Iterative Methods for Image Reconstruction

Jeffrey A. Fessler

EECS Department
The University of Michigan

ISBI Tutorial

Apr. 6, 2006

These annotated slides were prepared by Jeff Fessler for attendees of the ISBI tutorial on statistical image reconstruction methods.

The purpose of the annotation is to provide supplemental details, and particularly to provide extensive literature references for further study.

For a fascinating history of tomography, see [1]. For broad coverage of image science, see [2].

For further references on image reconstruction, see review papers and chapters, *e.g.*, [3–9].

Image Reconstruction Methods

(Simplified View)

Analytical	Iterative
(FBP)	(OSEM?)
(MR: iFFT)	(MR: CG?)

Image Reconstruction Methods / Algorithms

ANALYTICAL

- FBP
- BPF
- Gridding
- ...

ITERATIVE

- **Algebraic** (y = Ax)
  - ART
  - MART
  - SMART
  - ...
- **Statistical**
  - **(Weighted) Least Squares**
    - CG
    - CD
    - ISRA
    - ...
  - **Likelihood (e.g., Poisson)**
    - EM (etc.)
    - OSEM
    - SAGE
    - CG
    - Int. Point
    - GCA
    - PSCD
    - FSCD ...

Part of the goal is to bring order to this alphabet soup.

Outline of Part I

Part 0: Introduction / Overview / Examples

Part 1: Problem Statements
- Continuous-discrete vs continuous-continuous vs discrete-discrete

Part 2: Four of Five Choices for Statistical Image Reconstruction
- Object parameterization
- System physical modeling
- Statistical modeling of measurements
- Cost functions and regularization

Part 3: Fifth Choice: Iterative algorithms
- Classical optimization methods
- Considerations: nonnegativity, convergence rate, ...
- Optimization transfer: EM etc.
- Ordered subsets / block iterative / incremental gradient methods

Part 4: Performance Analysis
- Spatial resolution properties
- Noise properties
- Detection performance

Emphasis on general principles rather than specific empirical results.

The journals (and conferences like NSS/MIC!) are replete with empirical comparisons.

Although the focus of examples in this course are PET / SPECT / CT, most of the principles apply equally well to other tomography problems like MR image reconstruction, optical / diffraction tomography, etc.

History

- Successive substitution method vs direct Fourier (Bracewell, 1956)
- Iterative method for X-ray CT (Hounsfield, 1968)
- ART for tomography (Gordon, Bender, Herman, JTB, 1970)
- Richardson/Lucy iteration for image restoration (1972, 1974)
- Weighted least squares for 3D SPECT (Goitein, NIM, 1972)
- Proposals to use Poisson likelihood for emission and transmission tomography
  Emission: (Rockmore and Macovski, TNS, 1976)
  Transmission: (Rockmore and Macovski, TNS, 1977)
- Expectation-maximization (EM) algorithms for Poisson model
  Emission: (Shepp and Vardi, TMI, 1982)
  Transmission: (Lange and Carson, JCAT, 1984)
- Regularized (aka Bayesian) Poisson emission reconstruction
  (Geman and McClure, ASA, 1985)
- Ordered-subsets EM algorithm
  (Hudson and Larkin, TMI, 1994)
- Commercial introduction of OSEM for PET scanners circa 1997

Bracewell's classic paper on direct Fourier reconstruction also mentions a successive substitution approach [10]
X-ray CT patent: [11]
Early iterative methods for SPECT by Muehllehner [12] and Kuhl [13].
ART: [14–17]
Richardson/Lucy iteration for image restoration was not derived from ML considerations, but turns out to be the familiar ML-EM iteration [18, 19]
Emission: [20]
Transmission: [21]
General expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) [22]
Emission EM algorithm: [23]
Transmission EM algorithm: [24]
Bayesian method for Poisson emission problem: [25]
OSEM [26]

Prior to the proposals for Poisson likelihood models, the Lawrence Berkeley Laboratory had proposed and investigated weighted least-squares (WLS) methods for SPECT (in 3D!) using iterative algorithms; see (Goitein, 1972) [27] and (Budinger and Gullberg, 1974) [28]. These methods became widely available in 1977 through the release of the Donner RECLBL package [29].

Of course there was lots of work ongoing based on "algebraic" reconstruction methods in the 1970s and before. But until WLS methods were proposed, this work was largely not "statistical."

Why Statistical Methods?

- Object constraints (*e.g.*, nonnegativity, object support)
- Accurate physical models (less bias $\Longrightarrow$ improved quantitative accuracy)
  (*e.g.*, nonuniform attenuation in SPECT)
  improved spatial resolution?
- Appropriate statistical models (less variance $\Longrightarrow$ lower image noise)
  (FBP treats all rays equally)
- Side information (*e.g.*, MRI or CT boundaries)
- Nonstandard geometries (*e.g.*, irregular sampling or "missing" data)

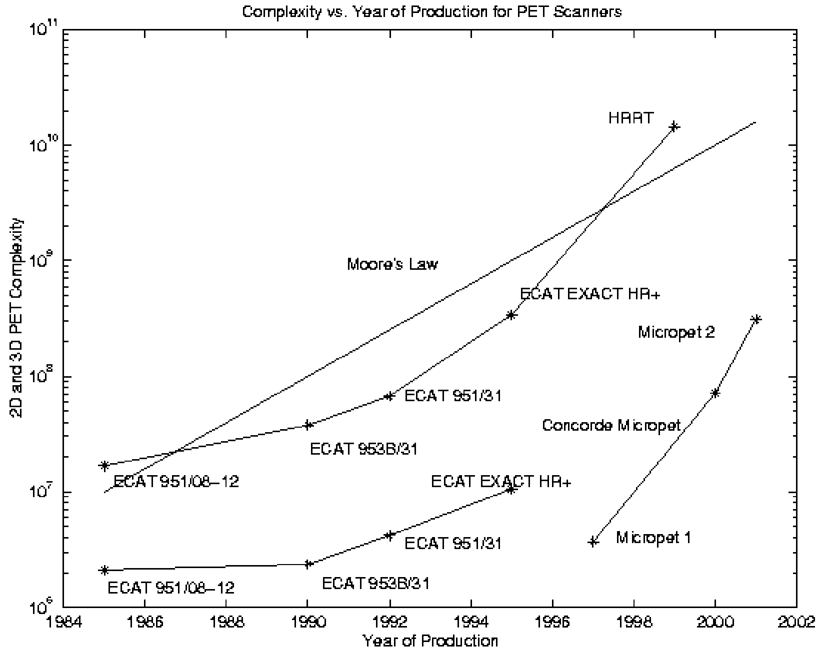
Disadvantages?
- Computation time
- Model complexity
- Software complexity

Analytical methods (a different short course!)
- Idealized mathematical model
  - Usually geometry only, greatly over-simplified physics
  - Continuum measurements (discretize/sample *after* solving)
- No statistical model
- Easier analysis of properties (due to linearity)
  e.g., Huesman (1984) FBP ROI variance for kinetic fitting

There is a continuum of physical system models that tradeoff accuracy and compute time. The "right" way to model the physics is usually too complicated, so one uses approximations. The sensitivity of statistical methods to those approximations needs more investigation.

FBP has its faults, but its properties (good and bad) are very well understood and hence predictable, due to its linearity. Spatial resolution, variance, ROI covariance (Huesman [30]), and autocorrelation have all been thoroughly analyzed (and empirical results agree with the analytical predictions). Only recently have such analyses been provided for *some* nonlinear reconstruction methods *e.g.*, [31–42].

What about Moore's Law?



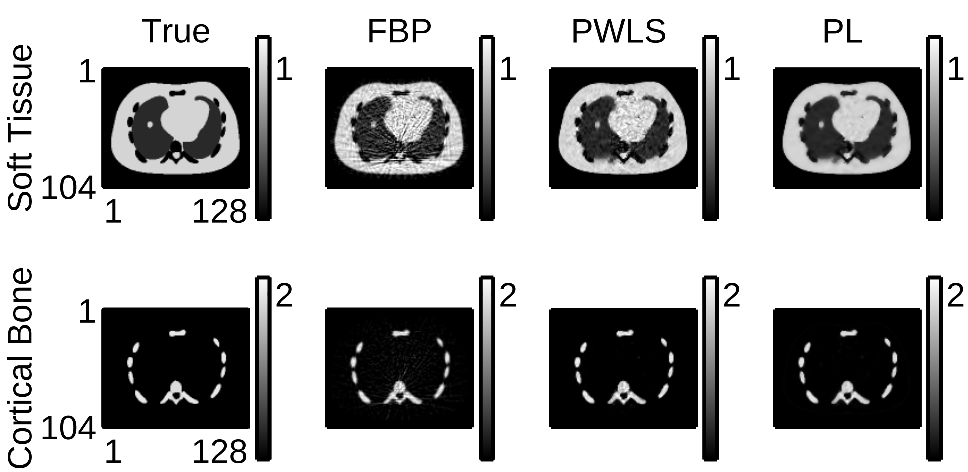
In this graph complexity is the number of lines of response (number of rays) acquired. The ECAT scanners can operate either in 2D mode (with septa in place) or 3D mode (with septa retracted) so those scanners have two points each.

I got this graph from Richard Leahy; it was made by Evren Asma. Only CTI scanners and their relatives are represented. Another such graph appeared in [43].

There is considerable ongoing effort to reduce or minimize the compute time by more efficient algorithms.

Moore's law for computing power increases will not alone solve all of the compute problems in image reconstruction. The problems increase in difficulty at nearly the same rate as the increase in compute power. (Consider the increased amount of data in 3D PET scanners relative to 2D.) (Or even the increased number of slices in 2D mode.) Or spiral CT, or fast dynamic MRI,... Therefore there is a need for further improvements in algorithms in addition to computer hardware advances.

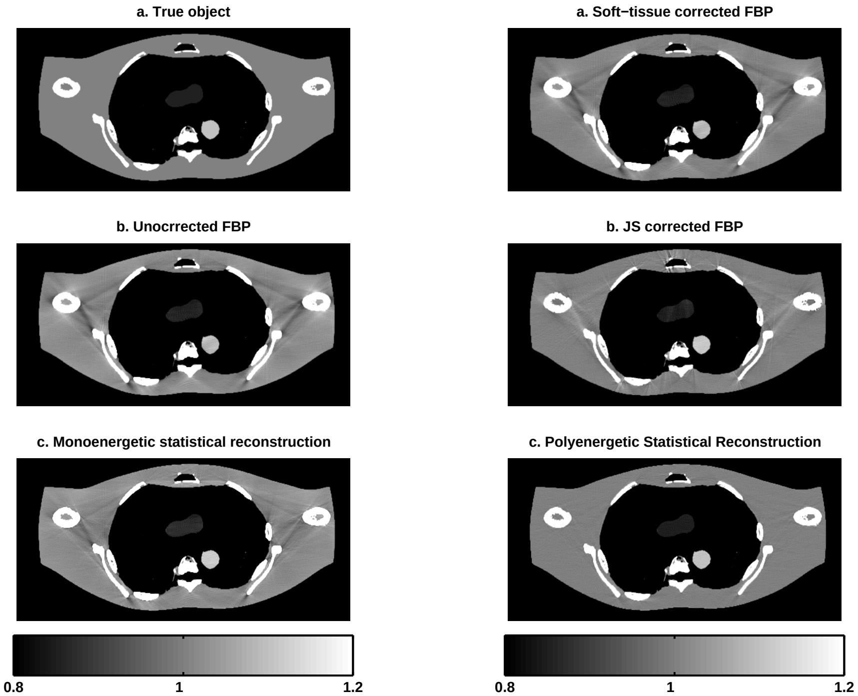
Benefit Example: Statistical Models



	NRMS Error	
Method	Soft Tissue	Cortical Bone
FBP	22.7%	29.6%
PWLS	13.6%	16.2%
PL	11.8%	15.8%

