors include the necessity of reconstructing 2D attenuation maps for reprojection to form 3D attenuation correction factors in septaless PET [7, 8], the widening availability of SPECT systems equipped with transmission sources [9], and the potential for reducing transmission noise in whole body PET images and in other protocols requiring short transmission scans [10]. An additional advantage of reconstructing attenuation maps in PET is that if the patient moves between the transmission and emission scan, and if one can estimate this motion, then one can calculate appropri-

ate attenuation correction factors by reprojecting the attenuation map at the proper angles.

The traditional approach to tomographic image reconstruction is based on the nonstatistical filtered backprojection method [11, 12]. The FBP method and the data-weighted least-squares method [13, 14] for transmission image reconstruction both lead to systematic biases for low-count scans [14–16]. These biases are due to the nonlinearity of the logarithm applied to the transmission data. To eliminate these biases, one can use statistical methods based on the Poisson measurement statistics. These method use the *raw measurements* rather than the logarithms thereof [14, 17–19]. Statistical methods also produce images with lower variance than FBP [14, 16, 20]. Thus, in this chapter we focus on statistical methods.

The organization of this chapter is as follows. Section 1.2 first reviews the low-count tomographic reconstruction problem. Section 1.3 gives an overview of the principles underlying optimization algorithms for image reconstruction. Section 1.4 through Section 1.7 describe in detail four categories of reconstruction algorithms: expectation maximization, coordinate ascent, paraboloidal surrogates, and direct algorithms. All these algorithms are presented for the Poisson statistical model; Section 1.8 summarizes alternatives to that approach. Section 1.9 briefly summarizes application of the algorithms to emission reconstruction. Section 1.10 gives an overview of some advanced topics. Section 1.11 presents illustrative results for real PET transmission scans.

A few of the algorithms presented are "new" in the sense that they are derived here under more realistic assumptions than were made in some of the original papers. And we have provided simple extensions to some algorithms that were not intrinsically monotonic as previously published, but can be made monotonic by suitable modifications.

1.2 The problem

In transmission tomography, the quantity of interest is the spatial distribution of the linear attenuation coefficient, denoted $\mu(\vec{x}, \mathcal{E})$, where $\vec{x} = (x_1, x_2, x_3)$ denotes spatial location in 3-space, and the argument $\mathcal{E}$ parameterizes the dependence of the attenuation coefficient on incident photon energy [12]. The units of μ are typically inverse centimeters (cm^{-1}). If the object (patient) is moving, or if there are variations due to, e.g., flowing contrast agent, then we could also denote the temporal dependence. For simplicity we assume the object is static in this chapter. The ideal transmission imaging modality would provide a complete description of $\mu(\vec{x}, \mathcal{E})$ for a wide range of energies, at infinitesimal spatial and temporal resolutions, at a modest price, and with no harm to the subject. In practice we settle for much less.

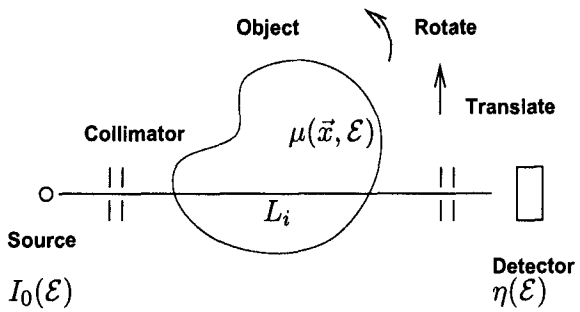

Figure 1.1: Transmission scanning geometry for a 1st-generation CT scanner.

1.2.1 Transmission measurements

The methods described in this chapter are applicable to general transmission geometries. However, the problem is simplest to describe in the context of a first-generation CT scanner as illustrated in Fig. 1.1. A collimated source of photons with intensity $I_0(\mathcal{E})$ is transmitted through the attenuating object and the transmitted photons are recorded by a detector with detector efficiency $\eta(\mathcal{E})$. The source and detector are translated and rotated around the object. We use the letter i to index the source/detector locations, where $i = 1, \ldots, N_Y$. Typically N_Y is the product of the number of radial positions assumed by the source for each angular position times the number of angular positions. For 2D acquisitions, in SPECT transmission scans $N_Y \approx 10^4$; in PET transmission scans, $N_Y \approx 10^5$; and in modern x-ray CT systems $N_Y \approx 10^6$.

For simplicity, we assume that the collimation eliminates scattered photons, which is called the "narrow beam" geometry. In PET and SPECT transmission scans, the source is usually a monoenergetic radioisotope[1] that emits gamma photons with a single energy $\mathcal{E}_0$, i.e.,

$$I_0(\mathcal{E}) = I_0 \delta(\mathcal{E} - \mathcal{E}_0),$$

where $\delta(\cdot)$ is the Dirac delta function. For simplicity, we assume this monoenergetic case hereafter. (In the polyenergetic case, one must consider effects such as *beam hardening* [21].)

The absorption and Compton scattering of photons by the object is governed by Beer's law. Let b_i denote the mean number of photons that would be recorded by the detector (for the ith source-detector position, hereafter referred to as a "ray") if the object were absent. This b_i depends on the scan duration, the source strength, and the detector efficiency at the source photon energy $\mathcal{E}_0$. The dependence on i reflects the fact that in modern systems there are multiple detectors, each of which

[1]Some gamma emitting radioisotopes produce photons at two or more distinct energies. If the detector has adequate energy resolution, then it can separate photons at the energy of interest from other photons, or bin the various energies separately.

can have its own efficiency. By Beer's law, the mean number of photons recorded for the ith ray ideally would be [12]

$$b_i \exp\left(- \int_{L_i} \mu_0(\vec{x}) \, dl \right) , \tag{1.1}$$

where L_i is the line or strip between the source and detector for the ith ray, and where

$$\mu_0(\vec{x}) \stackrel{\triangle}{=} \mu(\vec{x}, \mathcal{E}_0)$$

is the linear attenuation coefficient at the source photon energy. The number of photons actually recorded in practice differs from the ideal expression (1.1) in several ways. First, for a photon-counting detector[2], the number of recorded photons is a Poisson random variable [12]. Second, there will usually be additional "background" counts recorded due to Compton scatter [22], room background, random coincidences in PET [23,24], or emission crosstalk in SPECT [9,25–27]. Third, the detectors have finite width, so the infinitesimal line integral in (1.1) is an approximation. For accurate image reconstruction, one must incorporate these effects into the statistical model for the measurements, rather than simply using the idealized model (1.1).

Let Y_i denote the random variable representing the number of photons counted for the ith ray. A reasonable statistical model[3] for these transmission measurements is that they are independent Poisson random variables with means given by

$$E[Y_i] = b_i \exp\left(- \int_{L_i} \mu_0(\vec{x}) \, dl \right) + r_i , \tag{1.2}$$

where r_i denotes the mean number of background events (such as random coincidences, scatter, and crosstalk). In many papers, the r_i's are ignored or assumed to be zero. In this chapter, we assume that the r_i's are known, which in practice means that they are determined separately by some other means (such as smoothing a delayed-window sinogram in PET transmission scans [31]). The noise in these estimated r_i's is not considered here, and is a subject requiring further investigation and analysis. In some PET transmission scans, the random coincidences are subtracted from the measurements in real time. Statistical methods for treating this problem have been developed [32–34], and require fairly simple modifications of the algorithms presented in this chapter.

We assume the b_i's are known. In PET and SPECT centers, these are determined by periodic "blank scans": transmission scans with nothing but air in the

[2]For a current integrating detector, such as those used in commercial x-ray CT scanners, the measurement noise is a mixture of Poisson photon statistics and gaussian electronic noise.

[3]Due to the effects of detector deadtime in PET and SPECT, the measurement distributions are not exactly Poisson [28–30], but the Poisson approximation seems adequate in practice.

scanner portal. Since no patient is present, these scans can have fairly long durations (typically a couple of hours, run automatically in the middle of the night). Thus the estimated b_i's computed from such a long scan have much less variability than the transmission measurements (Y_i's). Therefore, we ignore the variability in these estimated b_i's. Accounting for the small variability in the b_i estimates is another open problem (but one likely to be of limited practical significance).

1.2.2 Reconstruction problem

After acquiring a transmission scan, the tomographic reconstruction problem is to estimate $\mu_0(\vec{x})$ from a realization $\{y_i = Y_i\}_{i=1}^{N_Y}$ of the measurements. This collection of measurements is usually called a *sinogram*[4] [35]. The conventional approach to this problem is to first estimate the ith line integral from the model (1.2) and then to apply the FBP algorithm to the collection of line-integral estimates. Specifically, let

$$l_i^{\text{true}} \triangleq \int_{L_i} \mu_0(\vec{x})\, dl$$

denote the true line integral along the ith ray. Conventionally, one forms an estimate $\hat{l}_i$ of l_i^{true} by computing the logarithm of the measured sinogram as follows:

$$\hat{l}_i = \begin{cases} \log\left(\frac{b_i}{Y_i - r_i}\right), & Y_i > r_i \\ ?, & Y_i \leq r_i\,. \end{cases} \tag{1.3}$$

One then reconstructs an estimate $\hat{\mu}^{\text{FBP}}$ from $\{\hat{l}_i\}_{i=1}^{N_Y}$ using FBP [35]. There are several problems with this approach. First, the logarithm is not defined when $Y_i \leq r_i$, which can happen frequently in low-count transmission scans. (Typically one must substitute some artificial value (denoted "?" above) for such rays, or interpolate neighboring rays [36], which can lead to biases.) Second, the above procedure yields *biased* estimates of the line-integral. By Jensen's inequality, since $-\log x$ is a concave function, for any random variable X,

$$E[-\log X] \geq -\log E[X]$$

(see [37], p. 50), so when applied to (1.2) and (1.3)

$$\begin{aligned} E[\hat{l}_i] &= E\left[\log\left(\frac{b_i}{Y_i - r_i}\right)\right] = E\left[-\log\left(\frac{Y_i - r_i}{b_i}\right)\right] \\ &\geq -\log\left(E\left[\frac{Y_i - r_i}{b_i}\right]\right) = \int_{L_i} \mu_0(\vec{x})\, dl = l_i^{\text{true}}\,. \end{aligned} \tag{1.4}$$

Thus, the logarithm in (1.3) *systematically over-estimates* the line integral on average. This over-estimation has been verified empirically [14, 15]. One can show

[4]When the ray measurements are organized as a 2D array according their radial and angular coordinates, the projection of a point object appears approximately as a sinusoidal trace in the array.

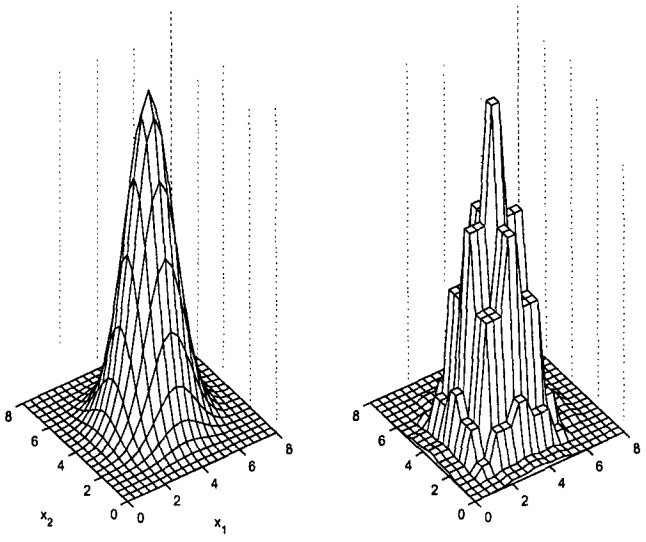

Figure 1.2: Illustration of 2D function $\mu_0(x_1, x_2)$ parameterized using the pixel basis (1.6).

analytically that the bias increases as the counts decrease [14], so the logarithm is particularly unsuitable for low-count scans. A third problem with (1.3) is that the variances of the $\hat{l}_i$'s can be quite nonuniform, so some rays are much more informative than other rays. The FBP method treats all rays equally, even those for which $Y_i - r_i$ is non-positive, which leads to noisy images corrupted by streaks originating from high variance $\hat{l}_i$'s. Noise is considered only as an afterthought by apodizing the ramp filter, which is equivalent to space-invariant smoothing. (There are a few exceptions where space-variant sinogram filtering has been applied, e.g., [38–40].) Fourth, the FBP method is poorly suited to nonstandard imaging geometries, such as truncated fan-beam or cone-beam scans, e.g., [41–46].

Since noise is a primary concern, the image reconstruction problem is naturally treated as a statistical estimation problem. Since we only have a finite number N_Y of measurements, it is natural to also represent $\mu_0(\vec{x})$ with a finite parameterization. Such parameterizations are reasonable in practice since ultimately the estimate of μ_0 will be viewed on a digital display with a finite number of pixels. After one has parameterized μ_0, the reconstruction problem becomes a statistical problem: estimate the parameters from the noisy measurements $\{Y_i\}_{i=1}^{N_Y}$.

A general approach to parameterizing the attenuation map is to expand it in terms of a finite basis expansion [47,48]:

$$\mu_0(\vec{x}) = \sum_{j=1}^{N_p} \mu_j \chi_j(\vec{x}) \,, \tag{1.5}$$

where N_p is the number of coefficients μ_j and basis functions $\chi_j(\vec{x})$. There are

many possible choices for the basis functions. We would like to choose basis functions that naturally represent nonnegative functions since $\mu_0 \geq 0$. We would also like basis functions that have compact support, since such a basis yields a very sparse system matrix $\boldsymbol{A}$ in (1.9) below. The conventional basis is just the "pixel" or "voxel" basis, which satisfies both of these requirements. The voxel basis $\chi_j(\vec{x})$ is 1 inside the jth voxel, and is 0 everywhere else. In two-space, one can express the pixel basis by

$$\chi_j(x,y) = \text{rect}\left(\frac{x-x_j}{\Delta}\right) \text{rect}\left(\frac{y-y_j}{\Delta}\right), \tag{1.6}$$

where (x_j, y_j) is the center of the jth pixel and Δ is the pixel width. This basis gives a piecewise-constant approximation to μ_0, as illustrated in Fig. 1.2. With any parameterization of the form (1.5), the problem of estimating $\mu_0(\vec{x})$ is reduced to the simpler problem of estimating the parameter vector $\underline{\mu} = [\mu_1, \ldots, \mu_{N_p}]'$ from the measurement vector $\underline{Y} = [Y_1, \ldots, Y_{N_Y}]'$, where "$'$" denotes vector and matrix transpose. Under the parameterization (1.5), the line integral in (1.2) becomes the following summation:

$$\int_{L_i} \mu_0(\vec{x})\, dl = \int_{L_i} \sum_{j=1}^{N_p} \mu_j \chi_j(\vec{x})\, dl = \sum_{j=1}^{N_p} \mu_j \int_{L_i} \chi_j(\vec{x})\, dl = \sum_{j=1}^{N_p} a_{ij}\mu_j ,$$

where

$$a_{ij} \stackrel{\triangle}{=} \int_{L_i} \chi_j(\vec{x})\, dl$$

is the line integral[5] along the ith ray through the jth basis function. This simplification yields the following discrete-discrete measurement model:

$$Y_i \sim \text{Poisson}\left\{\bar{y}_i(\underline{\mu}_{\text{true}})\right\}, \; i = 1, \ldots, N_Y , \tag{1.7}$$

where the ensemble mean of the ith measurement is denoted

$$\bar{y}_i(\underline{\mu}) \stackrel{\triangle}{=} b_i e^{-[\boldsymbol{A}\underline{\mu}]_i} + r_i \tag{1.8}$$

$$[\boldsymbol{A}\underline{\mu}]_i \stackrel{\triangle}{=} \sum_{j=1}^{N_p} a_{ij}\mu_j , \tag{1.9}$$

where $\boldsymbol{A} = \{a_{ij}\}$. The remainder of this chapter will be based on the Poisson measurement model (1.7). (See Appendix 1.14 for a review of Poisson statistics.)

[5] In practice, we use normalized strip integrals [49, 50] rather than line integrals to account for finite detector width [51]. Regardless, the units of a_{ij} are length units (mm or cm), whereas the units of the μ_j's are inverse length.

1.2.3 Likelihood-based estimation

Maximum-likelihood (ML) estimation is a natural approach[6] for finding $\underline{\mu}$ from a particular measurement realization $\underline{Y} = \underline{y}$ when a statistical model such as (1.7) is available. The ML estimate $\hat{\underline{\mu}}$ is defined as follows:

$$\hat{\underline{\mu}} = \arg\max_{\underline{\mu} \geq \underline{0}} L(\underline{\mu}) \,, \quad L(\underline{\mu}) \stackrel{\triangle}{=} \log P[\underline{Y} = \underline{y}; \underline{\mu}] \,.$$

For the Poisson model (1.7), the measurement joint probability mass function is

$$P[\underline{Y} = \underline{y}; \underline{\mu}] = \prod_{i=1}^{N_Y} P[Y_i = y_i; \underline{\mu}] = \prod_{i=1}^{N_Y} e^{-\bar{y}_i(\underline{\mu})} [\bar{y}_i(\underline{\mu})]^{y_i} / y_i! \,. \tag{1.10}$$

The ML method seeks the object (as described by the parameter vector $\underline{\mu}$) that maximizes the probability of having observed the particular measurements that were recorded. The first paper to propose a ML approach for transmission tomography appears to be due to Rockmore and Macovski in 1977 [47]. However, the pseudoinverse method described in [47] in general does not find the maximizer of the likelihood $L(\underline{\mu})$.

For independent transmission measurements, we can use (1.8) and (1.10) to express the log-likelihood in the following convenient form:

$$L(\underline{\mu}) \equiv \sum_{i=1}^{N_Y} h_i([\boldsymbol{A}\underline{\mu}]_i) \tag{1.11}$$

where we use "$\equiv$" hereafter for expressions that are equal up to irrelevant constants independent of $\underline{\mu}$, and where the marginal log-likelihood of the ith measurement is

$$h_i(l) \stackrel{\triangle}{=} y_i \log(b_i e^{-l} + r_i) - (b_i e^{-l} + r_i) \,. \tag{1.12}$$

A typical h_i is shown in Fig. 1.4 on page 17. For convenience later, we also list the derivatives of h_i here:

$$\dot{h}_i(l) \stackrel{\triangle}{=} \frac{d}{dl} h_i(l) = \left[1 - \frac{y_i}{b_i e^{-l} + r_i}\right] b_i e^{-l} \tag{1.13}$$

$$\ddot{h}_i(l) \stackrel{\triangle}{=} \frac{d^2}{dl^2} h_i(l) = -\left[1 - \frac{y_i r_i}{(b_i e^{-l} + r_i)^2}\right] b_i e^{-l} \,. \tag{1.14}$$

[6]The usual rationale for the ML approach is that ML estimators are asymptotically unbiased and asymptotically efficient (minimum variance) under very general conditions [37]. Such asymptotic properties alone would be a questionable justification for the ML approach in the case of low-count transmission scans. However, ML estimators often perform well even in the "non-asymptotic" regime. We are unaware of any data-fit measure for low-count transmission scans that outperforms the log-likelihood, but there is no known proof of optimality of the log-likelihood in this case, so the question is an open one.

The algorithms described in the following sections are based on various strategies for finding the maximizer of $L(\underline{\mu})$. Several of the algorithms are quite general in the sense that one can easily modify them to apply to many objective functions of the form (1.11), even when h_i has a functional form different from the form (1.12) that is specific to transmission measurements. Thus, even though the focus of this chapter is transmission imaging, many of the algorithms and comments apply equally to emission reconstruction and to other inverse problems.

Maximizing the log-likelihood $L(\cdot)$ alone leads to unacceptably noisy images, because tomographic image reconstruction is an ill-conditioned problem. Roughly speaking, this means that there are many choices of attenuation maps $\mu_0(\vec{x})$ that fit the measurements $\{Y_i\}_{i=1}^{N_Y}$ reasonably well. Even when the problem is parameterized, there are many choices of the vector $\underline{\mu}$ that fit the measurements $\{Y_i\}_{i=1}^{N_Y}$ reasonably well, where the fit is quantified by the log-likelihood $L(\underline{\mu})$. Not all of those images are useful or physically plausible. Thus, the likelihood alone does not adequately identify the "best" image. One effective remedy to this problem is to modify the objective function by including a *penalty function* that favors reconstructed images that are piecewise smooth. This process is called *regularization* since the penalty function improves the conditioning of the problem[7]. In this chapter we focus on methods that form an estimate $\hat{\underline{\mu}}$ of the true attenuation map $\underline{\mu}_{\text{true}}$ by maximizing a *penalized-likelihood* objective function of the following form:

$$\hat{\underline{\mu}} \stackrel{\triangle}{=} \arg\max_{\underline{\mu} \geq \underline{0}} \Phi(\underline{\mu}), \quad \Phi(\underline{\mu}) \stackrel{\triangle}{=} L(\underline{\mu}) - \beta R(\underline{\mu}), \tag{1.15}$$

where the objective function Φ includes a roughness penalty $R(\underline{\mu})$ discussed in more detail below. The parameter β controls the tradeoff between spatial resolution and noise: larger values of β generally lead to reduced noise at the price of reduced spatial resolution. Solving (1.15) is the primary subject of this chapter.

One benefit of using methods that are based on objective functions such as (1.15) is that for such methods, image quality is determined by the objective function rather than by the particular iterative algorithm, provided the iterative algorithm converges to the maximizer of the objective function. In particular, from the objective function one can analyze spatial resolution properties and bias, variance, and autocorrelation properties [16, 20, 54–59].

1.2.3.1 Connection to Bayesian perspective

By letting $\beta = 0$ in (1.15) and in the algorithms presented in the following sections, one has ML algorithms as a special case. Bayesian image reconstruc-

[7]For emission tomography, a popular alternative approach to "regularization" is simply to post-smooth the ML reconstruction image with a Gaussian filter. In the emission case, under the somewhat idealized assumption of a shift-invariant Gaussian blur model for the system, a certain commutability condition ((12) of [52] holds, which ensures that Gaussian post-filtering is equivalent to Gaussian sieves. It is unclear whether this equivalence holds in the transmission case, although some authors have implied that it does without proof, e.g. [53].

tion formulations also lead to objective functions of the form (1.15). Suppose one considers $\underline{\mu}$ to be a random vector drawn from a prior distribution $f(\underline{\mu})$ that is proportional to $e^{-\beta R(\underline{\mu})}$. (Such priors arise naturally in the context of Markov random field models for images [60].) One computes the maximum *a posteriori* (MAP) estimate of $\underline{\mu}$ by maximizing the posterior distribution $f(\underline{\mu}|\underline{Y})$. By Bayes rule:

$$f(\underline{\mu}|\underline{Y}) = f(\underline{Y}|\underline{\mu})f(\underline{\mu})/f(\underline{Y})$$

so the log posterior is

$$\log f(\underline{\mu}|\underline{Y}) \equiv \log f(\underline{Y}|\underline{\mu}) + \log f(\underline{\mu}) \equiv L(\underline{\mu}) - \beta R(\underline{\mu}) .$$

Thus MAP estimation is computationally equivalent to (1.15).

1.2.4 Penalty function

It has been considered by many authors to be reasonable to assume that the attenuation maps of interest are piecewise smooth functions. An extreme example of such assumptions is the common use of attenuation map segmentation in PET imaging to reduce the noise due to attenuation correction factors [61–63]. Under the piecewise smooth attenuation map assumption, it is reasonable for the penalty function $R(\underline{\mu})$ to discourage images that are too "rough." The simplest penalty function that discourages roughness considers the discrepancies between neighboring pixel values:

$$R(\underline{\mu}) = \sum_{j=1}^{N_p} \frac{1}{2} \sum_{k=1}^{N_p} w_{jk} \psi(\mu_j - \mu_k) , \tag{1.16}$$

where $w_{jk} = w_{kj}$. Ordinarily $w_{jk} = 1$ for the four horizontal and vertical neighboring pixels, $w_{jk} = 1/\sqrt{2}$ for diagonal neighboring pixels, and $w_{jk} = 0$ otherwise. One can also adopt the modifications described in [20, 54–57] to provide more uniform spatial resolution. The *potential function* ψ assigns a cost to $\mu_j - \mu_k$.