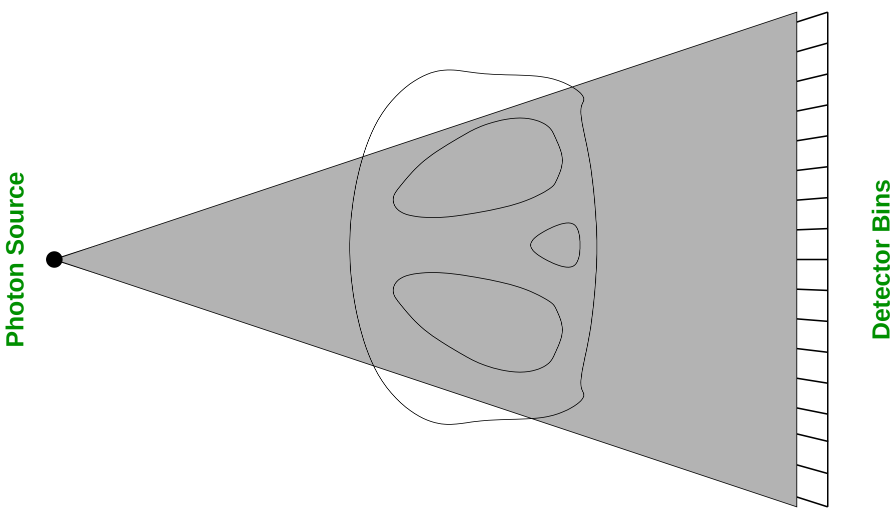
Conventional FBP reconstruction of dual-energy X-ray CT data does not account for the noise properties of CT measurements and results in significant noise propagation into the soft tissue and cortical bone component images. Statistical reconstruction methods greatly reduces this noise, improving quantitative accuracy [44]. This is of potential importance for applications like bone density measurements.

Benefit Example: Physical Models



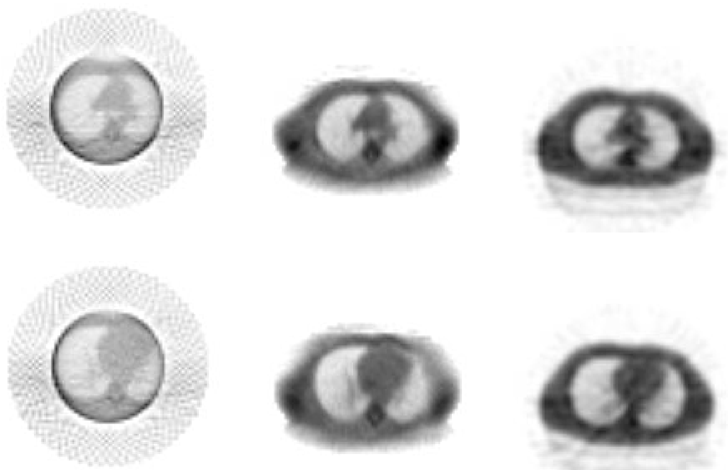
Conventional FBP ignores the polyenergetic X-ray source spectrum. Statistical/iterative reconstruction methods can build that spectrum into the model and nearly eliminate beam-hardening artifacts [45–47].

Benefit Example: Nonstandard Geometries



A SPECT transmission scan with 65cm distance between line source and standard Anger camera provides partially truncated sinogram views of most patients.

Truncated Fan-Beam SPECT Transmission Scan



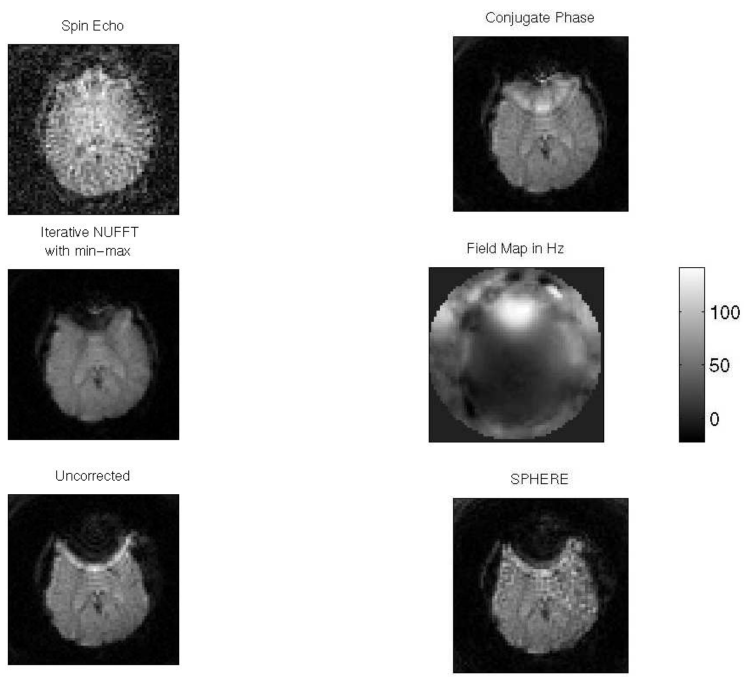
Truncated FBP | Truncated PWLS | Untruncated FBP

The FBP reconstruction method is largely ruined by the sinogram truncation.

Despite the partial truncation, each pixel is *partly* sampled by “line integrals” at some range of angles. With the benefit of spatial regularization, nonnegativity constraints, and statistical models, a statistical reconstruction method (PWLS in this case) can recover an attenuation map that is comparable to that obtained with an untruncated scan.

We have shown related benefits in PET with missing sinogram data due to detector gaps [48].

One Final Advertisement: Iterative MR Reconstruction



MR signal equation:

$$s(t) = \int f(\vec{\mathrm{x}}) \exp(-\imath\omega(\vec{\mathrm{x}})t) \exp\left(-\imath 2\pi \vec{k}(\vec{\mathrm{x}}) \cdot \vec{\mathrm{x}}\right) \mathrm{d}\vec{\mathrm{x}}$$

- Due to field inhomogeneity, signal is *not* Fourier transform of object.
- Measure off-resonance field-map $\omega(\vec{\mathrm{x}})$ using two displaced echos
- Penalized WLS cost function minimized by conjugate gradient
- System matrix $\boldsymbol{A}$ includes off-resonance effects
- Fast algorithm using NUFFT and time-segmentation

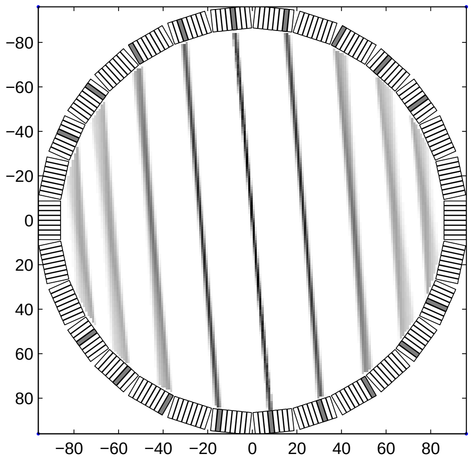
[49–51]

Hopefully that is enough motivation, so, on with the methodology!

Part 1: Problem Statement(s)

Example: in PET, the goal is to reconstruct radiotracer distribution $\lambda(\vec{x})$ from photon pair coincidence measurements $\{y_i\}_{i=1}^{n_d}$, given the detector sensitivity patterns $s_i(\vec{x})$, $i = 1, \ldots, n_d$, for each "line of response."

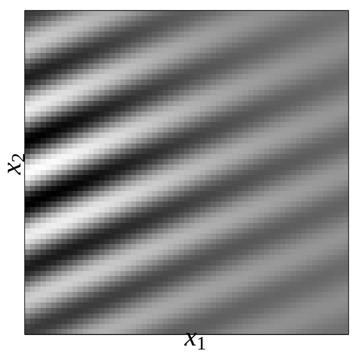
Statistical model: $y_i \sim \text{Poisson}\left\{ \int \lambda(\vec{x})\, s_i(\vec{x})\, d\vec{x} + r_i \right\}$



This part is much abbreviated from a short course I have given at NSS-MIC in which the PET / SPECT problem statements are described in detail. See my web site for the course notes if interested.

These sensitivity patterns account for the parallax and crystal penetration effects in ring PET systems.

Example: MRI "Sensitivity Pattern"



Each "k-space sample" involves the transverse magnetization $f(\vec{x})$ weighted by:

- sinusoidal (complex exponential) pattern corresponding to k-space location $\vec{k}$
- RF receive coil sensitivity pattern
- phase effects of field inhomogeneity
- spin relaxation effects.

$$y_i = \int f(\vec{x})\, s_i(\vec{x})\, d\vec{x} + \varepsilon_i, \quad i = 1, \ldots, n_d, \qquad s_i(\vec{x}) = c_{\mathrm{RF}}(\vec{x})\, e^{-\imath \omega(\vec{x}) t_i}\, e^{-t_i / T_2(\vec{x})}\, e^{-\imath 2\pi \vec{k}(t_i) \cdot \vec{x}}$$

Continuous-Discrete Models

Emission tomography: $y_i \sim \mathsf{Poisson}\{\int \lambda(\vec{x})\, s_i(\vec{x})\, d\vec{x} + r_i\}$

Transmission tomography (monoenergetic): $y_i \sim \mathsf{Poisson}\left\{b_i \exp\left(-\int_{\mathcal{L}_i} \mu(\vec{x})\, d\ell\right) + r_i\right\}$

Transmission (polyenergetic): $y_i \sim \mathsf{Poisson}\left\{\int I_i(\mathcal{E}) \exp\left(-\int_{\mathcal{L}_i} \mu(\vec{x}, \mathcal{E})\, d\ell\right) d\mathcal{E} + r_i\right\}$

Magnetic resonance imaging: $y_i = \int f(\vec{x})\, s_i(\vec{x})\, d\vec{x} + \varepsilon_i$

Discrete measurements $\boldsymbol{y} = (y_1, \ldots, y_{n_d})$
Continuous-space unknowns: $\lambda(\vec{x})$, $\mu(\vec{x})$, $f(\vec{x})$
Goal: estimate $f(\vec{x})$ given $\boldsymbol{y}$

Solution options:

- Continuous-continuous formulations (“analytical”)
- Continuous-discrete formulations
  usually $\hat{f}(\vec{x}) = \sum_{i=1}^{n_d} c_i\, s_i(\vec{x})$
- Discrete-discrete formulations $f(\vec{x}) \approx \sum_{j=1}^{n_p} x_j\, b_j(\vec{x})$

For a nice comparison of the options, see [9].

Part 2: Five Categories of Choices

- Object parameterization: function $f(\vec{r})$ vs finite coefficient vector $\boldsymbol{x}$
- System physical model: $\{s_i(\vec{r})\}$
- Measurement statistical model $y_i \sim \boxed{?}$
- Cost function: data-mismatch and regularization
- Algorithm / initialization

No perfect choices - one can critique all approaches!

Often these choices are made implicitly rather than explicitly. Leaving the choices implicit fortifies the common belief among non-experts that there are basically two kinds of reconstruction algorithms, FBP and “iterative.”

In fact, the choices one makes in the above five categories can affect the results significantly.

In my opinion, every paper describing iterative image reconstruction methods (or results thereof) should make as explicit as possible what choices were made in each of the above categories.

Choice 1. Object Parameterization

Finite measurements: $\{y_i\}_{i=1}^{n_d}$. Continuous object: $f(\vec{r})$. Hopeless?

"All models are wrong but some models are useful."

Linear *series expansion* approach. Replace $f(\vec{r})$ by $\boldsymbol{x} = (x_1, \ldots, x_{n_p})$ where

$$f(\vec{r}) \approx \tilde{f}(\vec{r}) = \sum_{j=1}^{n_p} x_j \, b_j(\vec{r}) \leftarrow \text{"basis functions"}$$

Forward projection:

$$\int s_i(\vec{r}) \, f(\vec{r}) \, d\vec{r} = \int s_i(\vec{r}) \left[\sum_{j=1}^{n_p} x_j \, b_j(\vec{r}) \right] d\vec{r} = \sum_{j=1}^{n_p} \left[\int s_i(\vec{r}) \, b_j(\vec{r}) \, d\vec{r} \right] x_j$$

$$= \sum_{j=1}^{n_p} a_{ij} x_j = [\boldsymbol{Ax}]_i, \text{ where } a_{ij} \triangleq \int s_i(\vec{r}) \, b_j(\vec{r}) \, d\vec{r}$$

- Projection integrals become finite summations.
- a_{ij} is contribution of jth basis function (*e.g.*, voxel) to ith measurement.
- The units of a_{ij} and x_j depend on the user-selected units of $b_j(\vec{r})$.
- The $n_d \times n_p$ matrix $\boldsymbol{A} = \{a_{ij}\}$ is called the *system matrix*.

In principle it is not entirely hopeless to reconstruction a continuous $f(\vec{r})$ from a finite set of measurements. This is done routinely in the field of nonparametric regression [52] (the generalization of linear regression that allows for fitting smooth functions rather than just lines). But it is complicated in tomography...

Van De Walle, Barrett, *et al.* [53] have proposed pseudoinverse calculation method for MRI reconstruction from a continuous-object / discrete-data formulation, based on the general principles of Bertero *et al.* [54]. If the pseudo-inverse could truly be computed once-and-for-all then such an approach could be practically appealing. However, in practice there are object-dependent effects, such as nonuniform attenuation in SPECT and magnetic field inhomogeneity in MRI, and these preclude precomputation of the required SVDs. So pseudo-inverse approaches are impractical for typical realistic physical models.

(Linear) Basis Function Choices

- Fourier series (complex / not sparse)
- Circular harmonics (complex / not sparse)
- Wavelets (negative values / not sparse)
- Kaiser-Bessel window functions (blobs)
- Overlapping circles (disks) or spheres (balls)
- Polar grids, logarithmic polar grids
- "Natural pixels" $\{s_i(\vec{r})\}$
- B-splines (pyramids)
- Rectangular pixels / voxels (rect functions)
- Point masses / bed-of-nails / lattice of points / "comb" function
- Organ-based voxels (*e.g.*, from CT in PET/CT systems)
- ...

See [55] for an early discussion.

Many published "projector / backprojector pairs" are not based explicitly on any particular choice of basis.

Some pixel-driven backprojectors could be interpreted implicitly as point-mass object models. This model works fine for FBP, but causes artifacts for iterative methods.

Mazur *et al.* [56] approximate the shadow of each pixel by a rect function, instead of by a trapezoid. "As the shapes of pixels are artifacts of our digitisation of continuous real-world images, consideration of alternative orientation or shapes for them seems reasonable." However, they observe slightly worse results that worsen with iteration!

Classic series-expansion reference [57]

Organ-based voxel references include [58–63]

Basis Function Considerations

Mathematical

- Represent $f(\vec{r})$ "well" with moderate n_p (approximation accuracy)
- *e.g.*, represent a constant (uniform) function
- Orthogonality? (not essential)
- Linear independence (ensures uniqueness of expansion)
- Insensitivity to shift of basis-function grid (approximate shift invariance)
- Rotation invariance

Computational

- "Easy" to compute a_{ij} values and/or $\boldsymbol{Ax}$
- If stored, the system matrix $\boldsymbol{A}$ should be sparse (mostly zeros).
- Easy to represent nonnegative functions *e.g.*, if $x_j \geq 0$, then $f(\vec{r}) \geq 0$.
  A sufficient condition is $b_j(\vec{r}) \geq 0$.

"Well" $\equiv$ approximation error less than estimation error

Many bases have the desirable approximation property that one can form arbitrarily accurate approximations to $f(\vec{r})$ by taking n_p sufficiently large. (This is related to *completeness*.) Exceptions include "natural pixels" (a finite set) and the point-lattice "basis" (usually).

Nonlinear Object Parameterizations

Estimation of intensity *and* shape (*e.g.*, location, radius, etc.)

Surface-based (homogeneous) models

- Circles / spheres
- Ellipses / ellipsoids
- Superquadrics
- Polygons
- Bi-quadratic triangular Bezier patches, ...

Other models

- Generalized series $f(\vec{r}) = \sum_j x_j b_j(\vec{r}, \boldsymbol{\theta})$
- Deformable templates $f(\vec{r}) = b(T_{\boldsymbol{\theta}}(\vec{r}))$
- ...

Considerations

- Can be considerably more parsimonious
- If correct, yield greatly reduced estimation error
- Particularly compelling in limited-data problems
- Often oversimplified (all models are wrong but...)
- Nonlinear dependence on location induces non-convex cost functions, complicating optimization

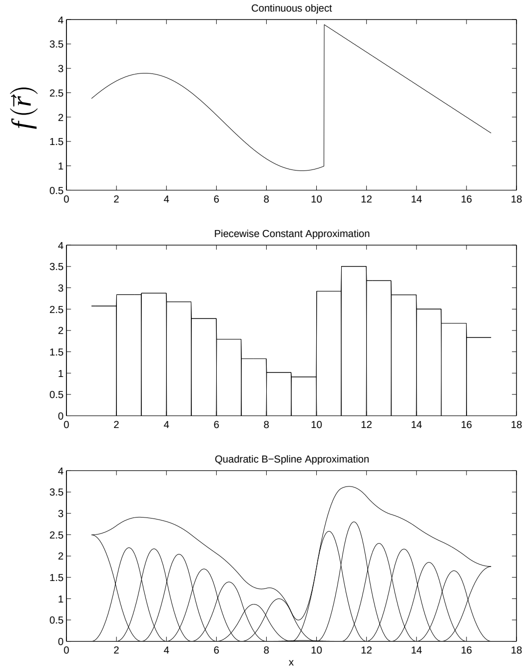
Disks [64, 65]

Polygons [66]

Generalized series [67]

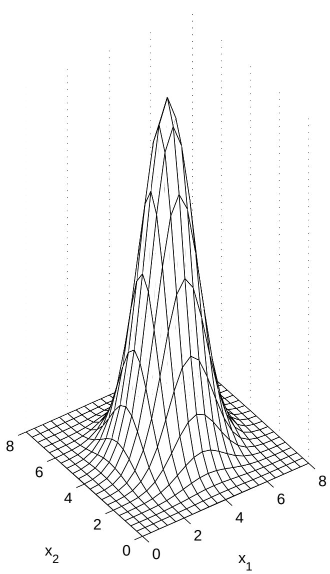
Bi-quadratic triangular Bezier patches [68]

Example Basis Functions - 1D

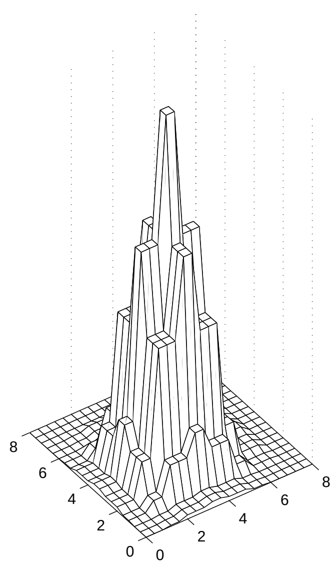


In the above example, neither the pixels nor the blobs are ideal, though both could reduce the average approximation error as low as needed by taking n_{p} sufficiently large.

Pixel Basis Functions - 2D



Continuous image $f(\vec{r})$



Pixel basis approximation $\sum_{j=1}^{n_{\mathrm{p}}} x_j \, b_j(\vec{r})$

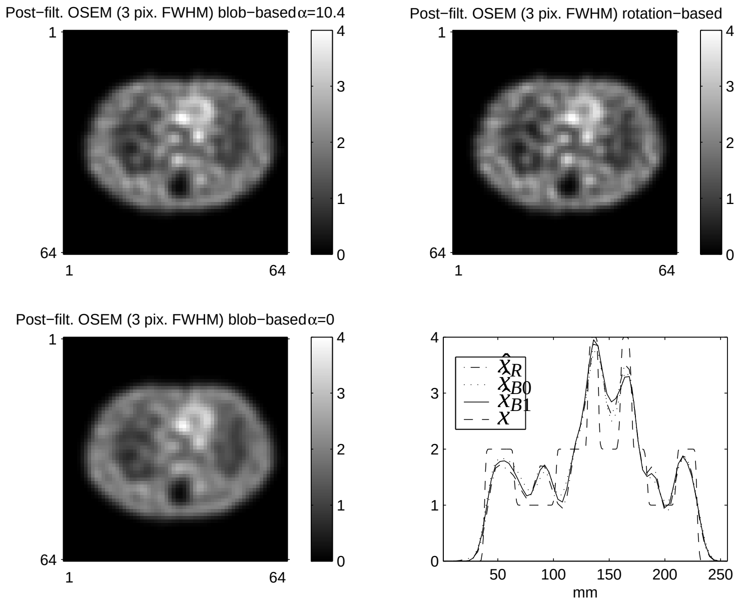
My tentative recommendation: use pixel / voxel basis.

- Simple
- Perfectly matched to digital displays
- Maximally sparse system matrix

Or use blobs (rotationally symmetric Kaiser-Bessel windows)

- Easy to compute projections “on the fly” due to rotational symmetry.
- Differentiable, nonnegative.
- Parsimony advantage using body-centered cubic packing

Blobs in SPECT: Qualitative



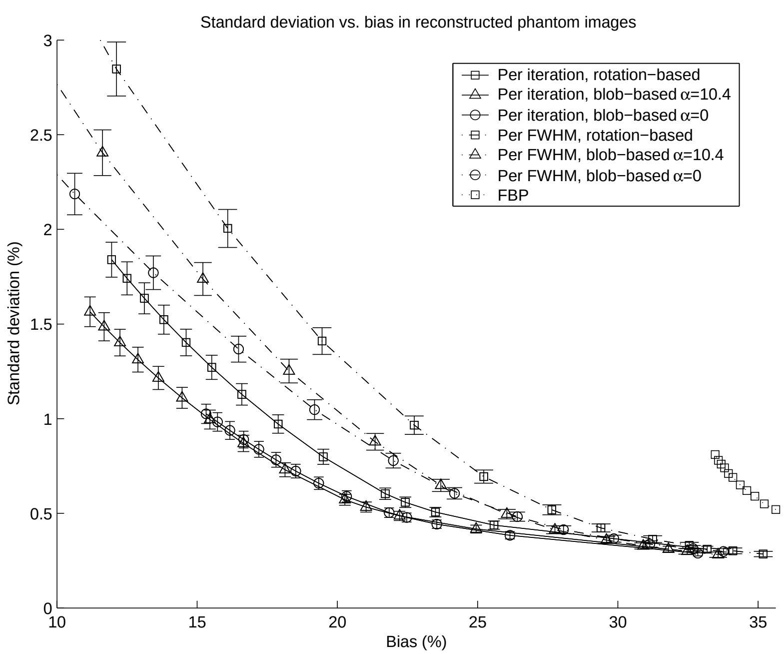
(2D SPECT thorax phantom simulations)

A slice and profiles through over-iterated and post-smoothed OSEM-reconstructed images of a single realization of noisy simulated phantom data. Superimposed on the profile of the true high-resolution phantom (x) are those of the images reconstructed with the rotation-based model ($\hat{x}_R$, NMSE = 4.12%), the blob-based model with $\alpha = 0$ ($\hat{x}_{B0}$, NMSE = 2.99%), and the blob-based model with $\alpha = 10.4$ ($\hat{x}_{B1}$, NMSE = 3.60%).