For the results presented in Section 1.11, we used a penalty function of the form (1.16). However, the methods we present all apply to much more general penalty functions. Such generality is needed for penalty functions such as the weak-plate prior of [64] or the local averaging function considered in [65]. One can express most[8] of the penalty functions that have been used in tomographic reconstruction in the following very general form:

$$R(\underline{\mu}) = \sum_{k=1}^{K} \psi_k([\boldsymbol{C}\underline{\mu}]_k) \tag{1.17}$$

[8] One exception is the median root prior [66].

where $\boldsymbol{C}$ is a $K \times N_p$ penalty matrix and

$$[\boldsymbol{C}\underline{\mu}]_k = \sum_{j=1}^{N_p} c_{kj}\mu_j \,.$$

We assume throughout that the functions ψ_k are symmetric and differentiable.

The pairwise penalty (1.16) is the special case of (1.17) where $K \approx 2N_p$ and each row of $\boldsymbol{C}$ has one $+1$ and one -1 entry corresponding to some pair of pixels.

In this chapter we focus on quadratic penalty functions where $\psi_k(t) = \omega_k t^2/2$ for $\omega_k \geq 0$, so

$$R(\underline{\mu}) = \sum_{k=1}^{K} \omega_k \frac{1}{2}([\boldsymbol{C}\underline{\mu}]_k)^2 = \frac{1}{2}\underline{\mu}'\boldsymbol{C}'\boldsymbol{\Omega}\boldsymbol{C}\underline{\mu}\,, \tag{1.18}$$

where $\boldsymbol{\Omega} \stackrel{\triangle}{=} \mathrm{diag}\{\omega_k\}$. The second derivative of such a penalty is given by

$$\frac{\partial^2}{\partial \mu_j^2} R(\underline{\mu}) = \sum_{k=1}^{K} c_{kj}^2 \omega_k \,. \tag{1.19}$$

This focus is for simplifying the presentation; in practice nonquadratic penalty functions are often preferable for transmission image reconstruction e.g. [67, 68].

1.2.5 Concavity

From the second derivative expression (1.14), when $r_i = 0$, $\ddot{h}_i(l) = -b_i e^{-l}$, which is always nonpositive, so h_i is concave over all of $\mathbb{R}$ (and strictly concave if $b_i > 0$). From (1.11) one can easily verify that the Hessian matrix (the $N_p \times N_p$ matrix of second partial derivatives) of $L(\cdot)$ is:

$$\nabla^2 L(\underline{\mu}) = \boldsymbol{A}' \,\mathrm{diag}\left\{ \ddot{h}_i([\boldsymbol{A}\underline{\mu}]_i) \right\} \boldsymbol{A}\,, \tag{1.20}$$

where $\mathrm{diag}\{d_i\}$ is a $N_Y \times N_Y$ diagonal matrix with ith diagonal element d_i. Thus the log-likelihood is concave over all of $\mathbb{R}^{N_p}$ when $r_i = 0\ \forall i$. If the ψ_k's are all strictly convex and $L(\cdot)$ is concave, then the objective Φ is strictly concave under mild conditions on $\boldsymbol{A}$ [69]. Such concavity is central to the convergence proofs of the algorithms described below. In the case $r_i \neq 0$, the likelihood $L(\cdot)$ is not necessarily concave. Nevertheless, in our experience it seems to be unimodal. (Initializing monotonic iterative algorithms with different starting images seems to lead to the same final image.) In the non-concave case, we cannot guarantee global convergence to the global maximum for any of the algorithms described below. For the monotonic algorithms we usually can prove convergence to a local maximum [70]; if in addition the objective function is unimodal, then the only local maximum will be the global maximum, but proving that Φ is unimodal is an open problem.

1.3 Optimization algorithms

Ignoring the nonnegativity constraint, one could attempt to find $\hat{\underline{\mu}}$ analytically by zeroing the gradient of Φ. The partial derivatives of Φ are

$$\frac{\partial}{\partial \mu_j} \Phi(\underline{\mu}) = \sum_{i=1}^{N_Y} a_{ij} \dot{h}_i([\boldsymbol{A}\underline{\mu}]_i) - \beta \frac{\partial}{\partial \mu_j} R(\underline{\mu}) \,, \quad j = 1, \ldots, N_p \,, \tag{1.21}$$

where $\dot{h}_i$ was defined in (1.13). Unfortunately, even disregarding both the nonnegativity constraint and the penalty function, there are no closed-form solutions to the set of equations (1.21), except in the trivial case when $\boldsymbol{A} = \boldsymbol{I}$. Even when $\boldsymbol{A} = \boldsymbol{I}$ there are no closed-form solutions for nonseparable penalty functions. Thus iterative methods are required to find the maximizer $\hat{\underline{\mu}}$ of such objective functions.

1.3.1 Why so many algorithms?

Analytical solutions for the maximizer of (1.15) appear intractable, so one must use iterative algorithms. An *iterative algorithm* is a procedure that is initialized with an initial guess $\underline{\mu}^{(0)}$ of $\underline{\mu}$, and then recursively generates a sequence $\underline{\mu}^{(1)}, \underline{\mu}^{(2)}, \ldots$, also denoted $\{\underline{\mu}^{(n)}\}$. Ideally, the iterates $\{\underline{\mu}^{(n)}\}$ should rapidly approach the maximizer $\hat{\underline{\mu}}$. When developing algorithms for image reconstruction based on penalized-likelihood objective functions, there are many design considerations, most of which are common to any problem involving iterative methods. In particular, an algorithm designer should consider the impact of design choices on the following characteristics.

- Monotonicity ($\Phi(\underline{\mu}^{(n)})$ increases every iteration)
- Nonnegativity constraint ($\underline{\mu} \geq \underline{0}$)
- Parallelization
- Sensitivity to numerical errors
- Convergence rate (as few iterations as possible)
- Computation time per iteration (as few floating point operations as possible)
- Storage requirements (as little memory as possible)
- Memory bandwidth (data access)

Generic numerical methods such as steepest ascent do not exploit the specific structure of the objective function Φ, nor do they easily accommodate the nonnegativity constraint. Thus for fastest convergence, one must seek algorithms tailored to this type of problem. Some of the relevant properties of L include:

- $L(\underline{\theta})$ is a sum of scalar functions $h_i(\cdot)$.
- The h_i's have bounded curvature, and are concave when $r_i = 0$.
- The arguments of the functions h_i are inner products.
- The inner product coefficients are all nonnegative.

The cornucopia of algorithms that have been proposed in the image reconstruction literature exploit these properties (implicitly or explicitly) in different ways.

1.3.2 Optimization transfer principle

Before delving into the details of the many algorithms that have been proposed for maximizing Φ, we first describe a very useful and intuitive general principle that underlies almost all the methods. The principle is called *optimization transfer*. This idea was described briefly as a *majorization principle* in the limited context of 1D line searches in the classic text by Ortega and Rheinbolt [71, p. 253]. It was rediscovered and generalized to inverse problems in the recent work of De Pierro [72, 73] and Lange [69, 74]. Since the concept applies more generally than just to transmission tomography, we use $\underline{\theta}$ as the generic unknown parameter here.

The basic idea is illustrated in Fig. 1.3. Since Φ is difficult to maximize, at the nth iteration we can replace Φ with a *surrogate function* $\phi(\underline{\theta}; \underline{\theta}^{(n)})$ that is easier to maximize, i.e., the next iterate is defined as:

$$\underline{\theta}^{(n+1)} \stackrel{\triangle}{=} \arg\max_{\underline{\theta}} \phi(\underline{\theta}; \underline{\theta}^{(n)}) . \tag{1.22}$$

The maximization is restricted to the valid parameter space (e.g. $\underline{\theta} \geq \underline{0}$ for problems with nonnegative constraints). Maximizing $\phi(\cdot; \underline{\theta}^{(n)})$ will usually not lead directly to the global maximizer $\hat{\underline{\mu}}$. Thus one repeats the process iteratively, finding a new surrogate function ϕ at each iteration and then maximizing that surrogate function. If we choose the surrogate functions appropriately, then the sequence $\{\underline{\theta}^{(n)}\}$ should eventually converge to the maximizer $\hat{\underline{\mu}}$ [75].

Fig. 1.3 does not do full justice to the problem, since 1D functions are usually fairly easy to maximize. The optimization transfer principle is particularly compelling for problems where the dimension of $\underline{\theta}$ is large, such as in inverse problems like tomography.

It is very desirable to use algorithms that monotonically increase Φ each iteration, i.e., for which $\Phi(\underline{\theta}^{(n+1)}) \geq \Phi(\underline{\theta}^{(n)})$. Such algorithms are guaranteed to be stable, i.e., the sequence $\{\underline{\theta}^{(n)}\}$ will not diverge if Φ is concave. And generally such algorithms will converge to the maximizer $\hat{\underline{\mu}}$ if it is unique [70]. If we choose surrogate functions that satisfy

$$\Phi(\underline{\theta}) - \Phi(\underline{\theta}^{(n)}) \geq \phi(\underline{\theta}; \underline{\theta}^{(n)}) - \phi(\underline{\theta}^{(n)}; \underline{\theta}^{(n)}), \ \forall \underline{\theta}, \underline{\theta}^{(n)} , \tag{1.23}$$

then one can see immediately that the algorithm (1.22) monotonically increases Φ. To ensure monotonicity, it is not essential to find the exact maximizer in (1.22). It suffices to find a value $\underline{\theta}^{(n+1)}$ such that $\phi(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) \geq \phi(\underline{\theta}^{(n)}; \underline{\theta}^{(n)})$, since that alone will ensure $\Phi(\underline{\theta}^{(n+1)}) \geq \Phi(\underline{\theta}^{(n)})$ by (1.23). The various algorithms described in the sections that follow are all based on different choices of the surrogate function ϕ, and on different procedures for the maximization in (1.22).

Rather than working with (1.23), all the surrogate functions we present satisfy

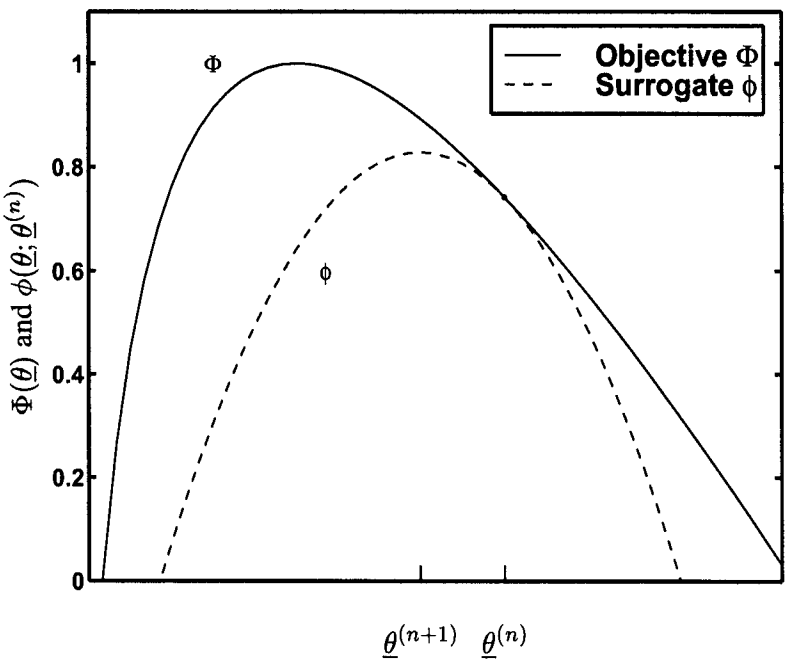

Figure 1.3: Illustration of optimization transfer in 1D.

the following conditions:

$$\phi(\underline{\theta}^{(n)}; \underline{\theta}^{(n)}) = \Phi(\underline{\theta}^{(n)}) \tag{1.24}$$

$$\nabla_{\underline{\theta}} \phi(\underline{\theta}; \underline{\theta}^{(n)})\Big|_{\underline{\theta}=\underline{\theta}^{(n)}} = \nabla \Phi(\underline{\theta})|_{\underline{\theta}=\underline{\theta}^{(n)}} \tag{1.25}$$

$$\phi(\underline{\theta}; \underline{\theta}^{(n)}) \leq \Phi(\underline{\theta}) \; \forall \underline{\theta} \geq \underline{0}. \tag{1.26}$$

Any surrogate function that satisfies these conditions will satisfy (1.23). (The middle condition follows from the outer two conditions when Φ and ϕ are differentiable.)

1.3.3 Convergence rate

The convergence rate of an iterative algorithm based on the optimization transfer principle can be analyzed qualitatively by considering Fig. 1.3. If the surrogate function ϕ has low curvature, then it appears as a "broad" graph in Fig. 1.3, which means that the algorithm can take large steps ($\underline{\theta}^{(n+1)} - \underline{\theta}^{(n)}$ can be large) which means that it reaches the maximizer faster. Conversely, if the surrogate function has high curvature, then it appears as a "skinny" graph, the steps are small, and many steps are required for convergence. So in general we would like to find low curvature surrogate functions, with the caveat that we want to maintain $\phi \leq \Phi$ to ensure monotonicity [76]. And of course we would also like the surrogate ϕ to be

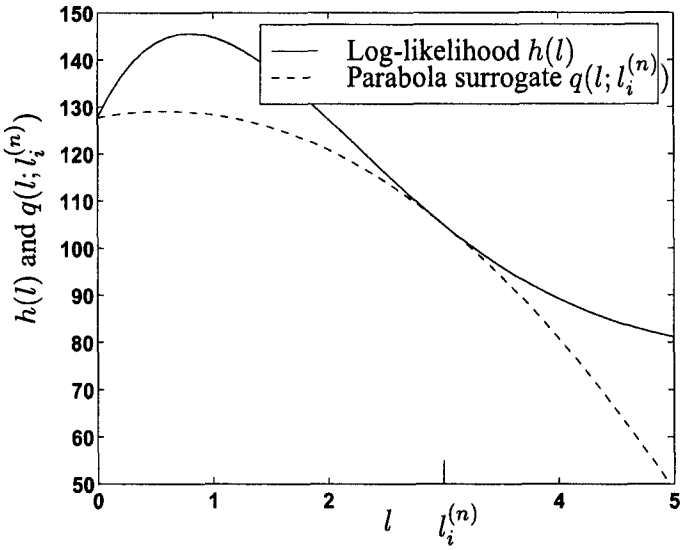

Figure 1.4: Ray log-likelihood $h_i(l)$ and parabola surrogate $q_i(l; l_i^{(n)})$ for $y_i = 50$, $b_i = 100$, $r_i = 5$, and $l_i^{(n)} = 3$.

easy to maximize for (1.22). Unfortunately, the criteria "low curvature" and "easy to maximize" are often incompatible, so we must compromise.

1.3.4 Parabola surrogate

Throughout this chapter we struggle with the ray log-likelihood function h_i defined in (1.12), which we rewrite here without the i's for simplicity:

$$h(l) = y \log(b\, e^{-l} + r) - (b\, e^{-l} + r)\,,$$

where $y, b, r \geq 0$ are known constants. Of the many possible surrogate functions that could replace $h(\cdot)$ in an optimization transfer approach, a choice that is particularly convenient is a parabola surrogate:

$$q(l; l_i^{(n)}) \triangleq h(l_i^{(n)}) + \dot{h}(l_i^{(n)})(l - l_i^{(n)}) - \frac{c}{2}(l - l_i^{(n)})^2\,, \tag{1.27}$$

for some choice of the curvature $c \geq 0$, where

$$l_i^{(n)} = [\boldsymbol{A}\underline{\mu}^{(n)}]_i \tag{1.28}$$

is the ith line integral through the estimated attenuation map at the nth iteration. The choice (1.27) clearly satisfies conditions (1.24) and (1.25), but we must carefully choose $c = c(l_i^{(n)}, y, b, r)$ to ensure that $q(l; l_i^{(n)}) \leq h(l)\ \forall l \geq 0$ so that (1.26) is satisfied. On the other hand, from the convergence rate description in the preceding section, we would like the curvature c to be as small as possible. In other

words, we would like to find

$$c(l_i^{(n)}, y, b, r) \triangleq$$

$$\min \left\{ c : h(l_i^{(n)}) + \dot{h}(l_i^{(n)})(l - l_i^{(n)}) - \frac{c}{2}(l - l_i^{(n)})^2 \leq h(l),\ \forall l \geq 0 \right\} .$$

In [77], we showed that the optimal curvature is as follows:

$$c(l_i^{(n)}, y, b, r) = \begin{cases} \left[-2 \dfrac{h(0) - h(l_i^{(n)}) + \dot{h}(l_i^{(n)}) l_i^{(n)}}{(l_i^{(n)})^2} \right]_+ , & l_i^{(n)} > 0 \\ \left[-\ddot{h}(l_i^{(n)}) \right]_+ , & l_i^{(n)} = 0 , \end{cases} \tag{1.29}$$

where $[x]_+$ is x for positive x and zero otherwise. Fig. 1.4 illustrates the surrogate parabola q in (1.27) with the optimal curvature (1.29).

One small inconvenience with (1.29) is that it changes every iteration since it depends on $l_i^{(n)}$. An alternative choice of the curvature that ensures $q \leq h$ is the maximum second derivative of $-h(l)$ over $[0, \infty)$. In [77] we show that

$$\max_{l \geq 0} \left[-\ddot{h}(l) \right]_+ = \left[-\ddot{h}(0) \right]_+ = \left[\left(1 - \frac{y}{b + r} \right) b \right]_+ . \tag{1.30}$$

We can precompute this curvature before iterating since it is independent of $l_i^{(n)}$. However, typically this curvature is much larger than the optimal choice (1.29), so the floating point operations (flops) saved by precomputing may be lost in increased number of iterations due to a slower convergence rate.

The surrogate parabola (1.27) with curvature (1.29) will be used repeatedly in this chapter, both for the derivation of recently developed algorithms, as well as for making minor improvements to older algorithms that were not monotonic as originally proposed. A similar approach applies to the emission reconstruction problem [78].

1.4 EM algorithms

The emission reconstruction algorithm derived by Shepp and Vardi in [79] and by Lange and Carson in [17] is often referred to as "the" EM algorithm in the nuclear imaging community. In fact, the expectation-maximization (EM) framework is a general method for developing many different algorithms [80]. The appeal of the EM framework is that it leads to iterative algorithms that in principle yield sequences of iterates that monotonically increase the objective function. Furthermore, in many statistical problems one can derive EM algorithms that are quite simple to implement. Unfortunately, the Poisson transmission reconstruction problem does not seem to be such a problem. Only one basic type of EM algorithm

has been proposed for the transmission problem, and that algorithm converges *very* slowly and has other difficulties described below. We include the description of the EM algorithm for completeness, but the reader who is not interested in the historical perspective could safely skip this section since we present much more efficient algorithms in subsequent sections.

We describe the general EM framework in the context of problems where one observes a realization $\underline{y}$ of a measurement vector $\underline{Y}$, and wishes to estimate a parameter vector $\underline{\theta}$ by maximizing the likelihood or penalized log-likelihood. To develop an EM algorithm, one must first postulate a hypothetical collection of random variables called the "complete data space" $\underline{X}$. These are random variables that, in general, were not observed during the experiment, but that might have simplified the estimation procedure had they been observed. The only requirement that the complete data space must satisfy is that one must be able to extract the observed data from $\underline{X}$, i.e. there must exist a function $h(\cdot)$ such that

$$\underline{Y} = h(\underline{X}) . \tag{1.31}$$

This is a trivial requirement since one can always include the random vector $\underline{Y}$ itself in the collection $\underline{X}$ of random variables.

Having judiciously chosen $\underline{X}$, an essential ingredient of any EM algorithm is the following conditional expectation of the log-likelihood of $\underline{X}$:

$$Q(\underline{\theta}; \underline{\theta}^{(n)}) = E[\log f(\underline{X}; \underline{\theta}) \mid \underline{Y} = \underline{y}; \underline{\theta}^{(n)}] = \int f(\underline{x} \mid \underline{Y} = \underline{y}; \underline{\theta}^{(n)}) \log f(\underline{x}; \underline{\theta}) \, d\underline{x}, \tag{1.32}$$

and, in the context of penalized-likelihood problems, the following function

$$\phi(\underline{\theta}; \underline{\theta}^{(n)}) \stackrel{\triangle}{=} Q(\underline{\theta}; \underline{\theta}^{(n)}) - \beta R(\underline{\theta}) , \tag{1.33}$$

where $R(\underline{\theta})$ is a penalty function. We refer to ϕ as an *EM-based surrogate function*, since one replaces the difficult problem of maximizing Φ with a sequence of (hopefully) simpler maximizations of ϕ. There are often alternative surrogate functions that have advantages over the EM-based functions, as described in Section 1.6.1. The surrogate function concept is illustrated in Fig. 1.3 on page 16.

An EM algorithm is initialized at an arbitrary point $\underline{\theta}^0$ and generates a sequence of iterates $\underline{\theta}^1, \underline{\theta}^2, \ldots$. Under fairly general conditions [81], the sequence $\{\underline{\theta}^{(n)}\}$ converges to the maximizer of the objective function $\Phi(\underline{\theta}) = L(\underline{\theta}) - \beta R(\underline{\theta})$. The EM recursion is as follows:

E-step: find $Q(\underline{\theta}; \underline{\theta}^{(n)})$ using (1.32) and $\phi(\underline{\theta}; \underline{\theta}^{(n)})$ using (1.33)

M-step: $\underline{\theta}^{(n+1)} = \arg\max_{\underline{\theta}} \phi(\underline{\theta}; \underline{\theta}^{(n)})$,

where the maximization is restricted to the set of valid parameters, as in (1.22). Many generalizations of this basic framework have been proposed, see for example [82–86] and a recent review paper [87].

By applying Jensen's inequality, one can show [80] that the EM-based surrogate function satisfies the monotonicity inequality (1.23) for all $\underline{\theta}$ and $\underline{\theta}^{(n)}$. Since the M-step ensures that

$$\phi(\underline{\theta}^{(n+1)}; \underline{\theta}^{(n)}) \geq \phi(\underline{\theta}^{(n)}; \underline{\theta}^{(n)}) ,$$

it follows from (1.23) that the above EM recursion is guaranteed to lead to monotone increases in the objective function Φ. Thus, the EM framework is a special case of the optimization transfer approach (1.22), where the surrogate function derives from statistical principles (1.32).

1.4.1 Transmission EM algorithm

Although Rockmore and Macovski proposed ML transmission reconstruction in 1977 [47], it took until 1984 for the first practical algorithm to appear, when Lange and Carson proposed a complete data space for a transmission EM algorithm [17]. Lange and Carson considered the case where $r_i = 0$. In this section we derive a transmission EM algorithm that generalizes that of [17]; we allow $r_i \neq 0$, and we consider arbitrary pixel orderings, as described below. The algorithm of [17] is a special case of what follows.

The "complete" data space that we consider for the case $r_i \neq 0$ is the following collection of random variables, all of which have Poisson marginal distributions:

$$\underline{X} = \{Y_{ik} \ : \ i = 0, 1, \ldots, N_Y, \ k = 1, \ldots, N_p\} \cup \{R_i \ : \ i = 1, \ldots, N_Y\} .$$

We assume that

$$R_i \sim \text{Poisson}\{r_i\}$$

and

$$Y_{i0} \sim \text{Poisson}\{b_i\} , \ i = 1, \ldots, N_Y ,$$

and that the R_i's are all mutually independent and statistically independent of all the Y_{ik}'s. Furthermore, Y_{ik} and Y_{lk} are independent for $i \neq l$. However, Y_{ik} and Y_{ij} are not independent. (The distributions of R_i and Y_{i0} do not depend on $\underline{\mu}$, so are of less importance in what follows.)

For each i, let $(j_{i,1}, \ldots, j_{i,N_p})$ be *any* permutation of the set of pixel indices $j = 1, \ldots, N_p$. Notationally, the simplest case is just when $j_{i,k} = k$, which corresponds to the algorithm considered in [88]. Lange and Carson [17] assign $j_{i,k}$ to the physical ordering corresponding to the ray connecting the source to the detector. Statistically, any ordering suffices, and it is an open (and probably academic) question whether certain orderings lead to faster convergence. Given such an ordering, we define the remaining Y_{ik}'s recursively by the following conditional distributions:

$$Y_{ik} \mid Y_{i,k-1}, Y_{i,k-2}, \ldots, Y_{i0} \sim \text{Binomial}\left(Y_{i,k-1}, e^{-\lambda_{i,k}}\right) ,$$

where

$$\lambda_{i,k} \stackrel{\triangle}{=} a_{ij_{i,k}} \mu_{j_{i,k}} .$$

Thus the conditional probability mass function (PMF) of Y_{ik} for $k = 1, \ldots, N_p$ is given by:

$$p(y_{ik} \,|y_{i,k-1}, y_{i,k-2}, \cdots, y_{i0}; \underline{\mu}) = p(y_{ik} \,|y_{i,k-1}; \underline{\mu})$$

$$= \left(\begin{array}{c} y_{i,k-1} \\ y_{ik} \end{array} \right) \left(e^{-\lambda_{i,k}}\right)^{y_{ik}} \left(1 - e^{-\lambda_{i,k}}\right)^{y_{i,k-1} - y_{ik}} , \qquad (1.34)$$

where $\left(\begin{array}{c} n \\ m \end{array} \right)$ is the binomial coefficient. An alternative way of writing this recursion is

$$Y_{ik} = \sum_{l=1}^{Y_{i,k-1}} Z_{ikl}, \quad k = 1, \ldots, N_p, \quad i = 1, \ldots, N_Y ,$$

where Z_{ikl} is a collection of independent 0-1 Bernoulli random variables with

$$P[Z_{ikl} = 1] = e^{-\lambda_{i,k}} .$$

Since a Bernoulli-thinned Poisson process remains Poisson (see Appendix 1.14), it follows that each Y_{ik} has a Poisson distribution:

$$Y_{ik} \sim \text{ Poisson}\left\{ b_i \prod_{l=1}^{k} e^{-\lambda_{i,l}} \right\} \equiv \text{ Poisson}\left\{ b_i e^{-\sum_{l=1}^{k} \lambda_{i,l}} \right\} . \qquad (1.35)$$

The special case $k = N_p$ in (1.35) yields

$$Y_{i,N_p} \sim \text{ Poisson}\left\{ b_i e^{-\sum_{l=1}^{N_p} \lambda_{i,l}} \right\} \equiv \text{ Poisson}\left\{ b_i e^{-[\boldsymbol{A}\underline{\mu}]_i} \right\} .$$

Therefore, the observed measurements are related to the complete data space by

$$Y_i = Y_{i,N_p} + R_i ,$$

so the condition (1.31) is satisfied. As noted in [89], there are multiple orderings of the Y_{ik}'s that can be considered, each of which would lead to a different update, but which would leave unchanged the limit if there is a unique maximizer (and provided this EM algorithm is globally convergent, which has never been established for the case $r_i \neq 0$).

Figure 1.5 illustrates a loose physical interpretation of the above complete data space. For the ith ordering, imagine a sequence of layers of material with attenuation coefficients $\mu_{j_{i,1}}, \ldots, \mu_{j_{i,N_p}}$ and thicknesses $a_{ij_{i,1}}, \ldots, a_{ij_{i,N_p}}$. Suppose a

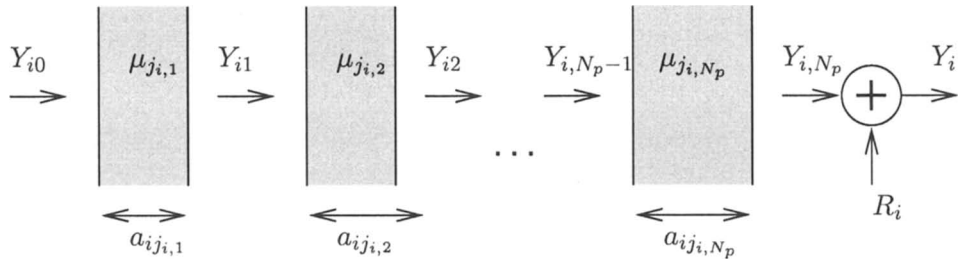

Figure 1.5: Pseudo-physical interpretation of transmission EM complete-data space.

Poisson number Y_{i0} of photons is transmitted into the first layer. The number that survive passage through that layer is Y_{i1}, which then proceed through the second layer and so on. The final number of photons exiting the sequence of layers is Y_{i,N_p}, and this number is added to the random coincidences R_i to form the observed counts Y_i. This interpretation is most intuitive when the pixels are ordered according to the actual passage of photons from source to detector (as in [17]), but a physical interpretation is not essential for EM algorithm development.

It follows from (1.35) and Appendix 1.14 that

$$Y_{i,N_p} | Y_{ik} \sim \text{Binomial}\left(Y_{ik}, \prod_{l=k+1}^{N_p} e^{-\lambda_{i,l}} \right) ,$$

since a cascade of independent Binomials is Binomial with the product of the success probabilities.

Having specified the complete-data space, the next step in developing an EM algorithm is to find the surrogate function $Q(\underline{\mu}; \underline{\mu}^{(n)})$ of (1.32). It follows from the above specifications that the joint probability mass function (PMF) of $\underline{X}$ is given by

$$p(\underline{X}; \underline{\mu}) = \prod_{i=1}^{N_Y} p(\{Y_{ik}\}_{k=0}^{N_p}; \underline{\mu}) \prod_{i=1}^{N_Y} p(R_i) .$$

By applying the chain rule for conditional probabilities [90] and using (1.34):

$$p(\{Y_{ik}\}_{k=0}^{N_p}; \underline{\mu}) = p(Y_{i0}) \prod_{k=1}^{N_p} p(Y_{ik} | Y_{i,k-1}; \underline{\mu})$$

so

$$\begin{aligned}\log p(\underline{X};\underline{\mu}) &\equiv \sum_{i=1}^{N_Y}\sum_{k=1}^{N_p}\log p(Y_{ik}\,|Y_{i,k-1};\underline{\mu}) \\ &\equiv \sum_{i=1}^{N_Y}\sum_{k=1}^{N_p} Y_{ik}\log\left(e^{-\lambda_{i,k}}\right)+(Y_{i,k-1}-Y_{ik})\log\left(1-e^{-\lambda_{i,k}}\right).\end{aligned}$$

Thus, following [17, 88], the EM-based surrogate function for the above complete data space has the following form:

$$\begin{aligned}Q(\underline{\mu};\underline{\mu}^{(n)}) &= \sum_{i=1}^{N_Y}\sum_{j=1}^{N_p}\bar{N}_{ij}^{(n)}\log\left(e^{-a_{ij}\mu_j}\right)+(\bar{M}_{ij}^{(n)}-\bar{N}_{ij}^{(n)})\log(1-e^{-a_{ij}\mu_j}) \\ &= \sum_{j=1}^{N_p} Q_j(\mu_j;\underline{\mu}^{(n)}),\end{aligned} \tag{1.36}$$

where

$$Q_j(\mu_j;\underline{\mu}^{(n)}) \triangleq \sum_{i=1}^{N_Y}\bar{N}_{ij}^{(n)}\log\left(e^{-a_{ij}\mu_j}\right)+(\bar{M}_{ij}^{(n)}-\bar{N}_{ij}^{(n)})\log(1-e^{-a_{ij}\mu_j}) \tag{1.37}$$

$$\bar{N}_{ij}^{(n)} \triangleq E[Y_{ik}\,|Y_i=y_i;\underline{\mu}^{(n)}]\Big|_{k\,:\,j_{i,k}=j} \tag{1.38}$$

$$\bar{M}_{ij}^{(n)} \triangleq E[Y_{i,k-1}\,|Y_i=y_i;\underline{\mu}^{(n)}]\Big|_{k\,:\,j_{i,k}=j}. \tag{1.39}$$

To complete the E-step of the EM algorithm, we must find the preceding conditional expectations. Using the law of iterated expectation [90]:

$$\begin{aligned}E[Y_{ik}\,|Y_i=y_i] &= E\left[E[Y_{ik}\,|R_i,Y_i=y_i]\,|\,Y_i=y_i\right] \\ &= E\left[E[Y_{ik}\,|Y_{i,N_p}=y_i-R_i]\,|\,Y_i=y_i\right] \\ &= E\left[E[Y_{ik}]-E[Y_{i,N_p}]+y_i-R_i\,|\,Y_i=y_i\right] \\ &= E[Y_{ik}]-E[Y_{i,N_p}]+y_i-y_i\frac{r_i}{r_i+E[Y_{i,N_p}]}\end{aligned} \tag{1.40}$$

using results from [17] and Appendix 1.14. From (1.35)

$$E[Y_{ij};\underline{\mu}^{(n)}]=\gamma_{ij}^{(n)} \triangleq b_i\prod_{l=1}^{k}e^{-a_{ij_{i,l}}\mu_{j_{i,l}}^{(n)}}=b_i e^{-\sum_{l=1}^{k}a_{ij_{i,l}}\mu_{j_{i,l}}^{(n)}}\Bigg|_{k:j_{i,k}=j}, \tag{1.41}$$

from which one can show using (1.40) that

$$\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)} = \gamma_{ij}^{(n)}(e^{a_{ij}\mu_j^{(n)}} - 1) \geq 0 . \tag{1.42}$$

Combining (1.36), (1.37), (1.40), and (1.41) yields an explicit expression for the EM surrogate $Q(\underline{\mu}; \underline{\mu}^{(n)})$, completing the E-step.

For the M-step of the EM algorithm, we must find the maximizer of $Q(\cdot; \underline{\mu}^{(n)})$. The function $Q(\cdot; \underline{\mu}^{(n)})$ is a *separable function* of the μ_j's, as shown in (1.36), so it is easier to maximize than Φ. Thus the M-step reduces to the N_p separable 1D maximization problems:

$$\mu_j^{(n+1)} = \arg\max_{\mu_j \geq 0} Q_j(\mu_j; \underline{\mu}^{(n)}) . \tag{1.43}$$

Unfortunately however, due to the transcendental functions in (1.36), there is no closed-form expression for the maximizer of $Q_j(\cdot; \underline{\mu}^{(n)})$. In fact, finding the maximizer of $Q_j(\cdot; \underline{\mu}^{(n)})$ is no easier than maximizing $\Phi(\underline{\mu})$ with respect to μ_j while holding the other parameters fixed, which is the coordinate ascent algorithm described in Section 1.5. Nevertheless, the EM algorithm is parallelizable, unlike the coordinate ascent algorithm, so we proceed here with its description. Zeroing the derivative of Q_j in (1.37) yields the following:

$$0 = \frac{d}{d\mu_j} Q_j(\mu_j; \underline{\mu}^{(n)}) = \sum_{i=1}^{N_Y} -\bar{N}_{ij}^{(n)} a_{ij} + (\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)}) \frac{a_{ij} e^{-a_{ij}\mu_j}}{1 - e^{-a_{ij}\mu_j}} , \tag{1.44}$$

the solution to which is $\mu_j^{(n+1)}$. Unfortunately, (1.44) can only be solved for μ_j analytically when the a_{ij}'s are all equal. However, Lange and Carson [17] noted that typically $a_{ij}\mu_j$ will be small (*much* less than unity), so that the following Taylor-series expansion around zero should be a reasonable approximation:

$$\frac{x e^{-x}}{1 - e^{-x}} \approx 1 - \frac{x}{2} \text{ for } x \approx 0 .$$

Applying this approximation to (1.44) with $x = a_{ij}\mu_j$ yields:

$$\begin{aligned} \sum_{i=1}^{N_Y} \bar{N}_{ij}^{(n)} a_{ij} &\approx \sum_{i=1}^{N_Y} (\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)}) \frac{1}{\mu_j} \left(1 - \frac{a_{ij}\mu_j}{2}\right) \\ &= \sum_{i=1}^{N_Y} (\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)}) \left(\frac{1}{\mu_j} - \frac{a_{ij}}{2}\right) . \end{aligned}$$

Solving this equality for μ_j yields the final iterative form for the ML transmission EM algorithm [17]:

$$\boxed{\mu_j^{(n+1)} \approx \frac{\sum_{i=1}^{N_Y} (\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)})}{\frac{1}{2} \sum_{i=1}^{N_Y} a_{ij} (\bar{M}_{ij}^{(n)} + \bar{N}_{ij}^{(n)})} .} \tag{1.45}$$

This algorithm is very slow to converge [69] and each iteration is very computationally expensive due to the large number of exponentiations required in (1.41). One exponentiation per nonzero a_{ij} is required.

Lange and Carson [17] also describe an update based on a second-order Taylor series, and they note that one can use their expansion to find upper and lower bounds for the exact value of μ_j that maximizes $Q_j(\cdot; \underline{\mu}^{(n)})$.

1.4.2 EM algorithms with approximate M-steps

Since the M-step of the transmission EM algorithm of [17] did not yield a closed form for the maximizer, Browne and Holmes [91] proposed a modified EM algorithm that used an approximate M-step based on image rotations using bilinear interpolation. Kent and Wright made a similar approximation [89]. An advantage of these methods is that (after interpolation) the a_{ij}'s are all equal, which is the case where one can solve (1.44) analytically. Specifically, if $a_{ij} = a_0$ for all i and j, then (1.44) simplifies to

$$\sum_{i=1}^{N_Y} \tilde{N}_{ij}^{(n)} = \sum_{i=1}^{N_Y} (\tilde{M}_{ij}^{(n)} - \tilde{N}_{ij}^{(n)}) \frac{1}{e^{a_0 \mu_j} - 1},$$

where $\tilde{N}_{ij}^{(n)}$ and $\tilde{M}_{ij}^{(n)}$ replace $\bar{M}_{ij}^{(n)}$ and $\bar{N}_{ij}^{(n)}$ respectively, in rotated coordinates. When solved for μ_j, this yields the iteration

$$\boxed{\mu_j^{(n+1)} = \frac{1}{a_0} \log \left(\frac{\sum_{i=1}^{N_Y} \tilde{M}_{ij}^{(n)}}{\sum_{i=1}^{N_Y} \tilde{N}_{ij}^{(n)}} \right),} \tag{1.46}$$

which is the logarithm of the ratio of (conditional expectations of) the number of photons entering the jth pixel to the number of photons leaving the jth pixel, divided by the pixel size. However, the interpolations required to form $\tilde{N}_{ij}^{(n)}$ and $\tilde{M}_{ij}^{(n)}$ presumably destroy the monotonicity properties of the EM algorithm. Although bookkeeping is reduced, these methods require the same (very large) number of exponentiations as the original transmission EM algorithm, so they are also impractical algorithms.

1.4.3 EM algorithm with Newton M-step

Ollinger [92,93] reported that the M-step approximation (1.45) proposed in [17] led to convergence problems, and proposed a 1D Newton's method for maximizing Q in the context of a GEM algorithm for the M-step. Since Newton's method is not guaranteed to converge, the step length was adjusted by a halving strategy to

ensure a monotone increase in the surrogate function:

$$\mu_j^{(n+1)} = \mu_j^{(n)} + \alpha_{j,n} \frac{\frac{d}{d\mu_j} Q_j(\mu_j; \underline{\mu}^{(n)}) \Big|_{\mu_j = \mu_j^{(n)}}}{\frac{d^2}{d\mu_j^2} Q_j(\mu_j; \underline{\mu}^{(n)}) \Big|_{\mu_j = \mu_j^{(n)}}}, \tag{1.47}$$

where one ensures that $Q_j(\mu_j^{(n+1)}; \underline{\mu}^{(n)}) \geq Q_j(\mu_j^{(n)}; \underline{\mu}^{(n)})$ by choosing $\alpha_{j,n}$ via a line-search. This line-search can require multiple evaluations of Q_j as each pixel is updated, which is relatively expensive. The large number of exponentiations required to compute $\bar{M}_{ij}^{(n)}$ and $\bar{N}_{ij}^{(n)}$ also remains a drawback.

From (1.44) and (1.42),

$$\begin{aligned}
\frac{d}{d\mu_j} Q_j(\mu_j; \underline{\mu}^{(n)}) \Big|_{\mu_j = \mu_j^{(n)}} &= \sum_{i=1}^{N_Y} a_{ij} \left[\frac{\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)}}{1 - e^{-a_{ij}\mu_j^{(n)}}} e^{-a_{ij}\mu_j^{(n)}} - \bar{N}_{ij}^{(n)} \right] \\
&= \sum_{i=1}^{N_Y} a_{ij} \left(\gamma_{ij}^{(n)} - \left[\gamma_{ij}^{(n)} - b_i e^{-l_i^{(n)}} + \frac{y_i b_i e^{-l_i^{(n)}}}{b_i e^{-l_i^{(n)}} + r_i} \right] \right) \\
&= \sum_{i=1}^{N_Y} a_{ij} \left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i} \right) b_i e^{-l_i^{(n)}} = \frac{d}{d\mu_j} L(\underline{\mu}) \Big|_{\underline{\mu} = \underline{\mu}^{(n)}},
\end{aligned}$$

so (1.25) is indeed satisfied. From (1.44) and (1.42),

$$-\frac{d^2}{d\mu_j^2} Q_j(\mu_j; \underline{\mu}^{(n)}) = \sum_{i=1}^{N_Y} (\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)}) \frac{a_{ij}^2 e^{-a_{ij}\mu_j}}{(1 - e^{-a_{ij}\mu_j})^2}, \tag{1.48}$$

so

$$-\frac{d^2}{d\mu_j^2} Q_j(\mu_j; \underline{\mu}^{(n)}) \Big|_{\mu_j = \mu_j^{(n)}} = \sum_{i=1}^{N_Y} a_{ij}^2 \gamma_{ij}^{(n)} / (1 - e^{-a_{ij}\mu_j^{(n)}}).$$

Thus, the ML EM Newton-Raphson (EM-NR) algorithm (1.47) becomes

$$\boxed{\mu_j^{(n+1)} = \mu_j^{(n)} + \alpha_{j,n} \frac{\sum_{i=1}^{N_Y} a_{ij} \left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i} \right) b_i e^{-l_i^{(n)}}}{\sum_{i=1}^{N_Y} a_{ij}^2 \gamma_{ij}^{(n)} / (1 - e^{-a_{ij}\mu_j^{(n)}})}.} \tag{1.49}$$

From (1.48), the curvature of Q_j becomes unbounded as $\mu_j \to 0$, which appears to preclude the use of parabola surrogates as described in Section 1.4.5.2 to form an intrinsically monotonic M-step (1.43).

Variations on the transmission EM algorithm continue to resurface at conferences, despite its many drawbacks. The endurance of the transmission EM algorithm can only be explained by its having "ridden on the coat tails" of the popular emission EM algorithm. The modern methods described in subsequent sections are entirely preferable to the transmission EM algorithm.

1.4.4 Diagonally-scaled gradient-ascent algorithms

Several authors, e.g. [94], noted that the emission EM algorithm can be expressed as a diagonally-scaled, gradient-ascent algorithm, with a particular diagonal scaling matrix that (almost miraculously) ensures monotonicity and preserves nonnegativity. (The EM-NR algorithm (1.49) has a similar form.) Based on an analogy with that emission EM algorithm, Lange et al. proposed a diagonally-scaled gradient-ascent algorithm for transmission tomography [95]. The algorithm can be expressed as the following recursion:

$$\boxed{\underline{\mu}^{(n+1)} = \underline{\mu}^{(n)} + D(\underline{\mu}^{(n)})\nabla' L(\underline{\mu}^{(n)}),} \tag{1.50}$$

where $D(\cdot)$ is some iteration-dependent diagonal matrix and where ∇' denotes the column gradient operator (cf (1.21)):

$$\left[\nabla' L(\underline{\mu})\right]_j = \frac{\partial}{\partial \mu_j} L(\underline{\mu}) = \sum_{i=1}^{N_Y} a_{ij} \left[1 - \frac{y_i}{b_i e^{-[\boldsymbol{A}\underline{\mu}]_i} + r_i}\right] b_i e^{-[\boldsymbol{A}\underline{\mu}]_i}. \tag{1.51}$$

Since the gradient of the objective function is evaluated once per iteration, the number of exponentiations required is roughly N_Y, far fewer than required by the transmission EM algorithm (1.45).

The choice of the $N_p \times N_p$ diagonal scaling matrix $D(\underline{\mu})$ critically affects convergence rate, monotonicity, and nonnegativity. Considering the case $r_i = 0$, Lange et al. [95] suggested the following diagonal scaling matrix, chosen so that (1.50) could be expressed as a multiplicative update in the case $r_i = 0$:

$$D(\underline{\mu})_{jj} = \frac{\mu_j}{\sum_{i=1}^{N_Y} a_{ij} y_i}. \tag{1.52}$$

The natural generalization of this choice to the general case where $r_i \neq 0$ is the diagonal matrix with the following expression for the jth diagonal element:

$$D(\underline{\mu})_{jj} = \frac{\mu_j}{\sum_{i=1}^{N_Y} a_{ij} y_i b_i e^{-[\boldsymbol{A}\underline{\mu}]_i} / (b_i e^{-[\boldsymbol{A}\underline{\mu}]_i} + r_i)}. \tag{1.53}$$

Using (1.51) and (1.53) one can rewrite the diagonally-scaled gradient ascent (DSGA) algorithm (1.50) as follows:

$$\boxed{\mu_j^{(n+1)} = \mu_j^{(n)} \frac{\sum_{i=1}^{N_Y} a_{ij} b_i e^{-l_i^{(n)}}}{\sum_{i=1}^{N_Y} a_{ij} y_i b_i e^{-l_i^{(n)}} / (b_i e^{-l_i^{(n)}} + r_i)}.} \tag{1.54}$$

This is a multiplicative update that preserves nonnegativity, and at least its positive fixed points are stationary points of the log-likelihood. However, the particular choice of diagonal scaling matrix (1.53) does not guarantee intrinsically monotone

increases in the likelihood function. In Section 1.6.2 below, we present one form of a paraboloidal surrogates algorithm that has the same general form as (1.50) but overcomes the limitations of (1.54) by choosing D appropriately.

Lange et al. also proposed modifications of the iteration (1.50) to include a separable penalty function and a line-search to enforce monotonicity [95] (for $r_i = 0$ case, but the ideas generalize easily to the $r_i \neq 0$ case).

Lange proposed another diagonally-scaled gradient-ascent algorithm in [96], based on the following diagonal scaling matrix:

$$D(\underline{\mu})_{jj} = \frac{\mu_j}{\sum_{i=1}^{N_Y} a_{ij} b_i e^{-[\boldsymbol{A}\underline{\mu}]_i} [\boldsymbol{A}\underline{\mu}]_i} . \tag{1.55}$$

Although the rationale for this choice was not given in [96], Lange was able to show that the algorithm has local convergence properties, but that it may not yield nonnegative estimates. Lange further modified the scaled-gradient algorithm in [97] to include nonseparable penalty functions, and a practical approximate line-search that ensures global convergence for $r_i = 0$.

Considering the case $r_i = 0$, Maniawski et al. [98] proposed the following over-relaxed unregularized version of the diagonally-scaled gradient-ascent algorithm (1.50):

$$\boxed{\mu_j^{(n+1)} = \mu_j^{(n)} \left[\omega \frac{\sum_{i=1}^{N_Y} a_{ij} b_i e^{-l_i^{(n)}}}{\sum_{i=1}^{N_Y} a_{ij} y_i} + (1 - \omega) \right] ,} \tag{1.56}$$

where ω was selected empirically to be $4 \cdot 10^{-8}$ times the total number of measured counts in a SPECT transmission scan. Like (1.54), this is a multiplicative update that preserves nonnegativity. One can also express the above algorithm more generally as follows:

$$\underline{\mu}^{(n+1)} = (1 - \omega)\underline{\mu}^{(n)} + \omega D(\underline{\mu}^{(n)}) \nabla' L(\underline{\mu}^{(n)}) ,$$

where $D(\underline{\mu})$ is chosen as in (1.52). No convergence analysis was discussed for the algorithm (1.56), although "fast convergence" was reported.