Figure taken from [69].

Blob expositions [70, 71].

Blobs in SPECT: Quantitative



Bottom line: in our experience in SPECT simulations comparing bias and variance of a small ROI, iterative reconstruction improved significantly over FBP, but blobs offered only a modest improvement over a rotation-based projector/backprojector that uses square pixels implicitly. And in some cases, a "blob" with shape parameter = 0, which is a (non-smooth) circ function performed best.

Discrete-Discrete Emission Reconstruction Problem

Having chosen a basis and *linearly* parameterized the emission density...

Estimate the emission density coefficient vector $\boldsymbol{x} = (x_1, \ldots, x_{n_p})$
(aka "image") using (something like) this statistical model:

$$y_i \sim \text{Poisson}\left\{ \sum_{j=1}^{n_p} a_{ij} x_j + r_i \right\}, \qquad i = 1, \ldots, n_d.$$

- $\{y_i\}_{i=1}^{n_d}$: observed counts from each detector unit
- $\boldsymbol{A} = \{a_{ij}\}$: system matrix (determined by system models)
- r_i's : background contributions (determined separately)

Many image reconstruction problems are "find $\boldsymbol{x}$ given $\boldsymbol{y}$" where

$$y_i = g_i([\boldsymbol{Ax}]_i) + \varepsilon_i, \qquad i = 1, \ldots, n_d.$$

Called the "discrete-discrete" estimation problem since both the measurement vector and the image vector are "discretized" (finite dimensional).

In contrast, FBP is derived from the "continuous-continuous" Radon transform model.

Choice 2. System Model, aka Physics

System matrix elements: $a_{ij} = \int s_i(\vec{r})\, b_j(\vec{r})\, d\vec{r}$

- scan geometry
- collimator/detector response
- attenuation
- scatter (object, collimator, scintillator)
- duty cycle (dwell time at each angle)
- detector efficiency / dead-time losses
- positron range, noncollinearity, crystal penetration, ...
- ...

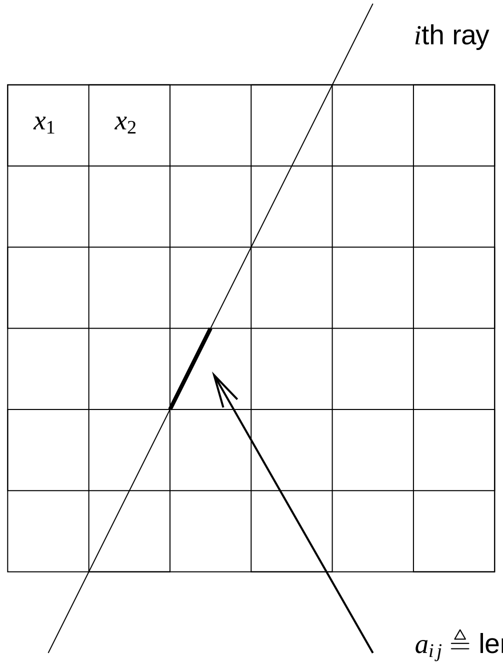
Considerations

- Improving system model can improve
  - Quantitative accuracy
  - Spatial resolution
  - Contrast, SNR, detectability
- Computation time (and storage vs compute-on-fly)
- Model uncertainties
  (*e.g.*, calculated scatter probabilities based on noisy attenuation map)
- Artifacts due to over-simplifications

For the pixel basis, a_{ij} is the probability that a decay in the jth pixel is recorded by the ith detector unit, or is proportional to that probability.

Attenuation enters into a_{ij} differently in PET and SPECT.

“Line Length” System Model for Tomography



Mathematically, the corresponding detector unit sensitivity pattern is

$$s_i(\vec{r}) = \delta\left(\vec{k}_i \cdot \vec{r} - \tau_i\right),$$

where δ denotes the Dirac impulse function.

This model is usually applied with the pixel basis, but can be applied to any basis.

Does not exactly preserve counts, *i.e.*, in general

$$\int f(\vec{r})\, \mathrm{d}\vec{r} \neq \sum_{i=1}^{n_\mathrm{d}} \sum_{j=1}^{n_\mathrm{p}} a_{ij} x_j$$

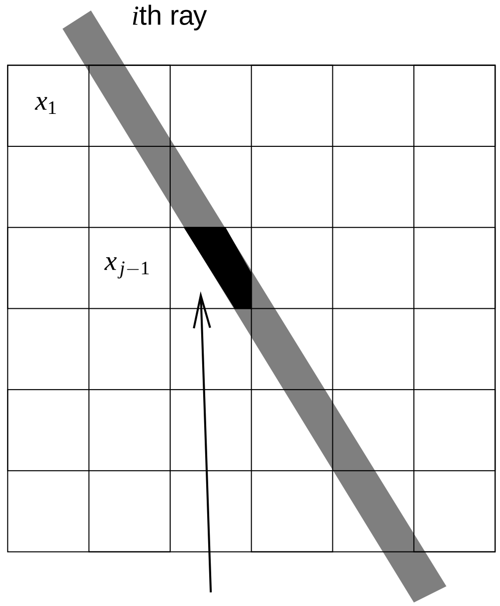
Leads to artifacts.

Units are wrong too. (Reconstructed $\boldsymbol{x}$ will have units inverse length.)

Perhaps reasonable for X-ray CT, but unnatural for emission tomography. (Line segment length is a probability?)

In short: I recommend using almost anything else!

“Strip Area” System Model for Tomography



Accounts for finite detector width.

Mathematically, the corresponding detector unit sensitivity pattern is

$$s_i(\vec{r}) = \operatorname{rect}\left(\frac{\vec{k}_i \cdot \vec{r} - \tau_i}{w}\right),$$

where w is the detector width.

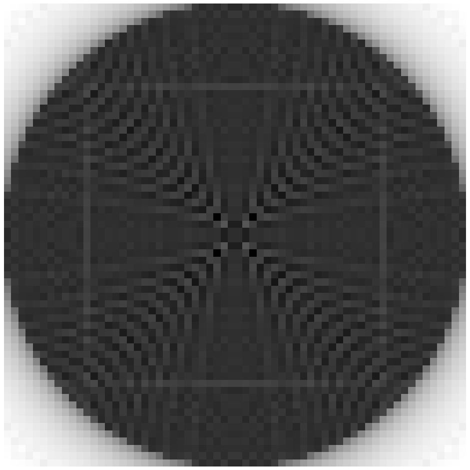
Can exactly preserve counts, since all areas are preserved, provided that the width w is an integer multiple of the center-to-center ray spacing.

Most easily applied to the pixel basis, but in principle applies to any choice.

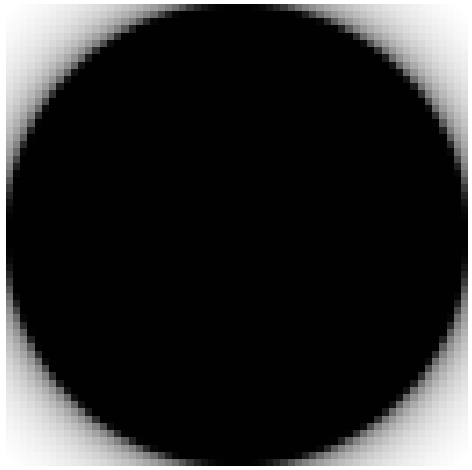
A little more work to compute than line-lengths, but worth the extra effort (particularly when pre-computed).

(Implicit) System Sensitivity Patterns

$$\sum_{i=1}^{n_{\mathrm{d}}} a_{ij} \approx s(\vec{r}_j) = \sum_{i=1}^{n_{\mathrm{d}}} s_i(\vec{r}_j)$$



Line Length



Strip Area

Backprojection of a uniform sinogram.

Explicitly:

$$\sum_{i=1}^{n_{\mathrm{d}}} a_{ij} = \sum_{i=1}^{n_{\mathrm{d}}} \int s_i(\vec{r})\, b_j(\vec{r})\, \mathrm{d}\vec{r} = \int \left[\sum_{i=1}^{n_{\mathrm{d}}} s_i(\vec{r})\right] b_j(\vec{r})\, \mathrm{d}\vec{r} = \int s(\vec{r})\, b_j(\vec{r})\, \mathrm{d}\vec{r} \approx s(\vec{r}_j)$$

where $\vec{r}_j$ is center of jth basis function.

Shows probability for each pixel that an emission from that pixel will be detected somewhere.

These nonuniformities propagate into the reconstructed images, except when sinograms are simulated from the same model of course.

Forward- / Back-projector "Pairs"

Forward projection (image domain to projection domain):

$$\bar{y}_i = \int s_i(\vec{r})\, f(\vec{r})\, \mathrm{d}\vec{r} = \sum_{j=1}^{n_{\mathrm{p}}} a_{ij} x_j = [\boldsymbol{A}\boldsymbol{x}]_i, \quad \text{or} \quad \bar{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{x}$$

Backprojection (projection domain to image domain):

$$\boldsymbol{A}'\boldsymbol{y} = \left\{ \sum_{i=1}^{n_{\mathrm{d}}} a_{ij} y_i \right\}_{j=1}^{n_{\mathrm{p}}}$$

The term "forward/backprojection pair" often corresponds to an implicit choice for the object basis and the system model.

Sometimes $\boldsymbol{A}'\boldsymbol{y}$ is implemented as $\boldsymbol{B}\boldsymbol{y}$ for some "backprojector" $\boldsymbol{B} \neq \boldsymbol{A}'$

Least-squares solutions (for example):

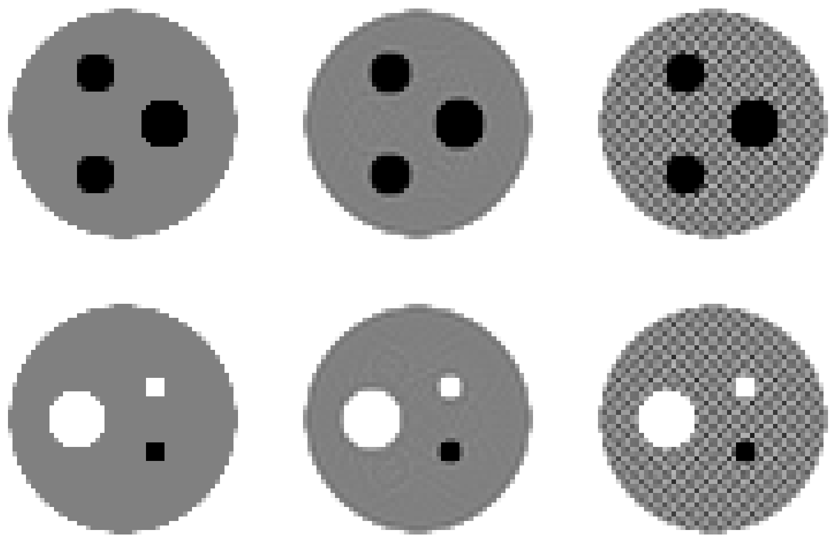
$$\hat{\boldsymbol{x}} = [\boldsymbol{A}'\boldsymbol{A}]^{-1}\boldsymbol{A}'\boldsymbol{y} \neq [\boldsymbol{B}\boldsymbol{A}]^{-1}\boldsymbol{B}\boldsymbol{y}$$

Algorithms are generally derived using a single $\boldsymbol{A}$ matrix, and usually the quantity $\boldsymbol{A}'\boldsymbol{y}$ appears somewhere in the derivation.

If the product $\boldsymbol{A}'\boldsymbol{y}$ is implemented by some $\boldsymbol{B}\boldsymbol{y}$ for $\boldsymbol{B} \neq \boldsymbol{A}'$, then all convergence properties, statistical properties, etc. of the theoretical algorithm may be lost by the implemented algorithm.

Mismatched Backprojector $B \neq A'$

x $\hat{x}$ (PWLS-CG) $\hat{x}$ (PWLS-CG)



Matched Mismatched

Note: when converting from .ps to .pdf, I get JPEG image compression artifacts that may corrupt these images. If I disable compression, then the files are 8x larger...

Noiseless 3D PET data, images are $n_x \times n_y \times n_z = 64 \times 64 \times 4$, with $n_u \times n_v \times n_\phi \times n_\theta = 62 \times 10 \times 60 \times 3$ projections. 15 iterations of PWLS-CG, initialized with the true image. True object values range from 0 to 2. Display windowed to [0.7, 1.3] to highlight artifacts.

In this case mismatch arises from a ray-driven forward projector but a pixel-driven back projector.

Another case where mismatch can arise is in "rotate and sum" projection / backprojection methods, if implemented carelessly.

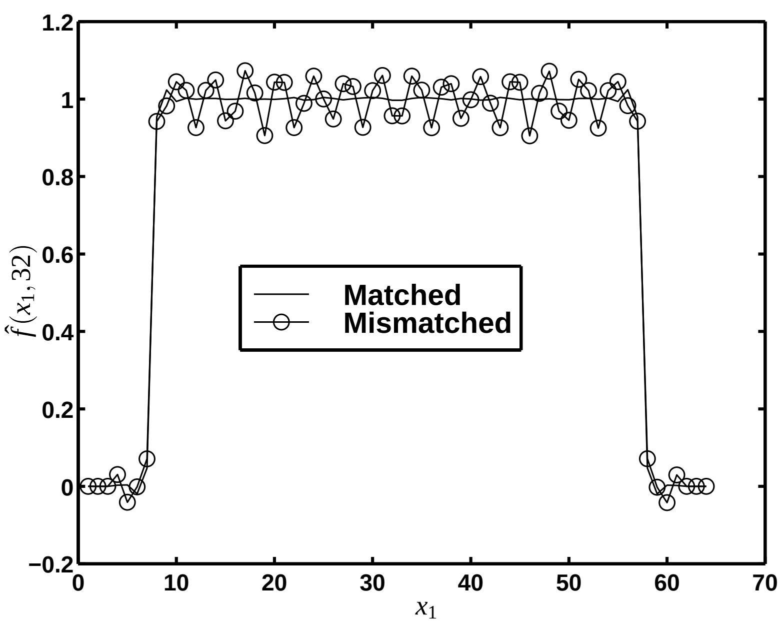
The problem with mismatched backprojectors arises in iterative reconstruction because multiple iterations are generally needed, so discrepancies between B and A' can accumulate.

Such discrepancies may matter more for regularized methods where convergence is desired, then for unregularized methods where one stops well before convergence [72], but this is merely speculation.

The deliberate use of mismatched projectors/backprojectors has been called the "dual matrix" approach [73, 74].

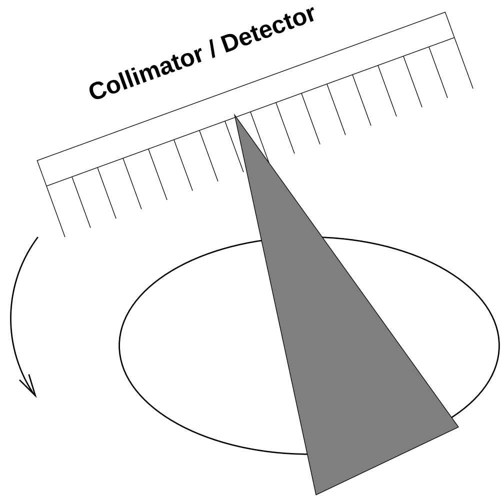
The importance of matching also arises in solving differential equations [75].

Horizontal Profiles



This was from noiseless simulated data!

SPECT System Modeling



Complications: nonuniform attenuation, depth-dependent PSF, Compton scatter

(MR system models discussed in Part II)

Numerous papers in the literature address aspects of the system model in the context of SPECT imaging. Substantial improvements in image quality and quantitative accuracy have been demonstrated by using appropriate system models.

Choice 3. Statistical Models

After modeling the system physics, we have a deterministic "model:"

$$y_i \approx g_i([\boldsymbol{A}\boldsymbol{x}]_i)$$

for some functions g_i, *e.g.*, $g_i(l) = l + r_i$ for emission tomography.

Statistical modeling is concerned with the " $\approx$ " aspect.

Considerations

- More accurate models:
  - can lead to lower variance images,
  - may incur additional computation,
  - may involve additional algorithm complexity (*e.g.*, proper transmission Poisson model has nonconcave log-likelihood)
- Statistical model errors (*e.g.*, deadtime)
- Incorrect models (*e.g.*, log-processed transmission data)

"Complexity" can just mean "inconvenience." It would certainly be more convenient to precorrect the sinogram data for effects such as randoms, attenuation, scatter, detector efficiency, etc., since that would save having to store those factors for repeated use during the iterations. But such precorrections destroy the Poisson statistics and lead to suboptimal performance (higher variance).

More accurate statistical models may also yield lower bias, but bias is often dominated by approximations in the system model (neglected scatter, etc.) and by resolution effects induced by regularization.

Statistical Model Choices for Emission Tomography

- "None." Assume $\boldsymbol{y} - \boldsymbol{r} = \boldsymbol{A}\boldsymbol{x}$. "Solve algebraically" to find $\boldsymbol{x}$.
- White Gaussian noise. Ordinary least squares: minimize $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2$ (This is the appropriate statistical model for MR.)
- Non-white Gaussian noise. Weighted least squares: minimize

$$\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|_{\boldsymbol{W}}^2 = \sum_{i=1}^{n_\mathrm{d}} w_i (y_i - [\boldsymbol{A}\boldsymbol{x}]_i)^2, \quad \text{where} \quad [\boldsymbol{A}\boldsymbol{x}]_i \triangleq \sum_{j=1}^{n_\mathrm{p}} a_{ij} x_j$$

  (*e.g.*, for Fourier rebinned (FORE) PET data)
- Ordinary Poisson model (ignoring or precorrecting for background)

$$y_i \sim \mathsf{Poisson}\{[\boldsymbol{A}\boldsymbol{x}]_i\}$$

- Poisson model

$$y_i \sim \mathsf{Poisson}\{[\boldsymbol{A}\boldsymbol{x}]_i + r_i\}$$

- Shifted Poisson model (for randoms precorrected PET)

$$y_i = y_i^{\text{prompt}} - y_i^{\text{delay}} \sim \mathsf{Poisson}\{[\boldsymbol{A}\boldsymbol{x}]_i + 2r_i\} - 2r_i$$

These are all for the emission case.

GE uses WLS for FORE data [76].

The shifted-Poisson model for randoms-precorrected PET is described in [77–80].

Snyder *et al.* used similar models for CCD imaging [81, 82].

Missing from the above list: deadtime model [83].

My recommendations.
- If the data is uncorrected, then use Poisson model above.
- If the data was corrected for random coincidences, use shifted Poisson model.
- If the data has been corrected for other stuff, consider using WLS, *e.g.* [84, 85].
- Try not to correct the data so that the first choice can be used!

Classic reason for WLS over Poisson was compute time. This has been obviated by recent algorithm advances. Now the choice should be made statistically.

Preprocessing: randoms subtraction, Fourier or multislice rebinning (3d to 2d), attenuation, scatter, detector efficiency, etc.

Shifted-Poisson Model for X-ray CT

A model that includes both photon variability and electronic readout noise:

$$y_i \sim \mathsf{Poisson}\{\bar{y}_i(\boldsymbol{\mu})\} + \mathsf{N}(0, \sigma^2)$$

Shifted Poisson approximation

$$\left[y_i + \sigma^2\right]_+ \sim \mathsf{Poisson}\left\{\bar{y}_i(\boldsymbol{\mu}) + \sigma^2\right\}$$

or just use WLS...

Complications:
- Intractability of likelihood for Poisson+Gaussian
- Compound Poisson distribution due to photon-energy-dependent detector signal.

X-ray statistical modeling is a current research area in several groups!

For Poisson+Gaussian, see [81, 82].

For compound Poisson distribution, see [86–88].

Choice 4. Cost Functions

Components:
- *Data-mismatch* term
- *Regularization* term (and regularization parameter β)
- Constraints (*e.g.*, nonnegativity)

Cost function:

$$\Psi(\boldsymbol{x}) = \text{DataMismatch}(\boldsymbol{y}, \boldsymbol{A}\boldsymbol{x}) + \beta\, \text{Roughness}(\boldsymbol{x})$$

Reconstruct image $\hat{\boldsymbol{x}}$ by minimization:

$$\hat{\boldsymbol{x}} \triangleq \arg\min_{\boldsymbol{x} \geq \boldsymbol{0}} \Psi(\boldsymbol{x})$$

Actually *several* sub-choices to make for Choice 4 ...

Distinguishes "statistical methods" from "algebraic methods" for "$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$."

β sometimes called *hyperparameter*

Why Cost Functions?

(vs "procedure" *e.g.*, adaptive neural net with wavelet denoising)

Theoretical reasons

ML is based on minimizing a cost function: the negative log-likelihood
- ML is asymptotically consistent
- ML is asymptotically unbiased
- ML is asymptotically efficient (under true statistical model...)
- Estimation: Penalized-likelihood achieves uniform CR bound asymptotically
- Detection: Qi and Huesman showed analytically that MAP reconstruction outperforms FBP for SKE/BKE lesion detection (T-MI, Aug. 2001)

Practical reasons
- Stability of estimates (if Ψ and algorithm chosen properly)
- Predictability of properties (despite nonlinearities)
- Empirical evidence (?)

Stability means that running "too many iterations" will not compromise image quality.

Asymptotically efficient means that the variance of ML estimator approaches that given by the Cramer-Rao lower bound, which is a bound on the variance of unbiased estimators.

But nuclear imaging is not asymptotic (too few counts), and system models are always approximate, and we regularize which introduces bias anyway.

Uniform CR bound generalizes CR bound to biased case [89, 90]

Bottom line: have not found anything better, seen plenty that are worse (LS vs ML in low count)

OSEM vs MAP [91, 92]

Qi and Huesman [42]

"Iterative FBP" methods are examples of methods that are not based on any cost function, and have not shared the popularity of ML and MAP approaches *e.g.*, [93–96].