1.4.5 Convex algorithm

De Pierro [72] described a non-statistical derivation of the emission EM algorithm using the concavity properties of the log-likelihood for emission tomography. Lange and Fessler [69] applied a similar derivation to the transmission log-likelihood for the case $r_i = 0$, yielding a "convex[9] algorithm" that, like the transmission EM algorithm, is guaranteed to monotonically increase $L(\underline{\mu})$ each iteration. As discussed in Section 1.2.5, the transmission log-likelihood is concave

[9]The algorithm name is unfortunate, since the algorithm itself is not convex, but rather the algorithm is derived by exploiting the concavity of the log-likelihood.

when $r_i = 0$, so De Pierro's convexity method could be applied directly in [69]. In the case $r_i \neq 0$, the log-likelihood is not concave, so De Pierro's convexity argument does not directly apply. Fessler [14] noted that even when $r_i \neq 0$, the marginal log-likelihood functions (the h_i's in (1.11)) are concave over a (typically) large interval of the real line, and thereby developed an "approximate" convex algorithm. However, the "convex algorithm" of [14] is not guaranteed to be globally monotonic.

Rather than presenting either the convex algorithm of [69], which is incomplete since it did not consider the case $r_i \neq 0$, or the algorithm of [14], which is non-monotone, we derive a new "convex" algorithm here. The algorithm of [69] falls out as a special case of this new algorithm by setting $r_i = 0$. The idea is to first use the EM algorithm to find a concave surrogate function Q_1 that "eliminates" the r_i terms, but is still difficult to maximize directly; then we apply De Pierro's convexity argument to Q_1 to find another surrogate function Q_2 that is easily maximized. The same idea was developed independently by Kim [99].

Consider a "complete" data space that is the collection of the following statistically independent random variables:

$$\underline{X} = \{\{N_i\}_{i=1}^{N_Y}, \{R_i\}_{i=1}^{N_Y}\},$$

where

$$\begin{aligned} N_i &\sim \text{Poisson}\left\{b_i e^{-[\boldsymbol{A}\underline{\mu}]_i}\right\} \\ R_i &\sim \text{Poisson}\{r_i\} \end{aligned}$$

and where the observed measurements are related to the elements of $\underline{X}$ by

$$Y_i = N_i + R_i\,,$$

so the condition (1.31) is satisfied. The complete-data log-likelihood is simply:

$$\log p(\underline{X};\underline{\mu}) \equiv \sum_{i=1}^{N_Y} N_i \log\left(b_i e^{-[\boldsymbol{A}\underline{\mu}]_i}\right) - \left(b_i e^{-[\boldsymbol{A}\underline{\mu}]_i}\right)\,,$$

since the distribution of the R_i's is a constant independent of $\underline{\mu}$, so can be ignored. Since by Appendix 1.14

$$\bar{N}_i^{(n)} \triangleq E[N_i\,|Y_i = y_i; \underline{\mu}^{(n)}] = y_i \frac{b_i e^{-l_i^{(n)}}}{b_i e^{-l_i^{(n)}} + r_i}\,,$$

the EM surrogate function is the following concave function:

$$Q_1(\underline{\mu};\underline{\mu}^{(n)}) = \sum_{i=1}^{N_Y} -\bar{N}_i^{(n)}[\boldsymbol{A}\underline{\mu}]_i - b_i e^{-[\boldsymbol{A}\underline{\mu}]_i} = \sum_{i=1}^{N_Y} g_i^{(n)}([\boldsymbol{A}\underline{\mu}]_i) \qquad (1.57)$$

where

$$g_i^{(n)}(l) \triangleq -\bar{N}_i^{(n)} l - b_i e^{-l} . \tag{1.58}$$

The form of (1.57) is identical to the form that (1.11) would have if all the r_i's were zero. Therefore, by this technique we can generalize any algorithm that has been derived for the $r_i = 0$ case to the realistic case where $r_i \neq 0$ simply by replacing y_i in the algorithm with $\bar{N}_i^{(n)}$. However, in general the convergence of an algorithm derived this way may be slower than methods based on direct maximization of $L(\underline{\mu})$ since the curvatures of the Q_1 components are smaller than those of L since $\bar{N}_i^{(n)} \leq y_i$. For rays where the random fraction is large, $\bar{N}_i^{(n)} \ll y_i$, leading to slow convergence rates. In statistical terms, the complete-data space $\underline{X}$ is much more informative than the observed data $\underline{Y}$ [100].

We could attempt to naively perform the M-step of the EM algorithm:

$$\underline{\mu}^{(n+1)} = \arg \max_{\underline{\mu} \geq \underline{0}} Q_1(\underline{\mu}; \underline{\mu}^{(n)}) ,$$

except that maximizing Q_1 is (almost[10]) as difficult as maximizing the original log-likelihood.

By differentiating twice, one can easily show that each $g_i^{(n)}$ is a concave function and that $Q_1(\cdot; \underline{\mu}^{(n)})$ is a concave function. Therefore, rather than maximizing Q_1 directly, we find a surrogate function for Q_1 by applying the convexity method of De Pierro [72]. The essence of De Pierro's method is the following clever expression for matrix-vector multiplication:

$$[\boldsymbol{A}\underline{\mu}]_i = \sum_{j=1}^{N_p} \alpha_{ij} \left[\frac{a_{ij}}{\alpha_{ij}} (\mu_j - \mu_j^{(n)}) + l_i^{(n)} \right] \tag{1.59}$$

where the projection of the current attenuation map estimate is given by

$$l_i^{(n)} \triangleq [\boldsymbol{A}\underline{\mu}^{(n)}]_i .$$

The expression (1.59) holds for any collection of α_{ij}'s, provided $\alpha_{ij} = 0$ only if $a_{ij} = 0$ and

$$\sum_{j=1}^{N_p} \alpha_{ij} = 1 .$$

If we choose nonnegative α_{ij}'s, then because each $g_i^{(n)}$ is concave, by (1.59):

$$g_i^{(n)}([\boldsymbol{A}\underline{\mu}]_i) = g_i^{(n)} \left(\sum_{j=1}^{N_p} \alpha_{ij} \left[\frac{a_{ij}}{\alpha_{ij}} (\mu_j - \mu_j^{(n)}) + l_i^{(n)} \right] \right) \geq \sum_{j=1}^{N_p} \alpha_{ij}\, g_{ij}^{(n)}(\mu_j; \underline{\mu}^{(n)}) , \tag{1.60}$$

[10] Q_1 is concave, unlike L.

where

$$g_{ij}^{(n)}(\mu_j;\underline{\mu}^{(n)}) \triangleq g_i^{(n)}\left(\frac{a_{ij}}{\alpha_{ij}}(\mu_j-\mu_j^{(n)})+l_i^{(n)}\right).$$

Thus, a suitable surrogate function for $Q_1(\underline{\mu};\underline{\mu}^{(n)})$ is

$$Q_2(\underline{\mu};\underline{\mu}^{(n)}) \triangleq \sum_{i=1}^{N_Y}\sum_{j=1}^{N_p}\alpha_{ij}\,g_{ij}^{(n)}(\mu_j;\underline{\mu}^{(n)}) = \sum_{j=1}^{N_p} Q_{2,j}(\mu_j;\underline{\mu}^{(n)}) \tag{1.61}$$

where

$$Q_{2,j}(\mu_j;\underline{\mu}^{(n)}) \triangleq \sum_{i=1}^{N_Y}\alpha_{ij}\,g_{ij}^{(n)}(\mu_j;\underline{\mu}^{(n)}) = \sum_{i=1}^{N_Y}\alpha_{ij}g_i^{(n)}\left(\frac{a_{ij}}{\alpha_{ij}}(\mu_j-\mu_j^{(n)})+l_i^{(n)}\right). \tag{1.62}$$

Since Q_2 is a *separable function*, its maximization reduces to N_p simultaneous maximization problems:

$$\mu_j^{(n+1)} = \arg\max_{\mu_j\geq 0} Q_{2,j}(\mu_j;\underline{\mu}^{(n)}). \tag{1.63}$$

Unfortunately there is not a closed-form analytical solution for the maximizer, so we must apply approximations, line searches, or the optimization transfer principle.

1.4.5.1 Convex-NR algorithms

A simple "solution" to (1.63) is to apply one or more Newton-Raphson steps, as in (1.47). Such an algorithm should be locally convergent, and can presumably be made globally convergent by a line-search modification of the type proposed by Lange [97]. From (1.62), (1.57), and (1.13):

$$\begin{aligned}
\frac{d}{d\mu_j}Q_{2,j}(\mu_j;\underline{\mu}^{(n)})\Big|_{\mu_j=\mu_j^{(n)}} &= \sum_{i=1}^{N_Y} a_{ij}\dot{g}_i^{(n)}(l_i^{(n)}) \\
&= \sum_{i=1}^{N_Y} a_{ij}\left[b_i e^{-l_i^{(n)}} - \bar{N}_i^{(n)}\right] = \sum_{i=1}^{N_Y} a_{ij}\left[b_i e^{-l_i^{(n)}} - y_i\frac{b_i e^{-l_i^{(n)}}}{b_i e^{-l_i^{(n)}}+r_i}\right] \\
&= \sum_{i=1}^{N_Y} a_{ij}\left(1-\frac{y_i}{b_i e^{-l_i^{(n)}}+r_i}\right) b_i e^{-l_i^{(n)}} = \frac{\partial}{\partial\mu_j}L(\underline{\mu})\Big|_{\underline{\mu}=\underline{\mu}^{(n)}},
\end{aligned}$$

so

$$\nabla' Q_2(\underline{\mu};\underline{\mu}^{(n)})\Big|_{\underline{\mu}=\underline{\mu}^{(n)}} = \nabla' L(\underline{\mu};\underline{\mu}^{(n)})\Big|_{\underline{\mu}=\underline{\mu}^{(n)}}.$$

Thus the new convex algorithm (1.63), based on one 1D Newton-Raphson step for each pixel, has the same form as the diagonally-scaled gradient-ascent algorithm (1.50), except that it uses the following diagonal scaling matrix:

$$[D(\underline{\mu}^{(n)})]_{jj} = \frac{1}{-\frac{d^2}{d\mu_j^2} Q_{2,j}(\mu_j; \underline{\mu}^{(n)})\Big|_{\mu_j = \mu_j^{(n)}}},$$

where from (1.62) and (1.58)

$$-\frac{d^2}{d\mu_j^2} Q_{2,j}(\mu_j; \underline{\mu}^{(n)})\Big|_{\mu_j = \mu_j^{(n)}} = -\sum_{i=1}^{N_Y} \frac{a_{ij}^2}{\alpha_{ij}} \ddot{g}_i^{(n)}(\mu_j; \underline{\mu}^{(n)}) = \sum_{i=1}^{N_Y} \frac{a_{ij}^2}{\alpha_{ij}} b_i e^{-l_i^{(n)}}. \tag{1.64}$$

In [69] and [14], the following choice for the α_{ij}'s was used, following [72]

$$\alpha_{ij} = \frac{a_{ij}\mu_j^{(n)}}{\sum_{k=1}^{N_p} a_{ik}\mu_k^{(n)}} = \frac{a_{ij}\mu_j^{(n)}}{l_i^{(n)}}. \tag{1.65}$$

Substituting into (1.64) etc. yields the *Convex-NR-1 algorithm*:

$$\boxed{\mu_j^{(n+1)} = \mu_j^{(n)} + \mu_j^{(n)} \frac{\sum_{i=1}^{N_Y} a_{ij} \left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i}\right) b_i e^{-l_i^{(n)}}}{\sum_{i=1}^{N_Y} a_{ij} l_i^{(n)} b_i e^{-l_i^{(n)}}}}. \tag{1.66}$$

The diagonal scaling matrix of this Convex-NR-1 algorithm is identical to (1.55), which is interesting since (1.55) was presented for the case $r_i = 0$. There are potential problems with the choice (1.65) when the μ_j's approach zero [69], so the following alternative choice, considered in [73] and [101], may be preferable:

$$\alpha_{ij} = \frac{a_{ij}}{\sum_{k=1}^{N_p} a_{ik}} = \frac{a_{ij}}{a_{i\cdot}}, \tag{1.67}$$

where $a_{i\cdot} \triangleq \sum_{j=1}^{N_p} a_{ij}$, for which (1.64) leads to

$$[D(\underline{\mu})]_{jj} = \frac{1}{\sum_{i=1}^{N_Y} a_{ij} a_{i\cdot} b_i e^{-l_i^{(n)}}}.$$

A small advantage of the choice (1.67) over (1.65) is that the $a_{i\cdot}$'s in the denominator of (1.67) are independent of $\underline{\mu}$ so they can be precomputed, unlike the denominator of (1.65).

Substituting the above into (1.50) yields the following *Convex-NR-2* algorithm:

$$\boxed{\mu_j^{(n+1)} = \mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij}\left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i}\right) b_i e^{-l_i^{(n)}}}{\sum_{i=1}^{N_Y} a_{ij} a_{i\cdot} b_i e^{-l_i^{(n)}}}}. \tag{1.68}$$

Each iteration requires one forward projection (to compute the $l_i^{(n)}$'s) and two back-projections (one each for the numerator and denominator). In general a line search would be necessary with this algorithm to ensure monotonicity and convergence.

1.4.5.2 Convex-PS algorithm

The function $Q_{2,j}$ in (1.62) cannot be maximized analytically, but we can apply the optimization transfer principle of Section 1.3.2 to derive the first intrinsically monotonic algorithm presented in this chapter. Using the surrogate parabola (1.27):

$$g_i^{(n)}(l) \geq q_i^{(n)}(l; l_i^{(n)}) \triangleq g_i^{(n)}(l_i^{(n)}) + \dot{g}_i^{(n)}(l_i^{(n)})(l - l_i^{(n)}) - \frac{c_i^{(n)}}{2}(l - l_i^{(n)})^2, \; \forall l \geq 0 \tag{1.69}$$

where from (1.29) the optimal curvature is

$$c_i^{(n)} = c(l_i^{(n)}; \bar{N}_i^{(n)}, b_i, 0) = \begin{cases} \left[2(b_i - l_i^{(n)} b_i e^{-l_i^{(n)}} - l_i^{(n)})/(l_i^{(n)})^2\right]_+, & l_i^{(n)} > 0 \\ b_i, & l_i^{(n)} = 0. \end{cases} \tag{1.70}$$

This suggests the following quadratic surrogate function

$$Q_{3,j}(\mu_j; \underline{\mu}^{(n)}) \triangleq \sum_{i=1}^{N_Y} q_i^{(n)}\left(\frac{a_{ij}}{\alpha_{ij}}(\mu_j - \mu_j^{(n)}) + l_i^{(n)}; l_i^{(n)}\right),$$

with corresponding algorithm

$$\mu_j^{(n+1)} = \arg\max_{\mu_j \geq 0} Q_{3,j}(\mu_j; \underline{\mu}^{(n)}). \tag{1.71}$$

Since $Q_{3,j}$ is quadratic, it is trivial to maximize analytically. Furthermore, since this is a 1D maximization problem for a concave function, to enforce the nonnegativity constraint we simply reset any negative pixels to zero. The derivatives of $Q_{3,j}$ are:

$$\begin{aligned} \frac{d}{d\mu_j} Q_{3,j}(\mu_j; \underline{\mu}^{(n)})\Big|_{\mu_j = \mu_j^{(n)}} &= \sum_{i=1}^{N_Y} a_{ij} \dot{q}_i^{(n)}(l_i^{(n)}; l_i^{(n)}) \\ &= \sum_{i=1}^{N_Y} a_{ij} \dot{g}_i^{(n)}(l_i^{(n)}; l_i^{(n)}) = \sum_{i=1}^{N_Y} a_{ij}(1 - \bar{N}_i^{(n)}/(b_i e^{-l_i^{(n)}}) b_i e^{-l_i^{(n)}} \\ &= \sum_{i=1}^{N_Y} a_{ij}(1 - y_i/(b_i e^{-l_i^{(n)}} + r_i) b_i e^{-l_i^{(n)}} = \frac{d}{d\mu_j} L(\underline{\mu})\Big|_{\underline{\mu} = \underline{\mu}^{(n)}}, \end{aligned}$$

$$\frac{d^2}{d\mu_j^2} Q_{3,j}(\mu_j; \underline{\mu}^{(n)}) \Big|_{\mu_j = \mu_j^{(n)}} = \sum_{i=1}^{N_Y} \frac{a_{ij}^2}{\alpha_{ij}} c_i^{(n)}.$$

Using the choice (1.67) for the α_{ij}'s yields the following *Convex-PS algorithm*:

$$\boxed{\mu_j^{(n+1)} = \left[\mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij} \left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i} \right) b_i e^{-l_i^{(n)}}}{\sum_{i=1}^{N_Y} a_{ij} a_{i\cdot} c_i^{(n)}} \right]_+}. \tag{1.72}$$

The $[\cdot]_+$ operation enforces the nonnegativity constraint. Since this is the first intrinsically monotonic algorithm presented in this chapter, we provide the following more detailed description of its implementation.

for $n = 0, 1, \ldots$ {

$$\begin{aligned} l_i^{(n)} &:= [\boldsymbol{A}\underline{\mu}^{(n)}]_i, \; i = 1, \ldots, N_Y \\ \dot{h}_i^{(n)} &:= \left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i} \right) b_i e^{-l_i^{(n)}}, \; i = 1, \ldots, N_Y \end{aligned} \tag{1.73}$$

compute $c_i^{(n)}$ using (1.70), $i = 1, \ldots, N_Y$
$e_i^{(n)} := a_{i\cdot} c_i^{(n)}$, $i = 1, \ldots, N_Y$
for $j = 1, \ldots, N_p$ {

$$\mu_j^{(n+1)} := \left[\mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij} \dot{h}_i^{(n)}}{\sum_{i=1}^{N_Y} a_{ij} e_i^{(n)}} \right]_+ \tag{1.74}$$

}

}.

This ML algorithm monotonically increases the log-likelihood function $L(\underline{\mu}^{(n)})$ each iteration.

To derive (1.72), we have used all three of the optimization transfer principles that are present in this chapter: the EM approach in (1.57), the convex separability approach in (1.61), and the parabola surrogate approach in (1.69). Undoubtably there are other algorithms awaiting discovery by using different combinations of these principles!

1.4.6 Ordered-subsets EM algorithm

Hudson and Larkin [102] proposed an "ordered subsets" modification of the emission EM algorithm in which one updates the image estimate using a sequence

of subsets of the measured data, subsampled by projection angle, rather than using all the measurements simultaneously. Manglos et al. [44] applied this concept to the transmission EM algorithm (1.45), yielding the iteration:

$$\boxed{\mu_j^{(n+1)} \triangleq \frac{\sum_{i \in \mathcal{S}_n} (\bar{M}_{ij}^{(n)} - \bar{N}_{ij}^{(n)})}{\frac{1}{2} \sum_{i \in \mathcal{S}_n} a_{ij} (\bar{M}_{ij}^{(n)} + \bar{N}_{ij}^{(n)})}}, \tag{1.75}$$

where $\mathcal{S}_n$ is a subset of the ray indices $\{1, \ldots, N_Y\}$ selected for the nth subiteration. This type of modification destroys the monotonicity properties of the EM algorithm, and typically the sequence of images asymptotically approaches a limit cycle [103–105]. However, at least in the emission case, the OSEM algorithm seems to produce visually appealing images fairly quickly and hence has become very popular.

Any of the algorithms described in this chapter could be easily modified to have a block-iterative form akin to (1.75) simply by replacing any ray summations (those over i) with partial summations over $i \in \mathcal{S}_n$. Since (1.75) requires the same number of exponentiations per iteration as the transmission EM algorithm (1.45), it is still impractical. However, block-iterative forms of some of the other algorithms described in this chapter are practical. In particular, Nuyts et al. proposed a block-iterative modification of a gradient-based method [106] (only in the case $r_i = 0$). Kamphius and Beekman [107] proposed a block-iterative version of (1.66) (only in the case $r_i = 0$). Erdoğan and Fessler propose a block-iterative version of the separable paraboloidal surrogates algorithm of Section 1.6.2 in [108, 109].

1.4.7 EM algorithms with nonseparable penalty functions

All the algorithms described above were given for the ML case (where $\beta = 0$). What happens if we want to include a nonseparable penalty function for regularization, for example in the Convex-PS algorithm? Considering (1.33), it appears that we should replace (1.71) with

$$\underline{\mu}^{(n+1)} = \arg \max_{\underline{\mu} \geq \underline{0}} Q_3(\underline{\mu}; \underline{\mu}^{(n)}) - \beta R(\underline{\mu}) \,. \tag{1.76}$$

Unfortunately, this is a nontrivial maximization since a nonseparable penalty $R(\underline{\mu})$ leads to coupled equations. One approach to circumventing this problem is the *generalized EM* (GEM) method [80, 110–112], in which one replaces the maximization in (1.76) with a few cycles of, for example, the coordinate ascent algorithm.

A clever alternative is to replace $R(\underline{\mu})$ in (1.76) with a separable surrogate function using a similar trick as in (1.60), which was proposed by De Pierro [73]. We discuss this approach in more detail in Section 1.6.2: see (1.92).

1.5 Coordinate-ascent algorithms

A simple and natural approach to finding the maximizer of $\Phi(\underline{\mu})$ is to sequentially maximize $\Phi(\underline{\mu})$ over each element μ_j of $\underline{\mu}$ using *the most recent values* for all

other elements of $\underline{\mu}$. A general coordinate ascent method has the following form:

$$\begin{aligned}
&\text{for } n = 0, 1, \ldots \{ \\
&\quad \texttt{for } j = 1, \ldots, N_p \{ \\
&\quad \mu_j^{(n+1)} := \arg\max_{\mu_j \geq 0} \Phi(\mu_1^{(n+1)}, \ldots, \mu_{j-1}^{(n+1)}, \mu_j, \mu_{j+1}^{(n)}, \ldots, \mu_{N_p}^{(n)}) \qquad (1.77) \\
&\quad \} \\
&\}
\end{aligned}$$

The operation in (1.77) is performed "in place," i.e., the new value of μ_j replaces the old value, so that the most recent values of all elements of $\underline{\mu}$ are always used. An early use of such a method for tomography was in [113].

Sauer and Bouman analyzed such algorithms using clever frequency domain arguments [13], and showed that sequential algorithms yield iterates whose high frequency components converge fastest. This is often ideal for tomography, since we can use a low-resolution FBP image as the initial guess, and then iterate to improve resolution and reduce noise, which is mostly high frequency errors. (Using a uniform or zero initial image for coordinate ascent is a very poor choice since low frequencies can converge very slowly.)

The long string of arguments in (1.77) is quite notationally cumbersome. For the remainder of this section, we use

$$\tilde{\underline{\mu}} \triangleq [\mu_1^{(n+1)} \ \cdots \ \mu_{j-1}^{(n+1)} \ \mu_j^{(n)} \ \mu_{j+1}^{(n)} \ \cdots \ \mu_{N_p}^{(n)}] \qquad (1.78)$$

as shorthand for the vector of the most recent parameter values. For simplicity, this notation leaves implicit the dependence of $\tilde{\underline{\mu}}$ on iteration n and pixel index j.

The general method described by (1.77) is not exactly an "algorithm," since the procedure for performing the 1D maximization is yet unspecified. In practice it is impractical to find the exact maximizer, even in the 1D problem (1.77), so we settle for methods that increase Φ.

1.5.1 Coordinate-ascent Newton-Raphson

If $\Phi(\underline{\mu})$ were a quadratic functional, then the natural approach to performing the maximization in (1.77) would be Newton's method. Since $\Phi(\underline{\mu})$ in (1.15) is nonquadratic, applying Newton's method to (1.77) does not guarantee monotonic increases in Φ, but one might still try it anyway and hope for the best. In practice monotonicity usually does not seem to be a problem, as suggested by the success of Bouman et al. with this approach [18, 114]. For such a *coordinate-ascent Newton-*

Raphson (CA-NR) algorithm, we replace (1.77) with the following update:

$$\mu_j^{(n+1)} = \left[\mu_j^{(n)} + \frac{\frac{\partial}{\partial \mu_j} \Phi(\underline{\mu}) \Big|_{\underline{\mu}=\tilde{\underline{\mu}}}}{-\frac{\partial^2}{\partial \mu_j^2} \Phi(\underline{\mu}) \Big|_{\underline{\mu}=\tilde{\underline{\mu}}}} \right]_+ , \tag{1.79}$$

The $[\cdot]_+$ operation enforces the nonnegativity constraint. The first partial derivative is given by (1.21), and the second is given by:

$$\frac{\partial^2}{\partial \mu_j^2} \Phi(\underline{\mu}) = \sum_{i=1}^{N_Y} a_{ij}^2 \ddot{h}_i([\boldsymbol{A}\underline{\mu}]_i) - \beta \frac{\partial^2}{\partial \mu_j^2} R(\underline{\mu}) , \tag{1.80}$$

where $\ddot{h}_i$ is given by (1.14).

Specifically, using (1.13), (1.14), (1.21), and (1.80), the update (1.79) of the CA-NR algorithm becomes

$$\mu_j^{(n+1)} = \left[\mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij} \left(1 - \frac{y_i}{b_i e^{-[\boldsymbol{A}\tilde{\underline{\mu}}]_i} + r_i} \right) b_i e^{-[\boldsymbol{A}\tilde{\underline{\mu}}]_i} - \beta \frac{\partial}{\partial \mu_j} R(\tilde{\underline{\mu}})}{\sum_{i=1}^{N_Y} a_{ij}^2 \left(1 - \frac{y_i r_i}{(b_i e^{-[\boldsymbol{A}\tilde{\underline{\mu}}]_i} + r_i)^2} \right) b_i e^{-[\boldsymbol{A}\tilde{\underline{\mu}}]_i} + \beta \frac{\partial^2}{\partial \mu_j^2} R(\tilde{\underline{\mu}})} \right]_+ . \tag{1.81}$$

Literally interpreted, this form of the CA-NR algorithm appears to be extremely inefficient computationally, because it appears to require that $\boldsymbol{A}\tilde{\underline{\mu}}$ be recomputed after *every* pixel is updated sequentially. This would lead to $O(N_p^2)$ flops per iteration, which is impractical.

In the following *efficient* implementation of CA-NR, we maintain a copy of $\tilde{\underline{l}} \triangleq \boldsymbol{A}\tilde{\underline{\mu}}$ as a "state vector," and update that vector after each pixel is updated.

Initialization: $\tilde{l} := \boldsymbol{A}\underline{\mu}^0$
for $n = 0, 1, \ldots$ {
  for $j = 1, \ldots, N_p$ {

$$\mu_j^{(n+1)} := \left[\mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij} \left(1 - \frac{y_i}{b_i e^{-\tilde{l}_i} + r_i} \right) b_i e^{-\tilde{l}_i} - \beta \frac{\partial}{\partial \mu_j} R(\underline{\tilde{\mu}})}{\sum_{i=1}^{N_Y} a_{ij}^2 \left(1 - \frac{y_i r_i}{(b_i e^{-\tilde{l}_i} + r_i)^2} \right) b_i e^{-\tilde{l}_i} + \beta \frac{\partial^2}{\partial \mu_j^2} R(\underline{\tilde{\mu}})} \right]_+ \tag{1.82}$$

$$\tilde{l}_i := \tilde{l}_i + a_{ij}(\mu_j^{(n+1)} - \mu_j^{(n)}), \; \forall i \; : \; a_{ij} \neq 0 \tag{1.83}$$

  }
}

The computational requirements per iteration are summarized in [77].

Bouman et al. also present a clever method to search for a zero-crossing to avoid using Newton-Raphson for the penalty part of the objective function [19].

The numerator in (1.82) is essentially a backprojection, and appears to be quite similar to the backprojection in the numerator of (1.68). One might guess then that coordinate ascent and an algorithm like Convex-NR in (1.68) would have similar computational requirements, but they do not. We can precompute the entire expression $\left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i} \right) b_i e^{-l_i^{(n)}}$ for each i in the numerator of (1.68) *before* starting the backprojection, which saves many flops and nonsequential memory accesses. In contrast, the numerator of (1.82) contains $\tilde{l}_i$'s that *change after each pixel is updated*, so that expression cannot be precomputed. During the "backprojection" step, one must access four arrays (nonsequentially): the y_i's, b_i's, r_i's, and $\tilde{l}_i$'s, in addition to the system matrix elements a_{ij}. And one must compute an exponentiation and a handful of addition and multiplications for *each nonzero* a_{ij}. For these reasons, coordinate ascent is quite expensive computationally per iteration. On the other hand, experience shows that if one considers the number of *iterations* required for "convergence," then CA-NR is among the best of all algorithms. The PSCA algorithm described in Section 1.6.4 below is an attempt to capture the convergence rate properties of CA-NR, but yet guaranteeing monotonicity and greatly reducing the flop counts per iteration.

An alternative approach to ensuring monotonicity would be to evaluate the objective function Φ after updating each pixel, and impose an interval search in the (hopefully relatively rare) cases where the objective function decreases. Unfortunately, evaluating Φ after every pixel adds considerable computational overhead.

1.5.2 Variation 1: Hybrid Poisson/polynomial approach

One approach to reducing the flops required by (1.82) is to replace some of the nonquadratic h_i's in the log-likelihood (1.11) with quadratic functions. Specifically, any given measured sinogram is likely to contain a mixture of high and low count rays. For high-count rays, a quadratic approximation to h_i should be adequate, e.g. a Gaussian approximation to the Poisson statistics. For low count rays, the Poisson h_i function (1.12) can be retained to avoid biases. This "hybrid Poisson/polynomial" approach was proposed in [14], and was shown to significantly reduce CPU time. However, implementation is somewhat inelegant since the system matrix $\boldsymbol{A}$ must be stored by sparse columns, and those sparse columns must be regrouped according to the indices of low and high count rays, which is a programming nuisance.

1.5.3 Variation 2: 1D parabolic surrogates

Besides CPU time, another potential problem with (1.82) is that it is not guaranteed to monotonically increase Φ, so divergence is possible. One can ensure monotonicity by applying the optimization transfer principle to the maximization problem (1.77). One possible approach is to use a parabolic surrogate for the 1D function $f(\mu_j) = \Phi(\mu_1^{(n+1)}, \ldots, \mu_{j-1}^{(n+1)}, \mu_j, \mu_{j+1}^{(n)}, \ldots, \mu_{N_p}^{(n)})$. For the fastest convergence rate, the optimal parabolic surrogate would have the lowest possible curvature, as discussed in Section 1.6.2 below. The surrogate parabola (1.27) with optimal curvature (1.29) can be applied to (1.77) to yield an algorithm of the form (1.82) but with a different expression in the denominator. Ignoring the penalty function, the ML coordinate ascent parabola surrogate (CA-PS) algorithm is

$$\boxed{\mu_j^{(n+1)} := \left[\mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij} \left(1 - \frac{y_i}{b_i e^{-\tilde{l}_i} + r_i} \right) b_i e^{-\tilde{l}_i}}{\sum_{i=1}^{N_Y} a_{ij}^2 \, c(\tilde{l}_i, y_i, b_i, r_i)} \right]_+ , \qquad (1.84)}$$

where $c(\cdot)$ was defined in (1.29). Unfortunately, this algorithm suffers from the same high CPU demands as (1.82), so is impractical. To incorporate a penalty function, one could follow a similar procedure as in Section 1.6.2 below.

Another approach to applying optimization transfer to (1.77) was proposed by Saquib et al. [115] and Zheng et al. [116], called "functional substitution." That method also yields a monotonic algorithm, for the case $r_i = 0$ since concavity of h_i is exploited in the derivation. The required flops are comparable to those of CA-NR. We can generalize the functional substitution algorithm of [116], to the case $r_i \neq 0$ by exploiting the EM surrogate described in Section 1.4.5 to derive a new monotic algorithm. Essentially one simply replaces y_i with $y_i b_i e^{-\tilde{l}_i} / (b_i e^{-\tilde{l}_i} + r_i)$ in the curvature terms in [116], yielding an algorithm that is identical to (1.82) but with a different denominator.

1.6 Paraboloidal surrogates algorithms

Coordinate ascent algorithms are *sequential update* algorithms because the pixels are updated in sequence. This leads to fast convergence, but requires column access of the system matrix $\boldsymbol{A}$, and makes parallelization quite difficult. In contrast, *simultaneous update* algorithms can update all pixels independently in parallel, such as the EM algorithms (1.45), (1.46), and (1.49), the scaled gradient ascent algorithms (1.50), (1.54), and (1.56), and the Convex algorithms (1.66), (1.68), and (1.72). However, a serious problem with *all* the simultaneous algorithms described above, except Convex-PS (1.72), is that they are not intrinsically monotonic. (They can all be forced to be monotonic by adding line searches, but this is somewhat inconvenient.) In this section we describe an approach based on the optimization transfer principle of Section 1.3.2 that leads to a simultaneous update algorithm that is also intrinsically monotonic, as well as a sequential algorithm that is intrinsically monotonic like CA-PS (1.84), but much more computationally efficient.

As mentioned in Section 1.3.4, a principal difficulty with maximizing (1.15) is the fact that the h_i's in (1.12) are nonquadratic. Maximization is much easier for quadratic functions, so it is natural to use the surrogate parabola described in (1.27) to construct a *paraboloidal surrogate function* for the log-likelihood $L(\underline{\mu})$ in (1.11).

Using (1.29), define

$$c_i^{(n)} = c(l_i^{(n)}, y_i, b_i, r_i)\,,$$

where $l_i^{(n)} = [\boldsymbol{A}\underline{\mu}^{(n)}]_i$ was defined in (1.28). For this choice of curvatures, the parabola

$$q_i(l; l_i^{(n)}) = h_i(l_i^{(n)}) + \dot{h}_i(l_i^{(n)})(l - l_i^{(n)}) - \frac{1}{2}c_i^{(n)}(l - l_i^{(n)})^2$$

is a surrogate for $h_i(\cdot)$ in the sense that $h_i(l) \geq q_i(l; l_i^{(n)})$ for all $l \geq 0$. Summing these 1D surrogate functions, as in (1.11), leads to the following surrogate function for the log-likelihood:

$$Q_1(\underline{\mu}; \underline{\mu}^{(n)}) \stackrel{\triangle}{=} \sum_{i=1}^{N_Y} q_i([\boldsymbol{A}\underline{\mu}]_i; [\boldsymbol{A}\underline{\mu}^{(n)}]_i)\,. \tag{1.85}$$

This is a surrogate for the log-likelihood in the sense that if we define

$$\phi_1(\underline{\mu}; \underline{\mu}^{(n)}) = Q_1(\underline{\mu}; \underline{\mu}^{(n)}) - \beta R(\underline{\mu})\,, \tag{1.86}$$

then ϕ_1 satisfies (1.24), (1.25), and (1.26).

When expanded, the paraboloidal surrogate function Q_1 in (1.85) has the following quadratic form:

$$Q_1(\underline{\mu}; \underline{\mu}^{(n)}) \equiv \nabla L(\underline{\mu}^{(n)})(\underline{\mu} - \underline{\mu}^{(n)}) - \frac{1}{2}(\underline{\mu} - \underline{\mu}^{(n)})'\boldsymbol{A}'\,\mathrm{diag}\left\{c_i^{(n)}\right\}\boldsymbol{A}(\underline{\mu} - \underline{\mu}^{(n)})\,. \tag{1.87}$$

Maximizing a quadratic form like (1.87) is potentially much easier than the log-likelihood (1.11).

1.6.1 Paraboloidal surrogate with Newton Raphson

For a quadratic penalty $R(\underline{\mu}) = \underline{\mu}' \boldsymbol{R} \underline{\mu}/2$, as in (1.18), the surrogate function ϕ_1 in (1.86) above is a quadratic form. Disregarding the nonnegativity constraint, in principle we can maximize ϕ_1 (as in (1.22)) by zeroing the gradient of ϕ_1:

$$\nabla' \phi_1(\underline{\mu}; \underline{\mu}^{(n)}) = \nabla' L(\underline{\mu}^{(n)}) - \beta \boldsymbol{R} \underline{\mu}^{(n)} - [\boldsymbol{A}' \operatorname{diag}\left\{c_i^{(n)}\right\} \boldsymbol{A} + \beta \boldsymbol{R}]^{-1} (\underline{\mu} - \underline{\mu}^{(n)}) .$$

This leads to the following paraboloidal surrogates Newton-Raphson (PS-NR) algorithm:

$$\underline{\mu}^{(n+1)} = \underline{\mu}^{(n)} + \left[-\nabla^2 \phi_1(\underline{\mu}; \underline{\mu}^{(n)}) \Big|_{\underline{\mu}=\underline{\mu}^{(n)}} \right]^{-1} \nabla'_{\underline{\mu}} \phi_1(\underline{\mu}; \underline{\mu}^{(n)}) \Big|_{\underline{\mu}=\underline{\mu}^{(n)}} ,$$

$$\boxed{\underline{\mu}^{(n+1)} = \underline{\mu}^{(n)} + [\boldsymbol{A}' \operatorname{diag}\left\{c_i^{(n)}\right\} \boldsymbol{A} + \beta \boldsymbol{R}]^{-1} [\nabla' L(\underline{\mu}^{(n)}) - \beta \boldsymbol{R} \underline{\mu}^{(n)}] .} \quad (1.88)$$

There are three problems with this algorithm. The matrix inverse is impractical, the method appears only to apply to quadratic penalty functions, and nonnegativity is not enforced. Fortunately all three of these limitations can be overcome, as we describe next.

1.6.2 Separable paraboloidal surrogates algorithm

In this section we derive an intrinsically monotonic algorithm that we believe to be the current "method of choice" for cases where one desired a *simultaneous* update without any line searches.

A difficulty in maximizing (1.86) is that in general both Q_1 and R are *nonseparable* functions of the elements of the parameter vector $\underline{\mu}$. However, since each q_i in (1.85) is concave, we can apply precisely the same convexity trick of De Pierro used in (1.59) and (1.60) to form a second *separable* surrogate function. Since

$$\begin{aligned} q_i([\boldsymbol{A}\underline{\mu}]_i; l_i^{(n)}) &= q_i \left(\sum_{j=1}^{N_p} \alpha_{ij} \left[\frac{a_{ij}}{\alpha_{ij}} (\mu_j - \mu_j^{(n)}) + l_i^{(n)} \right] ; l_i^{(n)} \right) \\ &\geq \sum_{j=1}^{N_p} \alpha_{ij} q_i \left(\frac{a_{ij}}{\alpha_{ij}} (\mu_j - \mu_j^{(n)}) + l_i^{(n)}; l_i^{(n)} \right) , \end{aligned}$$

the natural separable surrogate function for the log-likelihood is

$$L(\underline{\mu}) \geq Q_2(\underline{\mu}; \underline{\mu}^{(n)}) \triangleq \sum_{j=1}^{N_p} Q_{2,j}(\mu_j; \underline{\mu}^{(n)}) \quad (1.89)$$

where

$$Q_{2,j}(\mu_j; \underline{\mu}^{(n)}) \triangleq \sum_{i=1}^{N_Y} \alpha_{ij} q_i \left(\frac{a_{ij}}{\alpha_{ij}} (\mu_j - \mu_j^{(n)}) + l_i^{(n)}; l_i^{(n)} \right) . \tag{1.90}$$

Since Q_2 is a *separable function*, it is easily maximized.

We need to apply a similar trick to separate the penalty function. We assume that R has the form (1.18). Similar to (1.59) we have

$$[\boldsymbol{C}\underline{\mu}]_k = \sum_{j=1}^{N_p} \gamma_{kj} \left[\frac{c_{kj}}{\gamma_{kj}} (\mu_j - \mu_j^{(n)}) + [\boldsymbol{C}\underline{\mu}^{(n)}]_k \right] , \tag{1.91}$$

where $\gamma_{kj} \geq 0$ and $\sum_{j=1}^{N_p} \gamma_{kj} = 1$. So since $t^2/2$ is a convex function:

$$\begin{aligned} R(\underline{\mu}) &= \sum_{k=1}^{K} \omega_k \frac{1}{2} \left([\boldsymbol{C}\underline{\mu}]_k \right)^2 \\ &= \sum_{k=1}^{K} \omega_k \frac{1}{2} \left(\sum_{j=1}^{N_p} \gamma_{kj} \left[\frac{c_{kj}}{\gamma_{kj}} (\mu_j - \mu_j^{(n)}) + [\boldsymbol{C}\underline{\mu}^{(n)}]_k \right] \right)^2 \\ &\geq \sum_{k=1}^{K} \sum_{j=1}^{N_p} \gamma_{kj} \omega_k \frac{1}{2} \left(\frac{c_{kj}}{\gamma_{kj}} (\mu_j - \mu_j^{(n)}) + [\boldsymbol{C}\underline{\mu}^{(n)}]_k \right)^2 , \end{aligned} \tag{1.92}$$

so the natural surrogate function is

$$R(\underline{\mu}; \underline{\mu}^{(n)}) \triangleq \sum_{j=1}^{N_p} R_j(\mu_j; \underline{\mu}^{(n)}) ,$$

where

$$R_j(\mu_j; \underline{\mu}^{(n)}) \triangleq \sum_{k=1}^{K} \gamma_{kj} \omega_k \frac{1}{2} \left(\frac{c_{kj}}{\gamma_{kj}} (\mu_j - \mu_j^{(n)}) + [\boldsymbol{C}\underline{\mu}^{(n)}]_k \right)^2 .$$

Combining Q_2 and $R(\cdot; \underline{\mu}^{(n)})$ yields the following separable quadratic surrogate function:

$$\phi_2(\underline{\mu}; \underline{\mu}^{(n)}) \triangleq Q_2(\underline{\mu}; \underline{\mu}^{(n)}) - \beta R(\underline{\mu}; \underline{\mu}^{(n)}) = \sum_{j=1}^{N_p} \phi_{2,j}(\mu_j; \underline{\mu}^{(n)})$$

where

$$\phi_{2,j}(\mu_j; \underline{\mu}^{(n)}) \triangleq Q_{2,j}(\mu_j; \underline{\mu}^{(n)}) - \beta R_j(\mu_j; \underline{\mu}^{(n)}) .$$

Using the choice (1.67) for the α_{ij}'s, the surrogate curvature is

$$\frac{d^2}{d\mu_j^2} Q_{2,j}(\mu_j; \underline{\mu}^{(n)}) = \sum_{i=1}^{N_Y} a_{ij}^2 / \alpha_{ij} c_i^{(n)} = \sum_{i=1}^{N_Y} a_{ij} a_{i\cdot} c_i^{(n)} .$$

Similarly, if we choose $\gamma_{kj} = |c_{kj}|/c_{k\cdot}$, where $c_{k\cdot} \triangleq \sum_{j=1}^{N_p} |c_{kj}|$, then

$$r_j \triangleq \frac{d^2}{d\mu_j^2} R_j(\mu_j; \underline{\mu}^{(n)}) = \sum_{k=1}^{K} c_{kj}^2 \omega_k / \gamma_{kj} = \sum_{k=1}^{K} |c_{kj}| c_{k\cdot} \omega_k . \tag{1.93}$$

Since the surrogate is separable, the optimization transfer algorithm (1.22) becomes

$$\mu_j^{(n+1)} = \arg\max_{\mu_j \geq 0} \phi_{2,j}(\mu_j; \underline{\mu}^{(n)}), \; j = 1, \ldots, N_p .$$

Since $\phi_{2,j}(\mu_j; \underline{\mu}^{(n)})$ is quadratic, it is easily maximized by zeroing its derivative, leading to the following *separable paraboloidal surrogates* (SPS) algorithm:

$$\boxed{\mu_j^{(n+1)} = \left[\mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij} \dot{h}_i^{(n)} - \beta \sum_{k=1}^{K} c_{kj} \omega_k [\boldsymbol{C} \underline{\mu}^{(n)}]_k}{\sum_{i=1}^{N_Y} a_{ij} a_{i\cdot} c_i^{(n)} + \beta r_j} \right]_+ ,} \tag{1.94}$$

where $\dot{h}_i^{(n)}$ was defined in (1.73) and we precompute the r_j's in (1.93) before iterating. This algorithm is highly parallelizable, and can be implemented efficiently using the same structure as the Convex-PS algorithm (1.74). It is also easy to form an ordered subsets version (cf. Section 1.4.6) of the SPS algorithm [109].

See [109, 117] for the extension to nonquadratic penalty functions, which is based on a parabola surrogate for ψ_k proposed by Huber [118].

1.6.3 Ordered subsets revisited

One can easily form an ordered subsets version (cf. Section 1.4.6) of the SPS algorithm (1.94) by replacing the sums over i with sums over subsets of the rays, yielding the ordered subsets transmission (OSTR) algorithm described in [109]. Since ordered subsets algorithms are not guaranteed to converge, one may as well further abandon monotonicity and replace the denominator in the ordered subsets version of (1.94) with something that can be precomputed. Specifically, in [109] we recommend replacing the $c_i^{(n)}$'s in (1.94) with[11]

$$c_i = -\ddot{h}_i(\hat{l}_i) = \begin{cases} \frac{(y_i - r_i)^2}{y_i}, & y_i > 0 \\ 0, & y_i = 0 , \end{cases} \tag{1.95}$$

[11] This trick is somewhat similar in spirit to the method of Fisher scoring [119, 120], in which one replaces the Hessian with its expectation (the Fisher information matrix) to reduce computation in nonquadratic optimization problems.

where $\hat{l}_i \triangleq \log \frac{b_i}{y_i - r_i}$. For this "fast denominator" approximation, the OSTR-M algorithm becomes:

$$\boxed{\mu_j^{\text{new}} = \left[\mu_j + \frac{M \sum_{i \in \mathcal{S}} a_{ij} \dot{h}_i^{(n)} - \beta \sum_{k=1}^{K} c_{kj} \omega_k [C \underline{\mu}^{(n)}]_k}{d_j + w_j} \right]_+ ,} \tag{1.96}$$

where $\mathcal{S}$ is a cyclically chosen subset of the rays, formed by angular subsampling by a factor M, where $\dot{h}_i^{(n)}$ was defined in (1.73), and where we precompute

$$d_j \triangleq \sum_{i=1}^{N_Y} a_{ij} a_{i\cdot} c_i, \; j = 1, \ldots, N_p .$$

The results in [109] show that this algorithm does not quite find the maximizer of the objective function Φ, but the images are nearly as good as those produced by convergent algorithms in terms of mean squared error and segmentation accuracy.