Bayesian Framework

Given a prior distribution $\mathsf{p}(\boldsymbol{x})$ for image vectors $\boldsymbol{x}$, by Bayes' rule:

$$\text{posterior:}\quad \mathsf{p}(\boldsymbol{x}|\boldsymbol{y}) = \mathsf{p}(\boldsymbol{y}|\boldsymbol{x})\,\mathsf{p}(\boldsymbol{x})\,/\,\mathsf{p}(\boldsymbol{y})$$

so

$$\log \mathsf{p}(\boldsymbol{x}|\boldsymbol{y}) = \log \mathsf{p}(\boldsymbol{y}|\boldsymbol{x}) + \log \mathsf{p}(\boldsymbol{x}) - \log \mathsf{p}(\boldsymbol{y})$$

- $-\log \mathsf{p}(\boldsymbol{y}|\boldsymbol{x})$ corresponds to data mismatch term (negative log-likelihood)
- $-\log \mathsf{p}(\boldsymbol{x})$ corresponds to regularizing penalty function

Maximum a posteriori (MAP) estimator:

$$\hat{\boldsymbol{x}} = \arg\max_{\boldsymbol{x}} \log \mathsf{p}(\boldsymbol{x}|\boldsymbol{y}) = \arg\max_{\boldsymbol{x}} \log \mathsf{p}(\boldsymbol{y}|\boldsymbol{x}) + \log \mathsf{p}(\boldsymbol{x})$$

- Has certain optimality properties (provided $\mathsf{p}(\boldsymbol{y}|\boldsymbol{x})$ and $\mathsf{p}(\boldsymbol{x})$ are correct).
- Same form as Ψ

I avoid the Bayesian terminology because

- Images drawn from the "prior" distributions almost never look like real objects
- The risk function associated with MAP estimation seems less natural to me than a quadratic risk function. The quadratic choice corresponds to conditional mean estimation $\hat{\boldsymbol{x}} = \mathsf{E}[\boldsymbol{x}|\boldsymbol{y}]$ which is used very rarely by those who describe Bayesian methods for image formation.
- I often use penalty functions $R(\boldsymbol{x})$ that depend on the data $\boldsymbol{y}$, which can hardly be called "priors," *e.g.*, [36].

Choice 4.1: Data-Mismatch Term

Options (for emission tomography):

- Negative log-likelihood of statistical model. Poisson *emission* case:

$$-L(\boldsymbol{x};\boldsymbol{y}) = -\log \mathsf{p}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} ([\boldsymbol{A}\boldsymbol{x}]_i + r_i) - y_i \log([\boldsymbol{A}\boldsymbol{x}]_i + r_i) + \log y_i!$$

- Ordinary (unweighted) least squares: $\sum_{i=1}^{n_\mathrm{d}} \frac{1}{2}(y_i - \hat{r}_i - [\boldsymbol{A}\boldsymbol{x}]_i)^2$
- Data-weighted least squares: $\sum_{i=1}^{n_\mathrm{d}} \frac{1}{2}(y_i - \hat{r}_i - [\boldsymbol{A}\boldsymbol{x}]_i)^2/\hat{\sigma}_i^2,\ \hat{\sigma}_i^2 = \max\left(y_i + \hat{r}_i, \sigma_{\min}^2\right)$, (causes bias due to data-weighting).
- Reweighted least-squares: $\hat{\sigma}_i^2 = [\boldsymbol{A}\hat{\boldsymbol{x}}]_i + \hat{r}_i$
- Model-weighted least-squares (nonquadratic, but convex!)

$$\sum_{i=1}^{n_\mathrm{d}} \frac{1}{2}(y_i - \hat{r}_i - [\boldsymbol{A}\boldsymbol{x}]_i)^2/([\boldsymbol{A}\boldsymbol{x}]_i + \hat{r}_i)$$

- Nonquadratic cost-functions that are robust to outliers
- ...

Considerations

- Faithfulness to statistical model vs computation
- Ease of optimization (convex?, quadratic?)
- Effect of statistical modeling errors

Poisson probability mass function (PMF):

$$\mathsf{p}(\boldsymbol{y}|\boldsymbol{x}) = \prod_{i=1}^{n_\mathrm{d}} e^{-\bar{y}_i}\bar{y}_i^{y_i}/y_i! \quad\text{where}\quad \bar{\boldsymbol{y}} \triangleq \boldsymbol{A}\boldsymbol{x} + \boldsymbol{r}$$

Reweighted least-squares [97]

Model-weighted least-squares [98, 99]

$$f(l) = \frac{1}{2}(y - r - l)^2/(l + r) \qquad \ddot{f}(l) = y^2/(l + r)^3 > 0$$

Robust norms [100, 101]

Generally the data-mismatch term and the statistical model go hand-in-hand.

Choice 4.2: Regularization

Forcing too much “data fit” gives noisy images
Ill-conditioned problems: small data noise causes large image noise

Solutions:
- Noise-reduction methods
- True regularization methods

Noise-reduction methods
- Modify the *data*
  - Prefilter or “denoise” the sinogram measurements
  - Extrapolate missing (*e.g.*, truncated) data
- Modify an *algorithm* derived for an ill-conditioned problem
  - Stop algorithm before convergence
  - Run to convergence, post-filter
  - Toss in a filtering step every iteration or couple iterations
  - Modify update to “dampen” high-spatial frequencies

Dampen high-frequencies in EM [102]

FBP with an apodized ramp filter belongs in the “modify the algorithm” category. The FBP method is derived based on a highly idealized system model. The solution so derived includes a ramp filter, which causes noise amplification if used unmodified. Throwing in apodization of the ramp filter attempts to “fix” this problem with the FBP “algorithm.”

The fault is not with the *algorithm* but with the problem definition and cost function. Thus the fix should be to the latter, not to the algorithm.

The estimate-maximize smooth (EMS) method [103] uses filtering every iteration.

The continuous image $f(\vec{r})$- discrete data problem is *ill-posed*.

If the discrete-discrete problem has a full column rank system matrix $\boldsymbol{A}$, then that problem is well-posed, but still probably ill-conditioned.

Noise-Reduction vs True Regularization

Advantages of noise-reduction methods
- Simplicity (?)
- Familiarity
- Appear less subjective than using penalty functions or priors
- Only fiddle factors are # of iterations, or amount of smoothing
- Resolution/noise tradeoff usually varies with iteration
  (stop when image looks good - in principle)
- Changing post-smoothing does not require re-iterating

Advantages of true regularization methods
- Stability (unique minimizer & convergence $\Longrightarrow$ initialization independence)
- Faster convergence
- Predictability
- Resolution can be made object independent
- Controlled resolution (*e.g.*, spatially uniform, edge preserving)
- Start with reasonable image (*e.g.*, FBP) $\Longrightarrow$ reach solution faster.

Running many iterations followed by post-filtering seems preferable to aborting early by stopping rules [104, 105].

Lalush *et al.* reported small differences between post-filtering and MAP reconstructions with an entropy prior [106].

Slijpen and Beekman conclude that post-filtering slightly more accurate than “oracle” filtering between iterations for SPECT reconstruction [107].

True Regularization Methods

Redefine the *problem* to eliminate ill-conditioning,
rather than patching the data or algorithm!

Options

- Use bigger pixels (fewer basis functions)
  - Visually unappealing
  - Can only preserve edges coincident with pixel edges
  - Results become even less invariant to translations
- Method of sieves (constrain image roughness)
  - Condition number for "pre-emission space" can be even worse
  - Lots of iterations
  - Commutability condition rarely holds exactly in practice
  - Degenerates to post-filtering in some cases
- Change cost function by adding a roughness penalty / prior

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}), \qquad \Psi(\boldsymbol{x}) = Ł(\boldsymbol{x}) + \beta R(\boldsymbol{x})$$

  - Disadvantage: apparently subjective choice of penalty
  - Apparent difficulty in choosing penalty parameter(s), *e.g.*, β (*cf.* apodizing filter / cutoff frequency in FBP)

Big pixels [108]

Sieves [109, 110]

Lots of iterations for convergence [104, 111]

Penalty Function Considerations

- Computation
- Algorithm complexity
- Uniqueness of minimizer of $\Psi(\boldsymbol{x})$
- Resolution properties (edge preserving?)
- # of adjustable parameters
- Predictability of properties (resolution and noise)

Choices

- separable vs nonseparable
- quadratic vs nonquadratic
- convex vs nonconvex

There is a huge literature on different regularization methods. Of the many proposed methods, and many anecdotal results illustrating properties of such methods, only the "lowly" quadratic regularization method has been shown *analytically* to yield detection results that are superior to FBP [42].

Penalty Functions: Separable vs Nonseparable

Separable

- Identity norm: $R(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}'\boldsymbol{I}\boldsymbol{x} = \sum_{j=1}^{n_p} x_j^2/2$
  penalizes large values of $\boldsymbol{x}$, but causes "squashing bias"
- Entropy: $R(\boldsymbol{x}) = \sum_{j=1}^{n_p} x_j \log x_j$
- Gaussian prior with mean μ_j, variance σ_j^2: $R(\boldsymbol{x}) = \sum_{j=1}^{n_p} \frac{(x_j-\mu_j)^2}{2\sigma_j^2}$
- Gamma prior $R(\boldsymbol{x}) = \sum_{j=1}^{n_p} \mathsf{p}(x_j, \mu_j, \sigma_j)$ where $\mathsf{p}(x, \mu, \sigma)$ is Gamma pdf

The first two basically keep pixel values from "blowing up."
The last two encourage pixels values to be close to prior means μ_j.

General separable form: $$R(\boldsymbol{x}) = \sum_{j=1}^{n_p} f_j(x_j)$$

Slightly simpler for minimization, but these do not explicitly enforce smoothness.
The simplicity advantage has been overcome in newer algorithms.

The identity norm penalty is a form of Tikhinov-Miller regularization [112].

The Gaussian and Gamma bias the results towards the prior image. This can be good or bad depending on whether the prior image is correct or not! If the prior image comes from a normal database, but the patient is abnormal, such biases would be undesirable.

For arguments favoring maximum entropy, see [113]. For critiques of maximum entropy regularization, see [114–116].

A key development in overcoming the "difficulty" with nonseparable regularization was a 1995 paper by De Pierro: [117].

Penalty Functions: Separable vs Nonseparable

Nonseparable (partially couple pixel values) to penalize *roughness*

x_1	x_2	x_3
x_4	x_5	

Example

$$R(\boldsymbol{x}) = (x_2-x_1)^2 + (x_3-x_2)^2 + (x_5-x_4)^2 + (x_4-x_1)^2 + (x_5-x_2)^2$$

2	2	2
2	1	

$R(\boldsymbol{x}) = 1$

3	3	1
2	2	

$R(\boldsymbol{x}) = 6$

1	3	1
2	2	

$R(\boldsymbol{x}) = 10$

Rougher images $\Longrightarrow$ larger $R(\boldsymbol{x})$ values

If diagonal neighbors were included there would be 3 more terms in this example.

Roughness Penalty Functions

First-order neighborhood and pairwise pixel differences:

$$R(\boldsymbol{x}) = \sum_{j=1}^{n_{\mathrm{p}}} \frac{1}{2} \sum_{k \in \mathcal{N}_j} \psi(x_j - x_k)$$

$\mathcal{N}_j \triangleq$ *neighborhood* of jth pixel (*e.g.*, left, right, up, down)
ψ called the *potential function*

Finite-difference approximation to continuous roughness measure:

$$R(f(\cdot)) = \int \|\nabla f(\vec{r})\|^2 \, d\vec{r} = \int \left|\frac{\partial}{\partial x} f(\vec{r})\right|^2 + \left|\frac{\partial}{\partial y} f(\vec{r})\right|^2 + \left|\frac{\partial}{\partial z} f(\vec{r})\right|^2 d\vec{r}.$$

Second derivatives also useful:
(More choices!)

$$\left.\frac{\partial^2}{\partial x^2} f(\vec{r})\right|_{\vec{r}=\vec{r}_j} \approx f(\vec{r}_{j+1}) - 2 f(\vec{r}_j) + f(\vec{r}_{j-1})$$

$$R(\boldsymbol{x}) = \sum_{j=1}^{n_{\mathrm{p}}} \psi(x_{j+1} - 2x_j + x_{j-1}) + \cdots$$

For differentiable basis functions (*e.g.*, B-splines), one can find $\int \|\nabla f(\vec{r})\|^2 d\vec{r}$ exactly in terms of coefficients, *e.g.*, [118].

See Gindi *et al.* [119, 120] for comparisons of first and second order penalties.