1.6.4 Paraboloidal surrogates coordinate-ascent (PSCA) algorithm

A disadvantage of simultaneous updates like (1.94) is that they typically converge slowly since separable surrogate functions have high curvature and hence slow convergence rates (cf. Section 1.3.3). Thus, in [77, 121] we proposed to apply coordinate ascent to the quadratic surrogate function (1.86). (We focus on the quadratic penalty case here; the extension to the nonquadratic case is straightforward following a similar approach as in Section 1.6.2.) To apply CA to (1.86), we sequentially maximize $\phi_1(\underline{\mu}; \underline{\mu}^{(n)})$ over each element μ_j, using the most recent values for all other elements of $\underline{\mu}$, as in Section 1.5. We again adopt the shorthand (1.78) here. In its simplest form, this leads to a *paraboloidal surrogates coordinate ascent* (PSCA) algorithm having a similar form as (1.82), but with the inner update being:

$$\boxed{\mu_j^{(n+1)} := \left[\mu_j^{(n)} + \frac{\sum_{i=1}^{N_Y} a_{ij} \dot{q}_i^{(n)} - \beta \frac{\partial}{\partial \mu_j} R(\underline{\tilde{\mu}})}{d_j + \beta \frac{\partial^2}{\partial \mu_j^2} R(\underline{\tilde{\mu}})} \right]_+ ,} \tag{1.97}$$

where, before looping over j in each iteration, we precompute

$$d_j = \sum_{i=1}^{N_Y} a_{ij}^2 c_i^{(n)}, \; j = 1, \ldots, N_p , \tag{1.98}$$

and the following term is maintained as a "state vector" (analogous to the $\tilde{l}_i$'s in (1.83):

$$\dot{q}_i^{(n)} \triangleq \dot{h}_i(l_i^{(n)}) = \left(1 - \frac{y_i}{b_i e^{-l_i^{(n)}} + r_i} \right) b_i e^{-l_i^{(n)}} .$$

This precomputation saves many flops per iteration, yet still yields an intrinsically monotonic algorithm. Even greater computational savings are possibly by a "fast denominator" trick similar to (1.95), although one should then check for monotonicity after each iteration and redo the iteration using the monotonicity preserving denominators (1.98) in those rare cases where the objective function decreases. There are several "details" that are essential for efficient implementation; see [77].

1.6.5 Grouped coordinate ascent algorithm

We have described algorithms that update a single pixel at a time, as in the PSCA algorithm (1.97) above, or update all pixels simultaneously, as in the SPS algorithm (1.94) above. A problem with the sequential algorithms is that they are difficult to parallelize, whereas a problem with simultaneous algorithms is their slow convergence rates. An alternative is to update a *group* of pixels simultaneously. If the pixels are well separated spatially, then they may be approximately uncorrelated[12], which leads to separable surrogate functions that have lower curvature [101, 122, 123]. We call such methods *grouped coordinate ascent* (GCA) algorithms. The statistics literature has work on GCA algorithms e.g. [124], which in turn cites related algorithms dating to 1964! In tomography, one can apply the GCA idea directly to the log-likelihood (1.11) [101,117,123], or to the paraboloidal surrogate (1.85) [77].

1.7 Direct algorithms

The algorithms described above have all been developed, to some degree, by considering the specific form of the log-likelihood (1.11). It is reasonable to hypothesize that algorithms that are "tailor made" for the form of the objective function (1.15) in tomography should outperform (converge faster) general purpose optimization methods that usually treat the objective function as a "black box" in the interest of greatest generality. Nevertheless, general purpose optimization is a very active research area, and it behooves developers of image reconstruction algorithms to keep abreast of progress in that field. General purpose algorithms that are natural candidates for image reconstruction include the conjugate gradient algorithm and the quasi-Newton algorithm, described next.

1.7.1 Conjugate gradient algorithm

For unconstrained quadratic optimization problems, the preconditioned conjugate gradient (CG) algorithm [125] is particularly appealing because it converges rapidly[13] for suitably chosen preconditioners, e.g. [67]. For nonquadratic objective

[12]To be more precise: the submatrix of the Hessian matrix of Φ corresponding to a subset of spatially separated pixels is approximately diagonal.

[13]It is often noted that CG converges in N_p iterations in exact arithmetic, but this fact is essentially irrelevant in tomography since N_p is so large. More relevant is the fact that the convergence rate of CG is quite good with suitable preconditioners.

functions, or when constraints such as nonnegativity are desired, the CG method is somewhat less convenient due to the need to perform line searches. It may be possible to adopt the optimization transfer principles to simplify the line searches, cf. [67].

Mumcuoğlu et al. [31, 126] have been particularly successful in applying diagonally preconditioned conjugate gradients to both transmission and emission tomography. Their diagonal preconditioner was based on (1.52). They investigated both a penalty function approach to encourage nonnegativity [31, 126], as well as active set methods [127] for determining the set of nonzero pixels [128, 129].

An alternative approach to enforcing nonnegativity in gradient-based methods uses adaptive barriers [130].

1.7.2 Quasi-Newton algorithm

The ideal preconditioner for the conjugate gradient algorithm would be the inverse of the Hessian matrix, which would lead to superlinear convergence [131]. Unfortunately, in tomography the Hessian matrix is a large non-sparse matrix, so its inverse is impractical to compute and store. The basic idea of the quasi-Newton family of algorithms is to form low-rank approximations to the inverse of the Hessian matrix as the iterations proceed [85]. This approach has been applied by Kaplan et al. [132] to simultaneous estimation of SPECT attenuation and emission distributions, using the public domain software for limited memory, bound-constrained minimization (L-BFGS-B) [133]. Preconditioning has been found to accelerate such algorithms [132].

1.8 Alternatives to Poisson models

Some of the algorithms described above are fairly complex, and this complexity derives from the nonconvex, nonquadratic form of the transmission Poisson log-likelihood (1.11) and (1.12). It is natural then to ask whether there are simpler approaches that would give adequate results in practice. Every simpler approach that we are aware of begins by using the logarithmic transformation (1.3), which compensates for the nonlinearity of Beer's law (1.2) and leads then to a linear problem

$$\hat{l}_i \approx [\boldsymbol{A}\underline{\mu}]_i, \; i = 1, \ldots, N_Y \,. \tag{1.99}$$

Unfortunately, for low-count transmission scans, especially those contaminated by background events of any type ($r_i \neq 0$), the logarithm (1.3) simply cannot be used since $Y_i - r_i$ can be nonpositive for many rays. In medium to high-count transmission scans, the bias described in (1.4) should be small, so one could work with the estimated line integrals (the $\hat{l}_i$'s) rather than the raw transmission measurements (the Y_i's).

1.8.1 *Algebraic reconstruction methods*

A simple approach to estimating $\underline{\mu}$ is to treat (1.99) as a set of N_Y equations in N_p unknowns and try to "solve" for $\underline{\mu}$. This was the motivation for the algebraic reconstruction technique (ART) family of algorithms [11]. For noisy measurements the equations (1.99) are usually inconsistent, and ART "converges" to a limit cycle for inconsistent problems. One can force ART to converge by introducing appropriate strong underrelaxation [134]. However, the limit is the minimum-norm weighted least-squares solution for a particular norm that is unrelated to the measurement statistics. The Gauss-Markov theorem [90] states that estimator variance is minimized when the least-squares norm is chosen to be the inverse of the covariance matrix, so it seems preferable to approach (1.99) by first finding a statistically-motivated cost function, and then finding algorithms that minimize that cost function, rather than trying to "fix up" algorithms that were derived under the unrealistic assumption that (1.99) is a consistent system of equations.

1.8.2 *Methods to avoid*

A surprising number of investigators have applied the *emission* EM algorithm to "solve" (1.99), even though the statistics of $\hat{l}_i$ are *entirely different* from those of emission sinogram measurements, e.g. [15]. We strongly recommend avoiding this practice. Empirical results with simulated and phantom data show that this approach is inferior to methods such as OSTR which are based on the transmission statistical model (1.7).

Liang and Ye [135] present the following iteration for MAP reconstruction of attenuation maps without giving any derivation:

$$\mu_j^{(n+1)} = \mu_j \frac{\sum_{i=1}^{N_Y} a_{ij}}{\sum_{i=1}^{N_Y} a_{ij} [\boldsymbol{A}\underline{\mu}^{(n)}]_i / \log(b_i/y_i)} .$$

The iteration looks like an "upside down" emission EM algorithm. The convergence properties of this algorithm are unknown.

Zeng and Gullberg [136] proposed the following steepest ascent method with a fixed step-length parameter:

$$\mu_j^{(n+1)} = \mu_j^{(n)} + \alpha \left[\sum_{i=1}^{N_Y} a_{ij} (y_i/b_i - e^{-l_i^{(n)}}) + \beta \frac{\partial}{\partial \mu_j} R(\underline{\mu}^{(n)}) \right] ,$$

for an interesting choice of penalty $R(\underline{\mu})$ that encourages attenuation values near those of air/lung, soft tissue, or bone. Without a line search of the type studied by Lange [95, 97], monotonicity is not guaranteed. Even with a line search it is unlikely that this algorithm maximizes the objective function since its fixed points are not stationary points of Φ.

1.8.3 Weighted least-squares methods

Rather than simply treating (1.99) as a system of equations, we can use (1.99) as the rationale for a weighted least-squares cost function. There are several choices for the weights.

1.8.3.1 Model-weighted LS

By a standard propagation-of-errors argument, one can show from (1.3) and (1.7) that

$$\mathrm{Var}\left\{\hat{l}_i\right\} \approx \frac{\bar{y}_i}{(\bar{y}_i - r_i)^2}, \tag{1.100}$$

where $\bar{y}_i$ was defined in (1.8). A natural "model-weighted" least-squares cost function is then

$$\Phi(\underline{\mu}) = \sum_{i=1}^{N_Y} (\hat{l}_i - [\boldsymbol{A}\underline{\mu}]_i)^2 \frac{(\bar{y}_i(\underline{\mu}) - r_i)^2}{\bar{y}_i(\underline{\mu})}. \tag{1.101}$$

This type of cost function has been considered in [137]. Unfortunately, the above cost function is nonquadratic, so finding its minimizer is virtually as difficult as maximizing (1.15).

1.8.3.2 Data-weighted LS

A computationally simpler approach arises if we replace the estimate-dependent variance (1.100) with a data-based estimate by substituting the data Y_i for $\bar{y}_i$. This leads naturally to the following *data-weighted least-squares* cost function:

$$\Phi(\underline{\mu}) = \sum_{i=1}^{N_Y} (\hat{l}_i - [\boldsymbol{A}\underline{\mu}]_i)^2 w_i, \tag{1.102}$$

where $w_i = \frac{(Y_i - r_i)^2}{Y_i}$ is a precomputed weight. Minimizing Φ is straightforward since it is quadratic, so one can apply, for example, conjugate gradient algorithms or coordinate descent algorithms [13]. This approach gives more weight to those measurements that have lower variance, and less weight to the noiser measurements. This type of weighting can significantly reduce the noise in the reconstructed image relative to unweighted least squares. In fact, unweighted least squares estimates are essentially equivalent to FBP images. (The shift-invariant FBP methods treat all data equally, since noise is ignored.) However, as mentioned below (1.4), data-weighting leads to a systematic negative bias that increases as counts decrease [14, 15]. So (1.102) is only appropriate for moderate to high SNR problems.

One can also derive (1.102) by making a second-order Taylor expansion of the log-likelihood (1.12) about $\hat{l}_i$ [13, 14, 138, 139].

1.8.3.3 Reweighted LS

The two cost functions given above represent two extremes. In (1.102), the weights are fixed once-and-for-all prior to minimization, whereas in (1.101), the weights vary continuously as the estimate of $\underline{\mu}$ changes. A practical alternative is to first run any inexpensive algorithm (such as OSTR) for a few iterations and then reproject the estimated image $\underline{\mu}^{(n)}$ to form estimated line integrals $l_i^{(n)} = [\boldsymbol{A}\underline{\mu}^{(n)}]_i$. Then perform a second-order Taylor expansion of the log-likelihood (1.12) around $l_i^{(n)}$ to find a quadratic approximation that can be minimized easily. This approach should avoid the biases of data-weighted least-squares, and if iterated is known as *reweighted least squares* [140, 141].

1.9 Emission reconstruction

In emission tomography, the goal is to reconstruct an emission distribution $\lambda(\vec{x})$ from recorded counts of emitted photons. We again parameterize the emission distribution analogous to (1.5), letting λ_j denote the mean number of emissions from the jth voxel. The goal is to estimate $\underline{\lambda} = [\lambda_1, \ldots, \lambda_{N_p}]'$ from projection measurements $\underline{Y} = [Y_1, \ldots, Y_{N_Y}]'$. The usual Poisson measurement model is identical to (1.7), except that the measurement means are given by

$$\bar{y}_i = \sum_{j=1}^{N_p} a_{ij}\lambda_j + r_i = [\boldsymbol{A}\underline{\lambda}]_i + r_i \,, \tag{1.103}$$

where a_{ij} represents the probability that an emission from the jth voxel is recorded by the ith detector, and r_i again denotes additive background counts such as random coincidences and scatter. (Accurate models for the a_{ij}'s can lead to significant improvements in image spatial resolution and accuracy, e.g. [142, 143]. The log-likelihood has a similar form to (1.11):

$$L(\underline{\lambda}) = \sum_{i=1}^{N_Y} h_i([\boldsymbol{A}\underline{\lambda}]_i) \,,$$

where

$$h_i(l) = y_i \log(l + r_i) - (l + r_i) \,. \tag{1.104}$$

This h_i function is concave for $l \in (-r_i, \infty)$, and is strictly concave if $y_i > 0$. Since $\bar{y}_i$ is linearly related to the λ_j's (in contrast to the nonlinear relationship in (1.8) in the transmission case), the emission reconstruction problem is considerably easier than the transmission problem. Many of the algorithms described above apply to the emission problem, as well as to other inverse problems having log-likelihood functions of the general form (1.11). We describe in this section a few algorithms for maximizing the emission log-likelihood $L(\underline{\lambda})$. Extensions to the regularized problem are similar to those described for the transmission case.

1.9.1 EM Algorithm

One can derive the classical EM algorithm for the emission problem by a formal complete-data exposition [17], which is less complicated than the transmission case but still somewhat mysterious to many readers, or by fixed-point considerations [79] (which do not fully illustrate the monotonicity of the emission EM algorithm). Instead, we adopt the simple concavity-based derivation of De Pierro [72], which reinforces the surrogate function concepts woven throughout this chapter.

The key to the derivation is the following "multiplicative" trick, which applies if $\lambda_j^{(n)} > 0$:

$$[\boldsymbol{A}\underline{\lambda}]_i + r_i = \sum_{j=1}^{N_p} \left(\frac{a_{ij}\lambda_j^{(n)}}{\bar{y}_i^{(n)}} \right) \frac{\lambda_j}{\lambda_j^{(n)}} \bar{y}_i^{(n)} + \left(\frac{r_i}{\bar{y}_i^{(n)}} \right) \bar{y}_i^{(n)} . \tag{1.105}$$

The $N_p + 1$ terms in parentheses are nonnegative and sum to unity, so we can apply the concavity inequality. Since $g_i(l) \triangleq y_i \log l - l$ is concave on $(0, \infty)$, it follows that

$$\begin{aligned} L(\underline{\lambda}) &= \sum_{i=1}^{N_Y} g_i([\boldsymbol{A}\underline{\lambda}]_i + r_i) \\ &= \sum_{i=1}^{N_Y} g_i \left(\sum_{j=1}^{N_p} \left(\frac{a_{ij}\lambda_j^{(n)}}{\bar{y}_i^{(n)}} \right) \frac{\lambda_j}{\lambda_j^{(n)}} \bar{y}_i^{(n)} + \left(\frac{r_i}{\bar{y}_i^{(n)}} \right) \bar{y}_i^{(n)} \right) \\ &\geq \sum_{i=1}^{N_Y} \sum_{j=1}^{N_p} \left(\frac{a_{ij}\lambda_j^{(n)}}{\bar{y}_i^{(n)}} \right) g_i \left(\frac{\lambda_j}{\lambda_j^{(n)}} \bar{y}_i^{(n)} \right) + \left(\frac{r_i}{\bar{y}_i^{(n)}} \right) g_i(\bar{y}_i^{(n)}) \triangleq Q(\underline{\lambda}; \underline{\lambda}^{(n)}) . \end{aligned}$$

The surrogate function Q is separable:

$$Q(\underline{\lambda}; \underline{\lambda}^{(n)}) = \sum_{j=1}^{N_p} Q_j(\lambda_j; \underline{\lambda}^{(n)}), \quad Q_j(\lambda_j; \underline{\lambda}^{(n)}) \equiv \sum_{i=1}^{N_Y} \frac{a_{ij}\lambda_j^{(n)}}{\bar{y}_i^{(n)}} g_i \left(\frac{\lambda_j}{\lambda_j^{(n)}} \bar{y}_i^{(n)} \right) . \tag{1.106}$$

Thus the the following parallelizable maximization step is guaranteed to monotonically increase the log-likelihood $L(\underline{\lambda})$ each iteration:

$$\lambda_j^{(n+1)} = \arg \max_{\lambda_j \geq 0} Q_j(\lambda_j; \underline{\lambda}^{(n)}) . \tag{1.107}$$

The maximization is trivial:

$$\frac{\partial}{\partial \lambda_j} Q_j(\lambda_j; \underline{\lambda}^{(n)}) = \sum_{i=1}^{N_Y} a_{ij} \dot{g}_i \left(\frac{\lambda_j}{\lambda_j^{(n)}} \bar{y}_i^{(n)} \right) = \sum_{i=1}^{N_Y} a_{ij} \left[\frac{\lambda_j^{(n)}}{\lambda_j} \frac{y_i}{\bar{y}_i^{(n)}} - 1 \right] .$$

Equating to zero and solving for λ_j yields the famous update:

$$\boxed{\lambda_j^{(n+1)} = \lambda_j^{(n)} \frac{\sum_{i=1}^{N_Y} a_{ij} y_i / \bar{y}_i^{(n)}}{\sum_{i=1}^{N_Y} a_{ij}}, \; j = 1, \ldots, N_p.} \tag{1.108}$$

Unfortunately, the emission EM algorithm (1.108) usually converges painfully slowly. To understand this, consider the curvatures of the surrogate functions Q_j:

$$-\frac{\partial^2}{\partial \lambda_j^2} Q_j(\lambda_j; \underline{\lambda}^{(n)}) = \frac{\lambda_j^{(n)}}{\lambda_j^2} \sum_{i=1}^{N_Y} a_{ij} \frac{y_i}{\bar{y}_i^{(n)}}. \tag{1.109}$$

For any pixels converging towards zero, these curvatures grow without bound. This leads to very slow convergence; even sublinear convergence rates are possible [76].

1.9.2 An improved EM algorithm

One can choose a slightly better decomposition than (1.105) to get slightly faster converging EM algorithms [75]. First find any set of nonnegative constants $\{m_j\}_{j=1}^{N_p}$ that satisfy

$$r_i \geq \sum_{j=1}^{N_p} a_{ij} m_j, \forall i. \tag{1.110}$$

Then an alternative to (1.105) is:

$$[\boldsymbol{A}\underline{\lambda}]_i + r_i = \sum_{j=1}^{N_p} \left(\frac{a_{ij}(\lambda_j^{(n)} + m_j)}{\bar{y}_i^{(n)}} \right) \frac{\lambda_j}{\lambda_j^{(n)} + m_j} \bar{y}_i^{(n)} + \left(\frac{\hat{r}_i}{\bar{y}_i^{(n)}} \right) \bar{y}_i^{(n)}, \tag{1.111}$$

where $\hat{r}_i = r_i - \sum_{j=1}^{N_p} a_{ij} m_j \geq 0$. Again the terms in parentheses in (1.111) are nonnegative and sum to unity. So a similar derivation as that yielding (1.106) leads to a new surrogate function:

$$Q_j(\lambda_j; \underline{\lambda}^{(n)}) = \sum_{i=1}^{N_Y} \frac{a_{ij}(\lambda_j^{(n)} + m_j)}{\bar{y}_i^{(n)}} g_i \left(\frac{\lambda_j + m_j}{\lambda_j^{(n)}} \bar{y}_i^{(n)} \right).$$

Maximizing as in (1.107) leads to the following algorithm

$$\boxed{\lambda_j^{(n+1)} = \left[(\lambda_j^{(n)} + m_j) \frac{\sum_{i=1}^{N_Y} a_{ij} y_i / \bar{y}_i^{(n)}}{\sum_{i=1}^{N_Y} a_{ij}} - m_j \right]_+, \; j = 1, \ldots, N_p.} \tag{1.112}$$

This algorithm was derived by a more complicated EM approach in [75], and called ML-EM-3. The surrogate function derivation is simpler to present and understand, and more readily generalizable to alternative surrogates.

The curvatures of the second Q_j's just derived are smaller than those in (1.106), due to the m_j's (replace λ_j with $\lambda_j + m_j$ in the denominator of (1.109)). The convergence rate improves as the m_j's increase, but of course (1.110) must be satisfied to ensure monotonicity. Since the EM algorithm updates all parameters simultaneously, the m_j values must be "shared" among all pixels, and typically are fairly small due to (1.110). In contrast the SAGE algorithm [75] updates the pixels sequentially, which greatly relaxes the constraints on the m_j's, allowing larger values and hence faster convergence rates.

1.9.3 Other emission algorithms

Most of the other methods for developing reconstruction algorithms described in this chapter have counterparts for the emission problem. Monotonic acceleration is possible using line searches [94]. Replacing the sums over i in (1.108) with sums over subsets of the projections yields the emission OSEM algorithm [102]; see also the related variants RAMLA [144] and RBBI [103, 104]. Although the OSEM algorithm fails to converge in general, it often gives reasonable looking images in a small number of iterations when initialized with a uniform image. Sequential updates rather than parallel updates leads to the fast converging SAGE algorithms [75] and coordinate ascent algorithms [19], including paraboloidal surrogate variations thereof [78]. The conjugate gradient algorithm has been applied extensively to the emission problem and is particularly effective provided one carefully treats the nonnegativity constraints [31].

1.10 Advanced topics

In this section we provide pointers to the literature for several additional topics, all of which are active research areas.

1.10.1 Choice of regularization parameters

A common critique of penalized-likelihood and Bayesian methods is that one must choose (subjectively?) the regularization parameter β in (1.15). (In unregularized methods there are also free parameters that one must select, such as the number of iterations or the amount of post-filtering, but fiddling these factors to get visually pleasing images is perhaps easier than adjusting β, since each new β requires another run of the iterative algorithm.) A large variety of methods for automatically choosing β have been proposed, based on principles such as maximum likelihood or *cross validation* e.g., [145–154], most of which have been evaluated in terms of mean-squared error performance, which equally weights bias (squared) and variance, even though resolution and noise may have unequal importance in imaging problems.

For quadratic regularization methods, one can choose both β and $R(\underline{\mu})$ to control the spatial resolution properties and to relate the desired spatial resolution to an appropriate value of β by a predetermined table [20, 54–57].

In addition to the variety of methods for choosing β, there is an even larger variety of possible choices for the "potential functions" ψ_k in (1.17), ranging from quadratic to nonquadratic to nonconvex and even nondifferentiable. See [155] for a recent discussion.

The absolute value potential ($\psi_k(t) = |t|$) is particularly appealing in problems with piecewise constant attenuation maps. However, its nondifferentiability greatly complicates optimization [156–158].

1.10.2 Source-free attenuation reconstruction

In PET and SPECT imaging, the attenuation map is a nuisance parameter; the emission distribution is of greatest interest. This has spawned several attempts to estimate the attenuation map from the emission sinograms, without a separate transmission scan. See e.g., [132, 159, 160].

1.10.3 Dual energy imaging

We have focused on the case of monoenergetic imaging, by the assumption (1.23). For quantitative applications such as bone densitometry, one must account for the polyenergetic property of x-ray source spectra. A variety of methods have been proposed for dual energy image reconstruction, including (recently) statistical methods [161–163].

1.10.4 Overlapping beams

Some transmission scan geometries involve multiple transmission sources, and it is possible for a given detector element to record photons that originated in more than one of these sources, i.e., the beams of photons emitted from the various sources overlap on the detector. The transmission statistical model (1.8) must be generalized to account for such overlap, leading to new reconstruction algorithms [164–167].

1.10.5 Sinogram truncation and limited angles

In certain geometries, portions of the sinogram are missing due to geometric truncation (such as fan-beam geometries with a short focal length). In such cases, prior information plays an essential role in regularizing the reconstruction problem, e.g., [43]. Similarly, in *limited angle tomography* the sinograms are truncated due to missing angles. Nonquadratic regularization methods have shown considerable promise for such problems [68].

1.10.6 Parametric object priors

Throughout this chapter we have considered the image to be parameterized by the linear series expansion (1.5), and the associated regularization methods have used only fairly generic image properties, such as piecewise smoothness. In some applications, particularly when the counts are extremely low or the number of

projection views is limited, it can be desirable (or even essential) to apply much stronger prior information to the reconstruction problem. Simple parametric object models such as circles and ellipses (with unknown location, shape, and intensity parameters) have been used for certain applications such as angiography [168–172] or for analysis of imaging system designs, e.g., [173, 174]. Polygonal models have been applied to cardiac image reconstruction, e.g., [175]. More general and flexible object models based on deformable templates have also shown considerable promise and comprise a very active research area, e.g., [176–180]. (See Chapter 3.)

1.11 Example results

This section presents representative results of applying penalized likelihood image reconstruction to real PET transmission scan data, following [109]. Many more examples can be found in the references cited throughout this chapter.

We collected a 12–minute transmission scan (y_i's) on a Siemens/CTI ECAT EXACT 921 PET scanner with rotating rod sources of an anthropomorphic thorax phantom (Data Spectrum, Chapel Hill, NC). The sinogram size was 160 radial bins by 192 angles (over 180°), with 3mm radial spacing. The reconstructed images were 128×128 pixels that were 4.5mm on each side.

For the penalized-likelihood reconstructions we used a second order penalty function of the form (1.16) with the following potential function proposed in [97]:

$$\psi(t) = \delta^2 \left[\, |t/\delta| - \log\left(1 + |t/\delta|\right) \, \right] , \qquad (1.113)$$

where $\beta = 2^{10}$ and $\delta = 0.004\text{cm}^{-1}$ were chosen visually. This function approaches the quadratic $\psi(t) = t^2/2$ as $\delta \to \infty$, but provides a degree of edge preservation for finite δ. The derivative of ψ requires no transcendental functions, which is computationally desirable.

Fig. 1.6 presents a representative example of performance on real PET transmission data. The FBP image is noisy and blurry, since there are only 921K prompt coincidences in this scan [101]. In the upper right is the emission OSEM algorithm applied to the logarithm (1.3) and (1.99). As discussed in Section 1.8.2, this approach yields suboptimal images. The lower two images in Fig. 1.6 were reconstructed by penalized likelihood methods based on (1.15) and (1.16) with the penalty described above. The lower left image used 2 iterations of the OSTR-16 algorithm (1.96) (modified for a nonquadratic penalty as described in [77]). The lower right image used 10 iterations of the PSCA algorithm (1.97). Both penalized likelihood images have lower noise and better spatial resolution than FBP and ML-OSEM-8, as quantified in [77]. There are small differences between the nonconvergent OSTR image and the image reconstructed by the monotonic PSCA algorithm, but whether these differences are important in practice is an open question.

In PET the main purpose of the attenuation map is to form attenuation correction factors (ACFs) for the emission scan. Fig. 1.7 shows sagittal views (47 slices)

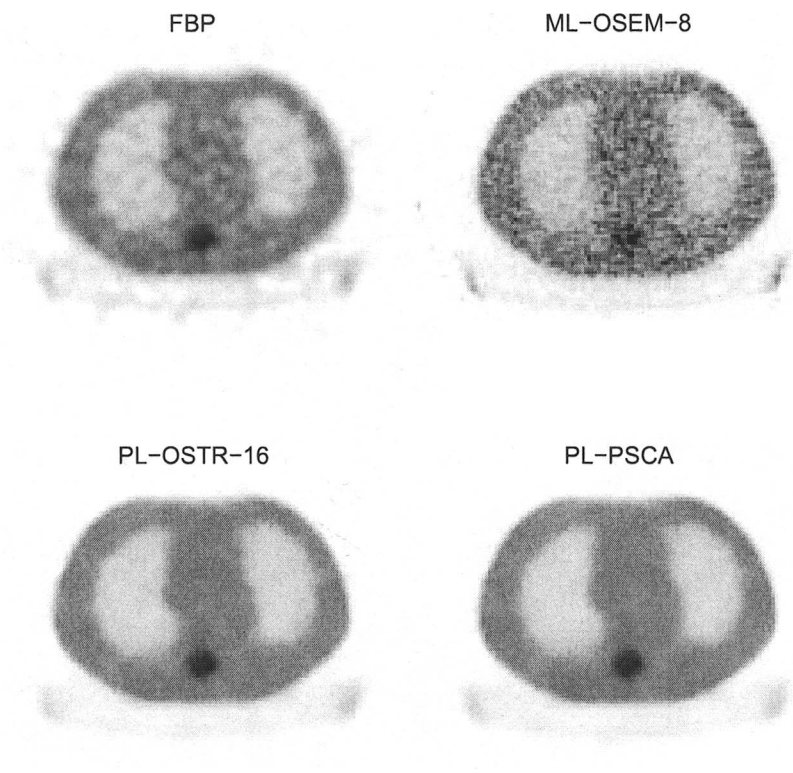

Figure 1.6: Reconstructed images of thorax phantom from 12–minute PET transmission scan.

of a patient injected with FDG and scanned with PET. In this case, a 2–minute transmission scan was emulated by binomial thinning of a 12–minute transmission scan [109]. For the subfigures labeled "T-PL" and "T-FBP" the ACFs were computed from attenuation maps reconstructed by penalized likelihood methods or by FBP respectively. For the subfigures labeled "E-PL" and "E-FBP," the emission data was reconstructed by penalized likelihood methods or by FBP respectively. The best image (upper left) is formed when both the emission and transmission images are reconstructed by statistical approaches. The second best image (upper right) is formed by using statistical reconstruction of the attenuation map, but ordinary FBP for the emission data. Clearly for such low-count transmission scans, reducing the noise in the ACFs is as important, if not more so, than how the emission images are reconstructed.

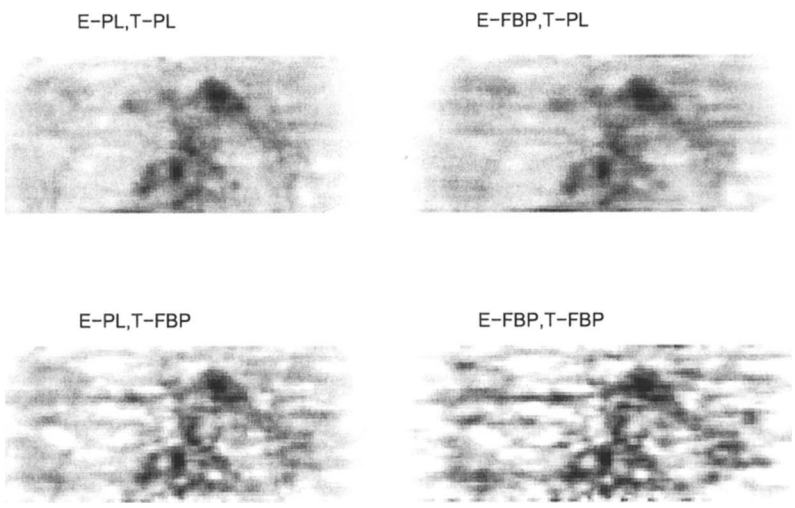

Figure 1.7: FDG PET emission images, reconstructed by both FBP (E-FBP) and penalized-likelihood (E-PL) methods. Attenuation correction was performed using attenuation maps generated either by transmission FBP (T-FBP) or transmission penalized-likelihood (T-PL) reconstructions. The use of statistical reconstruction methods significantly reduces image noise.

1.12 Summary

We have summarized a wide variety of algorithms for statistical image reconstruction from transmission measurements. Most of the ideas underlying these algorithms are applicable to emission tomography, as well as to image recovery problems in general.

There is a wide variety of algorithms in part because there is yet to have been found any algorithm that has all the desirable properties listed in Section 1.3.1. In cases where the system matrix $\boldsymbol{A}$ can easily be precomputed and stored, and a non-parallel computer is to be used, we recommend the PSCA algorithm of Section 1.6.4. For parallel computing, the conjugate gradient algorithm [31] is a reasonable choice, particularly if exact nonnegativity constraints can be relaxed. If an inexact maximum is acceptable, the OSTR algorithm of Section 1.6.3 is a very practical choice, and is likely to be widely applied given the popularity of the emission OSEM algorithm. Meanwhile, the search continues for an algorithm with the simplicity of OSTR that is parallelizable, monotone and fast converging, and can accommodate any form of system matrix.

1.13 Acknowledgements

The ideas in this chapter were greatly influenced by the dissertation research of Hakan Erdoğan [181], who also prepared Fig. 1.6 and 1.7. The author also gratefully acknowledges ongoing collaboration with Neal Clinthorne, Ed Ficaro, Ken Lange, and Les Rogers. The author thanks Ken Hanson for his careful reading of this chapter. This work was supported in part by NIH grants CA-60711 and CA-54362.

1.14 Appendix: Poisson properties

Suppose a source transmits N photons of a certain energy along a ray passing through an object towards a specified pixel on the detector. We assume N is a Poisson random variable with mean N_0:

$$P[N = k] = \frac{1}{k!} e^{-N_0} N_0^k .$$

Each of the N transmitted photons may either pass unaffected ("survive" passage) or may interact with the object. These are Bernoulli trials since the photons interact independently. From Beer's law we know that the probability of surviving passage is given by

$$p = e^{-\int \mu(z)\, dz} .$$

The number of photons M that pass unaffected through the object is a random variable, and from Beer's law:

$$P[M = m | N = n] = \binom{n}{m} p^m (1-p)^{n-m}, \; m = 0, \ldots, n .$$

Using total probability:

$$P[M = m] = \sum_n P[M = m | N = n] P[N = n]$$

$$= \sum_{n=m}^{\infty} \binom{n}{m} p^m (1-p)^{n-m} \frac{1}{n!} e^{-N_0} N_0^n = \frac{1}{m!} e^{-N_0 p} (N_0 p)^m, \; m = 0, 1, \ldots .$$

Therefore the distribution of photons that survive passage is also Poisson, with mean $E[M] = N_0 p$.

Furthermore, by applying Bayes' rule, for $n \geq m \geq 0$:

$$\begin{aligned} P[N=n|M=m] &= \frac{P[M=m|N=n]P[N=n]}{P[M=m]} \\ &= \frac{\binom{n}{m} p^m (1-p)^{n-m} \ \frac{1}{n!} e^{-N_0} N_0^n}{\frac{1}{m!} e^{-N_0 p} (N_0 p)^m} \\ &= \frac{1}{(n-m)!} (N_0 - N_0 p)^{n-m} e^{-(N_0 - N_0 p)} \\ &= \frac{1}{(n-m)!} (E[N] - E[M])^{n-m} e^{-(E[N]-E[M])} . \end{aligned}$$

Thus, conditioned on M, the random variable $N - M$ has a Poisson distribution with mean $E[N] - E[M]$. In particular,

$$E[N - M|M] = E[N] - E[M] ,$$

which is useful in deriving the transmission EM algorithm proposed in [17].

1.15 References

[1] M. M. Ter-Pogossian, M. E. Raichle, and B. E. Sobel, "Positron-emission tomography," *Scientific American*, vol. 243, pp. 171–181, Oct. 1980.

[2] J. M. Ollinger and J. A. Fessler, "Positron emission tomography," *IEEE Sig. Proc. Mag.*, vol. 14, pp. 43–55, Jan. 1997.

[3] T. F. Budinger and G. T. Gullberg, "Three dimensional reconstruction in nuclear medicine emission imaging," *IEEE Tr. Nuc. Sci.*, vol. 21, no. 3, pp. 2–20, 1974.

[4] T. H. Prettyman, R. A. Cole, R. J. Estep, and G. A. Sheppard, "A maximum-likelihood reconstruction algorithm for tomographic gamma-ray nondestructive assay," *Nucl. Instr. Meth. Phys. Res. A.*, vol. 356, pp. 407–52, Mar. 1995.

[5] G. Wang, D. L. Snyder, J. A. O'Sullivan, and M. W. Vannier, "Iterative deblurring for CT metal artifact reduction," *IEEE Tr. Med. Im.*, vol. 15, p. 657, Oct. 1996.

[6] R. M. Leahy and J. Qi, "Statistical approaches in quantitative positron emission tomography," *Statistics and Computing*, 1998.

[7] K. Wienhard, L. Eriksson, S. Grootoonk, M. Casey, U. Pietrzyk, and W. D. Heiss, "Performance evaluation of a new generation positron scanner ECAT EXACT," *J. Comp. Assisted Tomo.*, vol. 16, pp. 804–813, Sept. 1992.

[8] S. R. Cherry, M. Dahlbom, and E. J. Hoffman, "High sensitivity, total body PET scanning using 3D data acquisition and reconstruction," *IEEE Tr. Nuc. Sci.*, vol. 39, pp. 1088–1092, Aug. 1992.

[9] E. P. Ficaro, J. A. Fessler, W. L. Rogers, and M. Schwaiger, "Comparison of Americium-241 and Technicium-99m as transmission sources for attenuation correction of Thallium-201 SPECT imaging of the heart," *J. Nuc. Med.*, vol. 35, pp. 652–63, Apr. 1994.

[10] S. R. Meikle, M. Dahlbom, and S. R. Cherry, "Attenuation correction using count-limited transmission data in positron emission tomography," *J. Nuc. Med.*, vol. 34, pp. 143–150, Jan. 1993.

[11] G. T. Herman, *Image reconstruction from projections: The fundamentals of computerized tomography*. New York: Academic Press, 1980.

[12] A. Macovski, *Medical imaging systems*. New Jersey: Prentice-Hall, 1983.

[13] K. Sauer and C. Bouman, "A local update strategy for iterative reconstruction from projections," *IEEE Tr. Sig. Proc.*, vol. 41, pp. 534–548, Feb. 1993.

[14] J. A. Fessler, "Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans," *IEEE Tr. Im. Proc.*, vol. 4, pp. 1439–50, Oct. 1995.

[15] D. S. Lalush and B. M. W. Tsui, "MAP-EM and WLS-MAP-CG reconstruction methods for transmission imaging in cardiac SPECT," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1174–1178, 1993.

[16] J. A. Fessler, "Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography," *IEEE Tr. Im. Proc.*, vol. 5, pp. 493–506, Mar. 1996.

[17] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comp. Assisted Tomo.*, vol. 8, pp. 306–316, Apr. 1984.

[18] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proc. 27th Conf. Info. Sci. Sys., Johns Hopkins*, pp. 611–616, 1993.

[19] C. A. Bouman and K. Sauer, "A unified approach to statistical tomography using coordinate descent optimization," *IEEE Tr. Im. Proc.*, vol. 5, pp. 480–92, Mar. 1996.

[20] J. A. Fessler and W. L. Rogers, "Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs," *IEEE Tr. Im. Proc.*, vol. 5, pp. 1346–58, Sept. 1996.

[21] P. M. Joseph and R. D. Spital, "A method for correcting bone induced artifacts in computed tomography scanners," *J. Comp. Assisted Tomo.*, vol. 2, pp. 100–8, 1978.

[22] B. Chan, M. Bergström, M. R. Palmer, C. Sayre, and B. D. Pate, "Scatter distribution in transmission measurements with positron emission tomography," *J. Comp. Assisted Tomo.*, vol. 10, pp. 296–301, Mar. 1986.

[23] E. J. Hoffman, S. C. Huang, M. E. Phelps, and D. E. Kuhl, "Quantitation in positron emission computed tomography: 4 Effect of accidental coincidences," *J. Comp. Assisted Tomo.*, vol. 5, no. 3, pp. 391–400, 1981.

[24] M. E. Casey and E. J. Hoffman, "Quantitation in positron emission computed tomography: 7 a technique to reduce noise in accidental coincidence measurements and coincidence efficiency calibration," *J. Comp. Assisted Tomo.*, vol. 10, no. 5, pp. 845–850, 1986.

[25] E. P. Ficaro, W. L. Rogers, and M. Schwaiger, "Comparison of Am-241 and Tc-99m as transmission sources for the attenuation correction of Tl-201 cardiac SPECT studies," *J. Nuc. Med. (Abs. Book)*, vol. 34, p. 30, May 1993.

[26] E. P. Ficaro, J. A. Fessler, R. J. Ackerman, W. L. Rogers, J. R. Corbett, and M. Schwaiger, "Simultaneous transmission-emission Tl-201 cardiac SPECT: Effect of attenuation correction on myocardial tracer distribution," *J. Nuc. Med.*, vol. 36, pp. 921–31, June 1995.

[27] E. P. Ficaro, J. A. Fessler, P. D. Shreve, J. N. Kritzman, P. A. Rose, and J. R. Corbett, "Simultaneous transmission/emission myocardial perfusion tomography: Diagnostic accuracy of attenuation-corrected 99m-Tc-Sestamibi SPECT," *Circulation*, vol. 93, pp. 463–73, Feb. 1996.

[28] D. F. Yu and J. A. Fessler, "Mean and variance of photon counting with deadtime," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 1999.

[29] D. F. Yu and J. A. Fessler, "Mean and variance of singles photon counting with deadtime," *Phys. Med. Biol.*, 1999. Submitted.

[30] D. F. Yu and J. A. Fessler, "Mean and variance of coincidence photon counting with deadtime," *Phys. Med. Biol.*, 1999. Submitted.

[31] E. U. Mumcuoglu, R. Leahy, S. R. Cherry, and Z. Zhou, "Fast gradient-based methods for Bayesian reconstruction of transmission and emission PET images," *IEEE Tr. Med. Im.*, vol. 13, pp. 687–701, Dec. 1994.

[32] M. Yavuz and J. A. Fessler, "New statistical models for randoms-precorrected PET scans," in *Information Processing in Medical Im.* (J. Duncan and G. Gindi, eds.), vol. 1230 of *Lecture Notes in Computer Science*, pp. 190–203, Berlin: Springer Verlag, 1997.

[33] M. Yavuz and J. A. Fessler, "Statistical image reconstruction methods for randoms-precorrected PET scans," *Med. Im. Anal.*, vol. 2, no. 4, pp. 369–378, 1998.

[34] M. Yavuz and J. A. Fessler, "Penalized-likelihood estimators and noise analysis for randoms-precorrected PET transmission scans," *IEEE Tr. Med. Im.*, vol. 18, pp. 665–74, Aug. 1999.

[35] A. C. Kak and M. Slaney, *Principles of computerized tomographic imaging*. New York: IEEE Press, 1988.

[36] A. Celler, A. Sitek, E. Stoub, P. Hawman, R. Harrop, and D. Lyster, "Multiple line source array for SPECT transmission scans: Simulation, phantom and patient studies," *J. Nuc. Med.*, vol. 39, pp. 2183–9, Dec. 1998.

[37] E. L. Lehmann, *Theory of point estimation*. New York: Wiley, 1983.

[38] K. Sauer and B. Liu, "Nonstationary filtering of transmission tomograms in high photon counting noise," *IEEE Tr. Med. Im.*, vol. 10, pp. 445–452, Sept. 1991.

[39] J. A. Fessler, "Tomographic reconstruction using information weighted smoothing splines," in *Information Processing in Medical Im.* (H. H. Barrett and A. F. Gmitro, eds.), vol. 687 of *Lecture Notes in Computer Science*, pp. 372–86, Berlin: Springer Verlag, 1993.

[40] M. N. Wernick and C. T. Chen, "Superresolved tomography by convex projections and detector motion," *J. Opt. Soc. Am. A*, vol. 9, pp. 1547–1553, Sept. 1992.

[41] S. H. Manglos, "Truncation artifact suppression in cone-beam radionuclide transmission CT using maximum likelihood techniques: evaluation with human subjects," *Phys. Med. Biol.*, vol. 37, pp. 549–562, Mar. 1992.

[42] E. P. Ficaro and J. A. Fessler, "Iterative reconstruction of truncated fan beam transmission data," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 3, 1993.

[43] J. A. Case, T. S. Pan, M. A. King, D. S. Luo, B. C. Penney, and M. S. Z. Rabin, "Reduction of truncation artifacts in fan beam transmission imaging using a spatially varying gamma prior," *IEEE Tr. Nuc. Sci.*, vol. 42, pp. 2260–5, Dec. 1995.

[44] S. H. Manglos, G. M. Gagne, A. Krol, F. D. Thomas, and R. Narayanaswamy, "Transmission maximum-likelihood reconstruction with ordered subsets for cone beam CT," *Phys. Med. Biol.*, vol. 40, pp. 1225–41, July 1995.

[45] T.-S. Pan, B. M. W. Tsui, and C. L. Bryne, "Choice of initial conditions in the ML reconstruction of fan-beam transmission with truncated projection data," *IEEE Tr. Med. Im.*, vol. 16, pp. 426–38, Aug. 1997.

[46] G. L. Zeng and G. T. Gullberg, "An SVD study of truncated transmission data in SPECT," *IEEE Tr. Nuc. Sci.*, vol. 44, pp. 107–11, Feb. 1997.

[47] A. J. Rockmore and A. Macovski, "A maximum likelihood approach to transmission image reconstruction from projections," *IEEE Tr. Nuc. Sci.*, vol. 24, pp. 1929–1935, June 1977.

[48] Y. Censor, "Finite series expansion reconstruction methods," *Proc. IEEE*, vol. 71, pp. 409–419, Mar. 1983.

[49] R. Schwinger, S. Cool, and M. King, "Area weighted convolutional interpolation for data reprojection in single photon emission computed tomography," *Med. Phys.*, vol. 13, pp. 350–355, May 1986.

[50] S. C. B. Lo, "Strip and line path integrals with a square pixel matrix: A unified theory for computational CT projections," *IEEE Tr. Med. Im.*, vol. 7, pp. 355–363, Dec. 1988.

[51] J. A. Fessler, "ASPIRE 3.0 user's guide: A sparse iterative reconstruction library," Tech. Rep. 293, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, July 1995. Available from `http://www.eecs.umich.edu/~fessler`.

[52] D. L. Snyder, M. I. Miller, L. J. Thomas, and D. G. Politte, "Noise and edge artifacts in maximum-likelihood reconstructions for emission tomography," *IEEE Tr. Med. Im.*, vol. 6, pp. 228–238, Sept. 1987.

[53] J. A. Browne and T. J. Holmes, "Maximum likelihood techniques in X-ray computed tomography," in *Medical imaging systems techniques and applications: Diagnosis optimization techniques* (C. T. Leondes, ed.), vol. 3, pp. 117–46, Amsteldijk, Netherlands: Gordon and Breach, 1997.

[54] J. A. Fessler, "Resolution properties of regularized image reconstruction methods," Tech. Rep. 297, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Aug. 1995.

[55] J. A. Fessler and W. L. Rogers, "Uniform quadratic penalties cause nonuniform image resolution (and sometimes vice versa)," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 4, pp. 1915–1919, 1994.

[56] J. W. Stayman and J. A. Fessler, "Spatially-variant roughness penalty design for uniform resolution in penalized-likelihood image reconstruction," in *Proc. IEEE Intl. Conf. on Image Processing*, vol. 2, pp. 685–9, 1998.

[57] J. W. Stayman and J. A. Fessler, "Regularization for uniform spatial resolution properties in penalized-likelihood image reconstruction," *IEEE Tr. Med. Im.*, 1998.

[58] J. Qi and R. M. Leahy, "A theoretical study of the contrast recovery and variance of MAP reconstructions with applications to the selection of smoothing parameters," *IEEE Tr. Med. Im.*, vol. 18, pp. 293–305, Apr. 1999.

[59] J. Qi and R. M. Leahy, "Resolution and noise properties of MAP reconstruction for fully 3D PET," in *Proc. of the 1999 Intl. Mtg. on Fully 3D Im. Recon. in Rad. Nuc. Med.*, pp. 35–9, 1999.

[60] J. Besag, "On the statistical analysis of dirty pictures," *J. Royal Stat. Soc. Ser. B*, vol. 48, no. 3, pp. 259–302, 1986.

[61] S. C. Huang, R. E. Carson, M. E. Phelps, E. J. Hoffman, H. R. Schelbert, and D. E. Kuhl, "A boundary method for attenuation correction in positron computed tomography," *J. Nuc. Med.*, vol. 22, no. 1, pp. 627–637, 1981.