Penalty Functions: General Form

$$\boxed{R(\boldsymbol{x}) = \sum_k \psi_k([\boldsymbol{C}\boldsymbol{x}]_k)} \quad \text{where} \quad [\boldsymbol{C}\boldsymbol{x}]_k = \sum_{j=1}^{n_{\mathrm{p}}} c_{kj} x_j$$

Example:

x_1	x_2	x_3
x_4	x_5	

$$\boldsymbol{C}\boldsymbol{x} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} x_2 - x_1 \\ x_3 - x_2 \\ x_5 - x_4 \\ x_4 - x_1 \\ x_5 - x_2 \end{bmatrix}$$

$$\begin{aligned} R(\boldsymbol{x}) &= \sum_{k=1}^{5} \psi_k([\boldsymbol{C}\boldsymbol{x}]_k) \\ &= \psi_1(x_2 - x_1) + \psi_2(x_3 - x_2) + \psi_3(x_5 - x_4) + \psi_4(x_4 - x_1) + \psi_5(x_5 - x_2) \end{aligned}$$

This form is general enough to cover nearly all the penalty functions that have been used in tomography. Exceptions include priors based on nonseparable line-site models [121–124], and the median root "prior" [125, 126], both of which are nonconvex.

It is just coincidence that $\boldsymbol{C}$ is square in this example. In general, for a $n_x \times n_y$ image, there are $n_x(n_y - 1)$ horizontal pairs and $n_y(n_x - 1)$ vertical pairs, so $\boldsymbol{C}$ will be a $(2n_x n_y - n_x - n_y) \times (n_x n_x)$ *very* sparse matrix (for a first-order neighborhood consisting of horizontal and vertical cliques).

Concretely, for a $n_x \times n_y$ image ordered lexicographically, for a first-order neighborhood we use

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{I}_{n_y} \otimes \boldsymbol{D}_{n_x} \\ \boldsymbol{D}_{n_y} \otimes \boldsymbol{I}_{n_x} \end{bmatrix}$$

where $\otimes$ denotes the Kronecker product and $\boldsymbol{D}_n$ denotes the following $(n-1) \times n$ matrix:

$$\boldsymbol{D}_n \triangleq \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & \ddots & \ddots & 0 \\ 0 & 0 & 0 & -1 & 1 \end{bmatrix}.$$

Penalty Functions: Quadratic vs Nonquadratic

$$R(\boldsymbol{x}) = \sum_k \psi_k([\boldsymbol{C}\boldsymbol{x}]_k)$$

Quadratic ψ_k

If $\psi_k(t) = t^2/2$, then $R(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}'\boldsymbol{C}'\boldsymbol{C}\boldsymbol{x}$, a quadratic form.

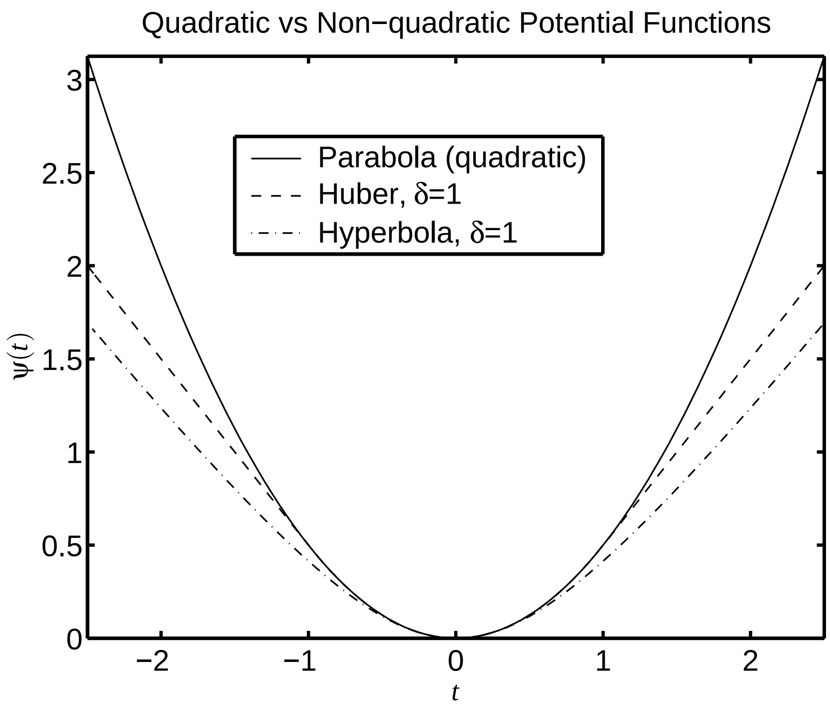
- Simpler optimization
- Global smoothing

Nonquadratic ψ_k

- Edge preserving
- More complicated optimization. (This is essentially solved in convex case.)
- Unusual noise properties
- Analysis/prediction of resolution and noise properties is difficult
- More adjustable parameters (*e.g.*, δ)

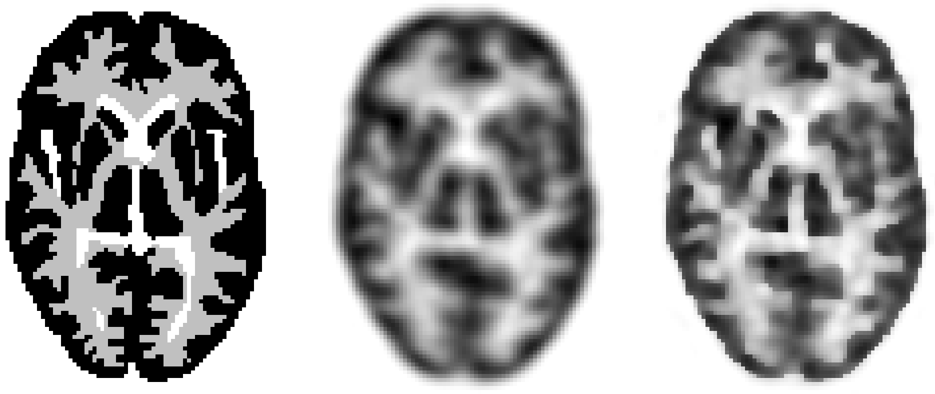
Example: Huber function. $\psi(t) \triangleq \begin{cases} t^2/2, & |t| \leq \delta \\ \delta|t| - \delta^2/2, & |t| > \delta \end{cases}$

Example: Hyperbola function. $\psi(t) \triangleq \delta^2 \left(\sqrt{1+(t/\delta)^2} - 1\right)$



Lower cost for large differences $\Longrightarrow$ edge preservation

Edge-Preserving Reconstruction Example

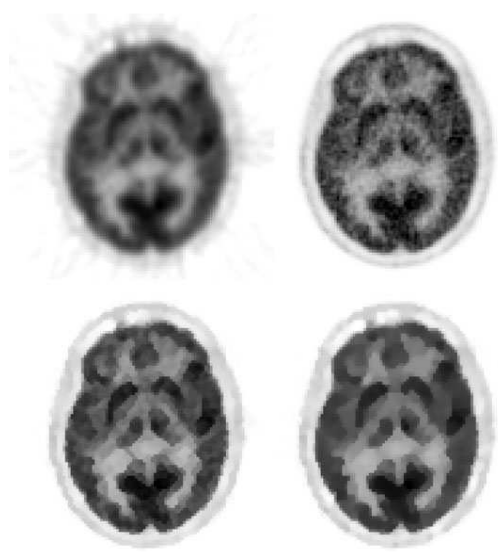


Phantom   Quadratic Penalty   Huber Penalty

In terms of ROI quantification, a nonquadratic penalty may outperform quadratic penalties for certain types of objects (especially phantom-like piecewise smooth objects). But the benefits of nonquadratic penalties for visual tasks is largely unknown.

The smaller δ is in the Huber penalty, the stronger the degree of edge preservation, and the more unusual the noise effects. In this case I used $\delta = 0.4$, for a phantom that is 0 in background, 1 in white matter, 4 in graymatter. Thus δ is one tenth the maximum value, as has been recommended by some authors.

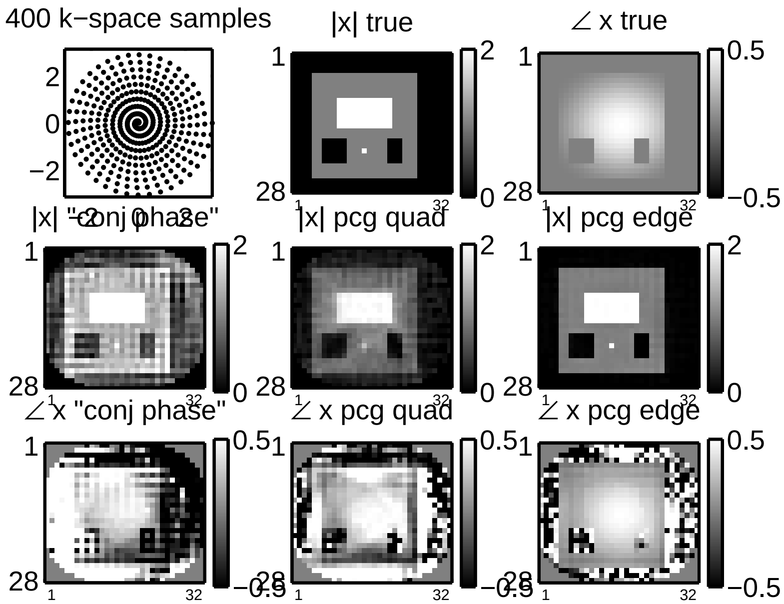
More "Edge Preserving" Regularization



Chlewicki *et al.*, PMB, Oct. 2004: "Noise reduction and convergence of Bayesian algorithms with blobs based on the Huber function and median root prior"

Figure taken from [127].

Piecewise Constant “Cartoon” Objects



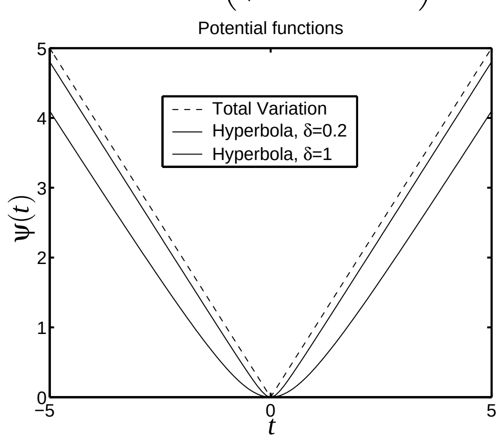
Total Variation Regularization

Non-quadratic roughness penalty:

$$\int \|\nabla f(\vec{r})\| \, d\vec{r} \approx \sum_k |[\boldsymbol{Cx}]_k|$$

Uses *magnitude* instead of *squared magnitude* of gradient.

Problem: $|\cdot|$ is not differentiable.
Practical solution: $$|t| \approx \delta\left(\sqrt{1+(t/\delta)^2}-1\right)$$ (hyperbola!)



To be more precise, in 2D: $\|\nabla f(x,y)\| = \sqrt{\left|\frac{\partial}{\partial x} f\right|^2 + \left|\frac{\partial}{\partial y} f\right|^2}$ so the *total variation* is

$$\iint \|\nabla f(x,y)\| \, dx\,dy \approx \sum_n \sum_m \sqrt{|f(n,m)-f(n-1,m)|^2 + |f(n,m)-f(n,m-1)|^2}$$

Total variation in image reconstruction [128–130]. A critique [131].