[62] E. Z. Xu, N. A. Mullani, K. L. Gould, and W. L. Anderson, "A segmented attenuation correction for PET," *J. Nuc. Med.*, vol. 32, pp. 161–165, Jan. 1991.

[63] M. Xu, W. K. Luk, P. D. Cutler, and W. M. Digby, "Local threshold for segmented attenuation correction of PET imaging of the thorax," *IEEE Tr. Nuc. Sci.*, vol. 41, pp. 1532–1537, Aug. 1994.

[64] S.-J. Lee, A. Rangarajan, and G. Gindi, "Bayesian image reconstruction in SPECT using higher order mechanical models as priors," *IEEE Tr. Med. Im.*, vol. 14, pp. 669–80, Dec. 1995.

[65] G. T. Herman, D. Odhner, K. D. Toennies, and S. A. Zenios, "A parallelized algorithm for image reconstruction from noisy projections," in *Large-Scale Numerical Optimization* (T. F. Coleman and Y. Li, eds.), pp. 3–21, Philadelphia: SIAM, 1990.

[66] S. Alenius, U. Ruotsalainen, and J. Astola, "Using local median as the location of the prior distribution in iterative emission tomography image reconstruction," *IEEE Tr. Nuc. Sci.*, vol. 45, pp. 3097–104, Dec. 1998.

[67] J. A. Fessler and S. D. Booth, "Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction," *IEEE Tr. Im. Proc.*, vol. 8, pp. 688–99, May 1999.

[68] A. H. Delaney and Y. Bresler, "Globally convergent edge-preserving regularized reconstruction: an application to limited-angle tomography," *IEEE Tr. Im. Proc.*, vol. 7, pp. 204–221, Feb. 1998.

[69] K. Lange and J. A. Fessler, "Globally convergent algorithms for maximum a posteriori transmission tomography," *IEEE Tr. Im. Proc.*, vol. 4, pp. 1430–8, Oct. 1995.

[70] R. R. Meyer, "Sufficient conditions for the convergence of monotonic mathematical programming algorithms," *J. Comput. System. Sci.*, vol. 12, pp. 108–21, 1976.

[71] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*. New York: Academic Press, 1970.

[72] A. R. De Pierro, "On the relation between the ISRA and the EM algorithm for positron emission tomography," *IEEE Tr. Med. Im.*, vol. 12, pp. 328–333, June 1993.

[73] A. R. De Pierro, "A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography," *IEEE Tr. Med. Im.*, vol. 14, pp. 132–137, Mar. 1995.

[74] K. Lange, *Numerical analysis for statisticians*. New York: Springer-Verlag, 1999.

[75] J. A. Fessler and A. O. Hero, "Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms," *IEEE Tr. Im. Proc.*, vol. 4, pp. 1417–29, Oct. 1995.

[76] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers, "On complete data spaces for PET reconstruction algorithms," *IEEE Tr. Nuc. Sci.*, vol. 40, pp. 1055–61, Aug. 1993.

[77] H. Erdoğan and J. A. Fessler, "Monotonic algorithms for transmission tomography," *IEEE Tr. Med. Im.*, vol. 18, pp. 801–14, Sept. 1999.

[78] J. A. Fessler and H. Erdoğan, "A paraboloidal surrogates algorithm for convergent penalized-likelihood emission image reconstruction," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1132–5, 1998.

[79] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Tr. Med. Im.*, vol. 1, pp. 113–122, Oct. 1982.

[80] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38, 1977.

[81] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, no. 1, pp. 95–103, 1983.

[82] X. L. Meng and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: A general framework," *Biometrika*, vol. 80, no. 2, pp. 267–278, 1993.

[83] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Tr. Sig. Proc.*, vol. 42, pp. 2664–77, Oct. 1994.

[84] C. H. Liu and D. B. Rubin, "The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence," *Biometrika*, vol. 81, no. 4, pp. 633–48, 1994.

[85] K. Lange, "A Quasi-Newton acceleration of the EM Algorithm," *Statistica Sinica*, vol. 5, pp. 1–18, Jan. 1995.

[86] D. A. van Dyk, X. L. Meng, and D. B. Rubin, "Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance," *Statistica Sinica*, vol. 5, pp. 55–76, Jan. 1995.

[87] X. L. Meng and D. van Dyk, "The EM algorithm - An old folk song sung to a fast new tune," *J. Royal Stat. Soc. Ser. B*, vol. 59, no. 3, pp. 511–67, 1997.

[88] J. A. Fessler, "EM and gradient algorithms for transmission tomography with background contamination," Tech. Rep. UM-PET-JF-94-1, Cyclotron PET Facility, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Dec. 1994. Available from `http://www.eecs.umich.edu/~fessler`.

[89] J. T. Kent and C. Wright, "Some suggestions for transmission tomography based on the EM algorithm," in *Stochastic Models, Statistical Methods, and Algorithms in Im. Analysis* (M. P. P Barone, A Frigessi, ed.), vol. 74 of *Lecture Notes in Statistics*, pp. 219–232, New York: Springer, 1992.

[90] H. Stark, *Image recovery, theory and application.* Orlando: Academic, 1987.

[91] J. A. Browne and T. J. Holmes, "Developments with maximum likelihood X-ray computed tomography," *IEEE Tr. Med. Im.*, vol. 12, pp. 40–52, Mar. 1992.

[92] J. M. Ollinger and G. C. Johns, "The use of maximum *a-posteriori* and maximum likelihood transmission images for attenuation correction PET," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1185–1187, 1992.

[93] J. M. Ollinger, "Maximum likelihood reconstruction of transmission images in emission computed tomography via the EM algorithm," *IEEE Tr. Med. Im.*, vol. 13, pp. 89–101, Mar. 1994.

[94] L. Kaufman, "Implementing and accelerating the EM algorithm for positron emission tomography," *IEEE Tr. Med. Im.*, vol. 6, pp. 37–51, Mar. 1987.

[95] K. Lange, M. Bahn, and R. Little, "A theoretical study of some maximum likelihood algorithms for emission and transmission tomography," *IEEE Tr. Med. Im.*, vol. 6, pp. 106–114, June 1987.

[96] K. Lange, "An overview of Bayesian methods in image reconstruction," in *Proc. SPIE 1351, Dig. Im. Synth. and Inverse Optics*, pp. 270–287, 1990.

[97] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Tr. Med. Im.*, vol. 9, pp. 439–446, Dec. 1990. Corrections, T-MI, 10:2(288), June 1991.

[98] P. J. Maniawski, H. T. Morgan, G. L. Gullberg, G. L. Zeng, A. E. Welch, and C. H. Tung, "Performance evaluation of a transmission reconstruction algorithm with simultaneous transmission-emission SPECT system in a presence of data truncation," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 4, pp. 1578–81, 1994.

[99] S. G. Kim, "Draft on some modifications of GCA algorithms for transmission CT," 1998. preprint Dec. 16, 1998.

[100] J. A. Fessler and A. O. Hero, "Complete-data spaces and generalized EM algorithms," in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, vol. 4, pp. 1–4, 1993.

[101] J. A. Fessler, E. P. Ficaro, N. H. Clinthorne, and K. Lange, "Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction," *IEEE Tr. Med. Im.*, vol. 16, pp. 166–75, Apr. 1997.

[102] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Tr. Med. Im.*, vol. 13, pp. 601–609, Dec. 1994.

[103] C. L. Byrne, "Block-iterative methods for image reconstruction from projections," *IEEE Tr. Im. Proc.*, vol. 5, pp. 792–3, May 1996.

[104] C. L. Byrne, "Convergent block-iterative algorithms for image reconstruction from inconsistent data," *IEEE Tr. Im. Proc.*, vol. 6, pp. 1296–1304, Sept. 1997.

[105] C. L. Byrne, "Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods," *IEEE Tr. Im. Proc.*, vol. 7, pp. 100–9, Jan. 1998.

[106] J. Nuyts, B. D. Man, P. Dupont, M. Defrise, P. Suetens, and L. Mortelmans, "Iterative reconstruction for helical CT: A simulation study," *Phys. Med. Biol.*, vol. 43, pp. 729–37, Apr. 1998.

[107] C. Kamphius and F. J. Beekman, "Accelerated iterative transmission CT reconstruction using an ordered subsets convex algorithm," *IEEE Tr. Med. Im.*, vol. 17, pp. 1001–5, Dec. 1998.

[108] H. Erdoğan, G. Gualtieri, and J. A. Fessler, "An ordered subsets algorithm for transmission tomography," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 1998. Inadvertently omitted from proceedings. Available from web page.

[109] H. Erdoğan and J. A. Fessler, "Ordered subsets algorithms for transmission tomography," *Phys. Med. Biol.*, vol. 44, pp. 2835–51, Nov. 1999.

[110] T. Hebert and R. Leahy, "A Bayesian reconstruction algorithm for emission tomography using a Markov random field prior," in *Proc. SPIE 1092, Med. Im. III: Im. Proc.*, pp. 458–4662, 1989.

[111] T. Hebert and R. Leahy, "A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors," *IEEE Tr. Med. Im.*, vol. 8, pp. 194–202, June 1989.

[112] T. J. Hebert and R. Leahy, "Statistic-based MAP image reconstruction from Poisson data using Gibbs priors," *IEEE Tr. Sig. Proc.*, vol. 40, pp. 2290–2303, Sept. 1992.

[113] G. Gullberg and B. M. W. Tsui, "Maximum entropy reconstruction with constraints: iterative algorithms for solving the primal and dual programs," in *Proc. Tenth Intl. Conf. on Information Processing in Medical Im.* (C. N. de Graaf and M. A. Viergever, eds.), pp. 181–200, New York: Plenum Press, 1987.

[114] C. A. Bouman, K. Sauer, and S. S. Saquib, "Tractable models and efficient algorithms for Bayesian tomography," in *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, vol. 5, pp. 2907–10, 1995.

[115] S. Saquib, J. Zheng, C. A. Bouman, and K. D. Sauer, "Provably convergent coordinate descent in statistical tomographic reconstruction," in *Proc. IEEE Intl. Conf. on Image Processing*, vol. 2, pp. 741–4, 1996.

[116] J. Zheng, S. Saquib, K. Sauer, and C. Bouman, "Functional substitution methods in optimization for Bayesian tomography," *IEEE Tr. Im. Proc.*, Mar. 1997. IEEE Tr. Image Proc.

[117] J. A. Fessler, "Grouped coordinate descent algorithms for robust edge-preserving image restoration," in *Proc. SPIE 3071, Im. Recon. and Restor. II*, pp. 184–94, 1997.

[118] P. J. Huber, *Robust statistics*. New York: Wiley, 1981.

[119] H. M. Hudson, J. Ma, and P. Green, "Fisher's method of scoring in statistical image reconstruction: comparison of Jacobi and Gauss-Seidel iterative schemes," *Stat. Meth. Med. Res.*, vol. 3, no. 1, pp. 41–61, 1994.

[120] S. L. Hillis and C. S. Davis, "A simple justification of the iterative fitting procedure for generalized linear models," *American Statistician*, vol. 48, pp. 288–289, Nov. 1994.

[121] H. Erdoğan and J. A. Fessler, "Accelerated monotonic algorithms for transmission tomography," in *Proc. IEEE Intl. Conf. on Image Processing*, vol. 2, pp. 680–4, 1998.

[122] J. A. Fessler, E. P. Ficaro, N. H. Clinthorne, and K. Lange, "Fast parallelizable algorithms for transmission image reconstruction," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 3, pp. 1346–50, 1995.

[123] K. D. Sauer, S. Borman, and C. A. Bouman, "Parallel computation of sequential pixel updates in statistical tomographic reconstruction," in *Proc. IEEE Intl. Conf. on Image Processing*, vol. 3, pp. 93–6, 1995.

[124] S. T. Jensen, S. Johansen, and S. L. Lauritzen, "Globally convergent algorithms for maximizing a likelihood function," *Biometrika*, vol. 78, no. 4, pp. 867–77, 1991.

[125] G. H. Golub and C. F. Van Loan, *Matrix computations*. Johns Hopkins Univ. Press, 1989.

[126] E. Mumcuoglu, R. Leahy, and S. Cherry, "A statistical approach to transmission image reconstruction from ring source calibration measurements in PET," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 910–912, 1992.

[127] E. Ü. Mumcuoğlu and R. M. Leahy, "A gradient projection conjugate gradient algorithm for Bayesian PET reconstruction," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 3, pp. 1212–6, 1994.

[128] J. J. Moré and G. Toraldo, "On the solution of large quadratic programming problems with bound constraints," *SIAM J. Optim.*, vol. 1, pp. 93–113, Feb. 1991.

[129] M. Bierlaire, P. L. Toint, and D. Tuyttens, "On iterative algorithms for linear least squares problems with bound constraints," *Linear Algebra and its Applications*, vol. 143, pp. 111–43, 1991.

[130] K. Lange, "An adaptive barrier method for convex programming," *Methods and Appl. of Analysis*, vol. 1, no. 4, pp. 392–402, 1994.

[131] M. Al-Baali and R. Fletcher, "On the order of convergence of preconditioned nonlinear conjugate gradient methods," *SIAM J. Sci. Comp.*, vol. 17, pp. 658–65, May 1996.

[132] M. S. Kaplan, D. R. Haynor, and H. Vija, "A differential attenuation method for simultaneous estimation of SPECT activity and attenuation distributions," *IEEE Tr. Nuc. Sci.*, vol. 46, pp. 535–41, June 1999.

[133] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comp.*, vol. 16, pp. 1190–1208, 1995.

[134] Y. Censor, P. P. B. Eggermont, and D. Gordon, "Strong underrelaxation in Kaczmarz's method for inconsistent systems," *Numerische Mathematik*, vol. 41, pp. 83–92, 1983.

[135] Z. Liang and J. Ye, "Reconstruction of object-specific attenuation map for quantitative SPECT," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1231–1235, 1993.

[136] G. L. Zeng and G. T. Gullberg, "A MAP algorithm for transmission computed tomography," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1202–1204, 1993.

[137] J. M. M. Anderson, B. A. Mair, M. Rao, and C. H. Wu, "A weighted least-squares method for PET," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 2, pp. 1292–6, 1995.

[138] C. Bouman and K. Sauer, "Nonlinear multigrid methods of optimization in Bayesian tomographic image reconstruction," in *SPIE Neural and Stoch. Methods in Image and Signal Proc.*, 1992.

[139] C. Bouman and K. Sauer, "A generalized Gaussian image model for edge-preserving MAP estimation," *IEEE Tr. Im. Proc.*, vol. 2, pp. 296–310, July 1993.

[140] P. J. Green, "Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives," *J. Royal Stat. Soc. Ser. B*, vol. 46, no. 2, pp. 149–192, 1984.

[141] M. B. Dollinger and R. G. Staudte, "Influence functions of iteratively reweighted least squares estimators," *J. Am. Stat. Ass.*, vol. 86, pp. 709–716, Sept. 1991.

[142] J. Qi, R. M. Leahy, C. Hsu, T. H. Farquhar, and S. R. Cherry, "Fully 3D Bayesian image reconstruction for the ECAT EXACT HR+," *IEEE Tr. Nuc. Sci.*, vol. 45, pp. 1096–1103, June 1998.

[143] J. Qi, R. M. Leahy, S. R. Cherry, A. Chatziioannou, and T. H. Farquhar, "High resolution 3D Bayesian image reconstruction using the microPET small-animal scanner," *Phys. Med. Biol.*, vol. 43, pp. 1001–14, Apr. 1998.

[144] J. A. Browne and A. R. D. Pierro, "A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography," *IEEE Tr. Med. Im.*, vol. 15, pp. 687–99, Oct. 1996.

[145] A. M. Thompson, J. C. Brown, J. W. Kay, and D. M. Titterington, "A study of methods or choosing the smoothing parameter in image restoration by regularization," *IEEE Tr. Patt. Anal. Mach. Int.*, vol. 13, no. 4, pp. 326–339, 1991.

[146] J. W. Hilgers and W. R. Reynolds, "Instabilities in the optimal regularization parameter relating to image recovery problems," *J. Opt. Soc. Am. A*, vol. 9, pp. 1273–1279, Aug. 1992.

[147] Y. Pawitan and F. O'Sullivan, "Data-dependent bandwidth selection for emission computed tomography reconstruction," *IEEE Tr. Med. Im.*, vol. 12, pp. 167–172, June 1993.

[148] F. O'Sullivan and Y. Pawitan, "Bandwidth selection for indirect density estimation based on corrupted histogram data," *J. Am. Stat. Ass.*, vol. 91, pp. 610–26, June 1996.

[149] C. R. Vogel, "Non-convergence of the L-curve regularization parameter selection method," *Inverse Prob.*, vol. 12, pp. 535–47, Aug. 1996.

[150] P. P. B. Eggermont and V. N. LaRiccia, "Nonlinearly smoothed EM density estimation with automated smoothing parameter selection for nonparametric deconvolution problems," *J. Am. Stat. Ass.*, vol. 92, pp. 1451–8, Dec. 1997.

[151] D. M. Higdon, J. E. Bowsher, V. E. Johnson, T. G. Turkington, D. R. Gilland, and R. J. Jaszczak, "Fully Bayesian estimation of Gibbs hyperparameters for emission computed tomography data," *IEEE Tr. Med. Im.*, vol. 16, p. 516, Oct. 1997.

[152] G. Sebastiani and F. Godtliebsen, "On the use of Gibbs priors for Bayesian image restoration," *Signal Processing*, vol. 56, pp. 111–18, Jan. 1997.

[153] Z. Zhou, R. M. Leahy, and J. Qi, "Approximate maximum likelihood hyperparameter estimation for Gibbs priors," *IEEE Tr. Im. Proc.*, vol. 6, pp. 844–61, June 1997.

[154] S. S. Saquib, C. A. Bouman, and K. Sauer, "ML parameter estimation for Markov random fields, with applications to Bayesian tomography," *IEEE Tr. Im. Proc.*, vol. 7, pp. 1029–44, July 1998.

[155] M. Nikolova, J. Idier, and A. Mohammad-Djafari, "Inversion of large-support ill-posed linear operators using a piecewise Gaussian MRF," *IEEE Tr. Im. Proc.*, vol. 7, pp. 571–85, Apr. 1998.

[156] K. Sauer and C. Bouman, "Bayesian estimation of transmission tomograms using local optimization operations," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 3, pp. 2089–2093, 1991.

[157] K. Sauer and C. Bouman, "Bayesian estimation of transmission tomograms using segmentation based optimization," *IEEE Tr. Nuc. Sci.*, vol. 39, pp. 1144–1152, Aug. 1992.

[158] R. Mifflin, D. Sun, and L. Qi, "Quasi-Newton bundle-type methods for nondifferentiable convex optimization," *SIAM J. Optim.*, vol. 8, pp. 583–603, May 1998.

[159] Y. Censor, D. E. Gustafson, A. Lent, and H. Tuy, "A new approach to the emission computerized tomography problem: simultaneous calculation of attenuation and activity coefficients," *IEEE Tr. Nuc. Sci.*, vol. 26, Jan. 1979.

[160] A. Welch, R. Clack, F. Natterer, and G. T. Gullberg, "Toward accurate attenuation correction in SPECT without transmission measurements," *IEEE Tr. Med. Im.*, vol. 16, p. 532, Oct. 1997.

[161] R. E. Alvarez and A. Macovski, "Energy-selective reconstructions in X-ray computed tomography," *Phys. Med. Biol.*, vol. 21, pp. 733–44, 1976.

[162] N. H. Clinthorne, "A constrained dual-energy reconstruction method for material-selective transmission tomography," *Nucl. Instr. Meth. Phys. Res. A.*, vol. 352, pp. 347–8, Dec. 1994.

[163] P. Sukovic and N. H. Clinthorne, "Penalized weighted least-squares image reconstruction in single and dual energy X-ray computed tomography," *IEEE Tr. Med. Im.*, 1999. Submitted.

[164] J. A. Fessler, D. F. Yu, and E. P. Ficaro, "Maximum likelihood transmission image reconstruction for overlapping transmission beams," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 1999.

[165] D. F. Yu, J. A. Fessler, and E. P. Ficaro, "Maximum likelihood transmission image reconstruction for overlapping transmission beams," *IEEE Tr. Med. Im.*, 1999. Submitted.

[166] J. E. Bowsher, M. P. Tornai, D. R. Gilland, D. E. G. Trotter, and R. J. Jaszczak, "An EM algorithm for modeling multiple or extended TCT sources," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 1999.

[167] A. Krol, J. E. Bowsher, S. H. Manglos, D. H. Feiglin, and F. D. Thomas, "An EM algorithm for estimating SPECT emission and transmission parameters from emission data only," *Phys. Med. Biol.*, 1999. Submitted.

[168] D. J. Rossi and A. S. Willsky, "Reconstruction from projections based on detection and estimation of objects—Parts i & II: Performance analysis and robustness analysis," *IEEE Tr. Acoust. Sp. Sig. Proc.*, vol. 32, pp. 886–906, Aug. 1984.

[169] D. J. Rossi, A. S. Willsky, and D. M. Spielman, "Object shape estimation from tomographic measurements—a performance evaluation," *Signal Processing*, vol. 18, pp. 63–88, Sept. 1989.

[170] Y. Bresler, J. A. Fessler, and A. Macovski, "Model based estimation techniques for 3-D reconstruction from projections," *Machine Vision and Applications*, vol. 1, no. 2, pp. 115–26, 1988.

[171] Y. Bresler, J. A. Fessler, and A. Macovski, "A Bayesian approach to reconstruction from incomplete projections of a multiple object 3-D domain," *IEEE Tr. Patt. Anal. Mach. Int.*, vol. 11, pp. 840–58, Aug. 1989.

[172] J. A. Fessler and A. Macovski, "Object-based 3-D reconstruction of arterial trees from magnetic resonance angiograms," *IEEE Tr. Med. Im.*, vol. 10, pp. 25–39, Mar. 1991.

[173] S. P. Müller, M. F. Kijewski, S. C. Moore, and B. L. Holman, "Maximum-likelihood estimation: a mathematical model for quantitation in nuclear medicine," *J. Nuc. Med.*, vol. 31, pp. 1693–1701, Oct. 1990.

[174] C. K. Abbey, E. Clarkson, H. H. Barrett, S. P. Müller, and F. J. Rybicki, "A method for approximating the density of maximum likelihood and maximum a posteriori estimates under a Gaussian noise model," *Med. Im. Anal.*, vol. 2, no. 4, pp. 395–403, 1998.

[175] P. C. Chiao, W. L. Rogers, N. H. Clinthorne, J. A. Fessler, and A. O. Hero, "Model-based estimation for dynamic cardiac studies using ECT," *IEEE Tr. Med. Im.*, vol. 13, pp. 217–26, June 1994.

[176] Y. Amit and K. Manbeck, "Deformable template models for emission tomography," *IEEE Tr. Med. Im.*, vol. 12, pp. 260–268, June 1993.

[177] K. M. Hanson, "Bayesian reconstruction based on flexible priors," *J. Opt. Soc. Am. A*, vol. 10, pp. 997–1004, May 1993.

[178] X. L. Battle, G. S. Cunningham, and K. M. Hanson, "Tomographic reconstruction using 3D deformable models," *Phys. Med. Biol.*, vol. 43, pp. 983–90, 1998.

[179] G. S. Cunningham, K. M. Hanson, and X. L. Battle, "Three-dimensional reconstructions from low-count SPECT data using deformable models," *Optics Express*, vol. 2, pp. 227–36, 1998. http://www.osa.org.

[180] X. L. Battle and Y. Bizais, "3D attenuation map reconstruction using geometrical models and free form deformations," in *Proc. of the 1999 Intl. Mtg. on Fully 3D Im. Recon. in Rad. Nuc. Med.*, pp. 181–184, 1999.

[181] H. Erdoğan, *Statistical image reconstruction algorithms using paraboloidal surrogates for PET transmission scans*. Ph.D. thesis, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Ann Arbor, MI., July 1999.