Penalty Functions: Convex vs Nonconvex

Convex
- Easier to optimize
- Guaranteed unique minimizer of Ψ (for convex negative log-likelihood)

Nonconvex
- Greater degree of edge preservation
- Nice images for piecewise-constant phantoms!
- Even more unusual noise properties
- Multiple extrema
- More complicated optimization (simulated / deterministic annealing)
- Estimator $\hat{\boldsymbol{x}}$ becomes a discontinuous function of data $\boldsymbol{Y}$

Nonconvex examples
- "broken parabola"

$$\psi(t) = \min(t^2,\ t_{\max}^2)$$

- true median root prior:

$$R(\boldsymbol{x}) = \sum_{j=1}^{n_p} \frac{(x_j - \mathrm{median}_j(\boldsymbol{x}))^2}{\mathrm{median}_j(\boldsymbol{x})} \quad \text{where } \mathrm{median}_j(\boldsymbol{x}) \text{ is local median}$$

Exception: orthonormal wavelet threshold *denoising* via nonconvex potentials!

The above form is not exactly what has been called the median root prior by Alenius *et al.* [126]. They have used $\mathrm{median}_j(\boldsymbol{x}^{(n)})$ which is not a true prior since it depends on the previous iteration. Hsiao, Rangarajan, and Ginda have developed a very interesting prior that is similar to the "medial root prior" but is convex [132].

For nice analysis of nonconvex problems, see the papers by Mila Nikolova [133].

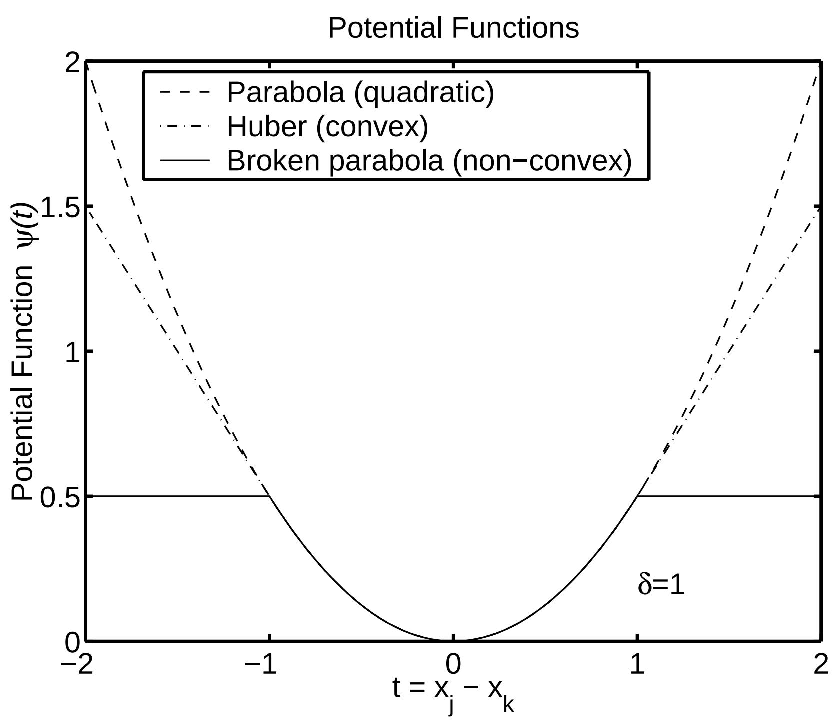
For orthonormal wavelet denoising, the cost functions [134] usually have the form

$$\Psi(\boldsymbol{x}) = \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 + \sum_{j=1}^{n_p} \psi(x_j)$$

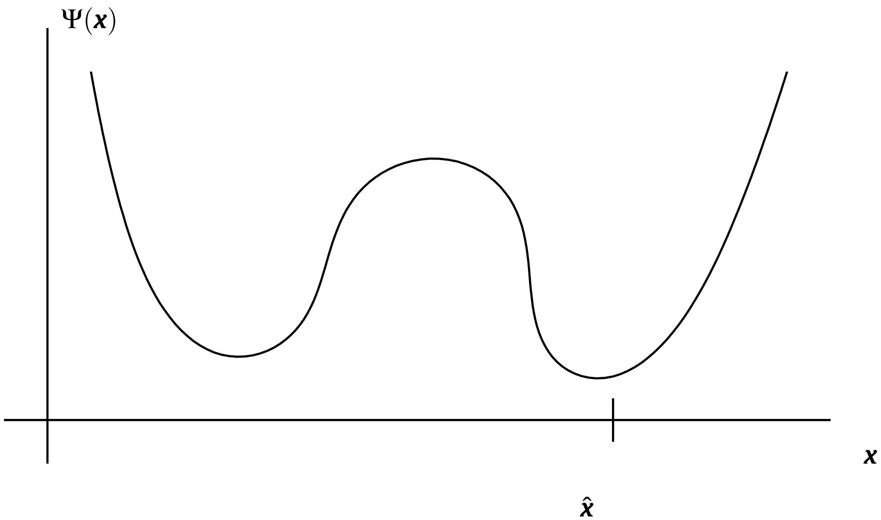
where $\boldsymbol{A}$ is an orthonormal. When $\boldsymbol{A}$ is orthonormal we can write: $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 = \|\boldsymbol{A}'\boldsymbol{y} - \boldsymbol{x}\|^2$, so

$$\Psi(\boldsymbol{x}) = \sum_{j=1}^{n_p} (x_j - [\boldsymbol{A}'\boldsymbol{y}]_j)^2 + \psi(x_j)$$

which separates completely into n_p 1-D minimization problems, each of which has a *unique* minimizer for all useful potential functions.



Local Extrema and Discontinuous Estimators



Small change in data $\Longrightarrow$ large change in minimizer $\hat{\boldsymbol{x}}$.
Using convex penalty functions obviates this problem.

[101] discuss discontinuity

Augmented Regularization Functions

Replace roughness penalty $R(\boldsymbol{x})$ with $R(\boldsymbol{x}|\boldsymbol{b}) + \alpha R(\boldsymbol{b})$,
where the elements of $\boldsymbol{b}$ (often binary) indicate boundary locations.

- Line-site methods
- Level-set methods

Joint estimation problem:

$$(\hat{\boldsymbol{x}}, \hat{\boldsymbol{b}}) = \arg\min_{\boldsymbol{x},\boldsymbol{b}} \Psi(\boldsymbol{x}, \boldsymbol{b}), \qquad \Psi(\boldsymbol{x}, \boldsymbol{b}) = \text{Ł}(\boldsymbol{x})[\boldsymbol{x};\boldsymbol{y}] + \beta R(\boldsymbol{x}|\boldsymbol{b}) + \alpha R(\boldsymbol{b}).$$

Example: b_{jk} indicates the presence of edge between pixels j and k:

$$R(\boldsymbol{x}|\boldsymbol{b}) = \sum_{j=1}^{n_p} \sum_{k \in \mathcal{N}_j} (1 - b_{jk}) \frac{1}{2} (x_j - x_k)^2$$

Penalty to discourage too many edges (*e.g.*):

$$R(\boldsymbol{b}) = \sum_{jk} b_{jk}.$$

- Can encourage local edge continuity
- May require annealing methods for minimization

Line-site methods: [121–124].
Level-set methods: [135–137].

For the simple *non-interacting* line-site penalty function $R(\boldsymbol{b})$ given above, one can perform the minimization over $\boldsymbol{b}$ analytically, yielding an equivalent regularization method of the form $R(\boldsymbol{x})$ with a broken parabola potential function [138].

More sophisticated line-site methods use neighborhoods of line-site variables to encourage local boundary continuity [121–124].

The convex median prior of Hsiao *et al.* uses augmented regularization but does not require annealing [132].

Modified Penalty Functions

$$R(\boldsymbol{x}) = \sum_{j=1}^{n_p} \frac{1}{2} \sum_{k \in \mathcal{N}_j} w_{jk} \psi(x_j - x_k)$$

Adjust weights $\{w_{jk}\}$ to
- Control resolution properties
- Incorporate anatomical side information (MR/CT)
  (avoid smoothing across anatomical boundaries)

Recommendations
- Emission tomography:
  - Begin with quadratic (nonseparable) penalty functions
  - Consider modified penalty for resolution control and choice of β
  - Use modest regularization and post-filter more if desired
- Transmission tomography (attenuation maps), X-ray CT
  - consider convex nonquadratic (*e.g.*, Huber) penalty functions
  - choose δ based on attenuation map units (water, bone, etc.)
  - choice of regularization parameter β remains nontrivial,
    learn appropriate values by experience for given study type

---

Resolution properties [36, 139–141].

Side information (a very incomplete list) [142–153].

Choice 4.3: Constraints

- Nonnegativity
- Known support
- Count preserving
- Upper bounds on values
  e.g., maximum μ of attenuation map in transmission case

Considerations
- Algorithm complexity
- Computation
- Convergence rate
- Bias (in low-count regions)
- . . .

Sometimes it is stated that the ML-EM algorithm "preserves counts." This only holds when $r_i = 0$ in the statistical model. The count-preserving property originates from the likelihood, not the algorithm. The ML estimate, under the Poisson model, happens to preserve counts. It is fine that ML-EM does so every iteration, but that does not mean that it is superior to other algorithms that get to the optimum $\hat{\boldsymbol{x}}$ faster without necessarily preserving counts along the way.

I do not recommend artificially renormalizing each iteration to try to "preserve counts."

Open Problems

- Performance prediction for nonquadratic penalties
- Effect of nonquadratic penalties on detection tasks
- Choice of regularization parameters for nonquadratic regularization

Deadtime statistics are analyzed in [154,155]. Bottom line: in most SPECT and PET systems with paralyzable deadtime, the measurements are non-Poisson, but the mean and variance are nearly identical. So presumably the Poisson statistical model is adequate, provided the deadtime losses are included in the system matrix $\boldsymbol{A}$.

Some of these types of questions are being addressed, *e.g.*, effects of sensitivity map errors (a type of system model mismatch) in list-mode reconstruction [156]. Qi's bound on system model error relative to data error: [157].

Summary

- 1. Object parameterization: function $f(\vec{r})$ vs vector $\boldsymbol{x}$
- 2. System physical model: $s_i(\vec{r})$
- 3. Measurement statistical model $Y_i \sim \boxed{?}$
- 4. Cost function: data-mismatch / regularization / constraints

$$\boxed{\textbf{Reconstruction Method} \triangleq \textbf{Cost Function + Algorithm}}$$

Naming convention: "criterion"-"algorithm":
- ML-EM, MAP-OSL, PL-SAGE, PWLS+SOR, PWLS-CG, . . .

Part 3. Algorithms

Method = Cost Function + Algorithm

Outline
- Ideal algorithm
- Classical general-purpose algorithms
- Considerations:
  - nonnegativity
  - parallelization
  - convergence rate
  - monotonicity
- Algorithms tailored to cost functions for imaging
  - Optimization transfer
  - EM-type methods
  - Poisson emission problem
  - Poisson transmission problem
- Ordered-subsets / block-iterative algorithms
  - Recent convergent versions

Choosing a cost function is an important part of imaging science.

Choosing an algorithm should be mostly a matter of computer science (numerical methods).

Nevertheless, it gets a lot of attention by imaging scientists since our cost functions have forms that can be exploited to get faster convergence than general-purpose methods.

Why iterative algorithms?

- For nonquadratic Ψ, no closed-form solution for minimizer.
- For quadratic Ψ with nonnegativity constraints, no closed-form solution.
- For quadratic Ψ without constraints, closed-form solutions:

$$\text{PWLS:} \quad \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2_{\boldsymbol{W}^{1/2}} + \boldsymbol{x}'\boldsymbol{R}\boldsymbol{x} = [\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \boldsymbol{R}]^{-1}\boldsymbol{A}'\boldsymbol{W}\boldsymbol{y}$$

$$\text{OLS:} \quad \hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2 = [\boldsymbol{A}'\boldsymbol{A}]^{-1}\boldsymbol{A}'\boldsymbol{y}$$

  Impractical (memory and computation) for realistic problem sizes.
  $\boldsymbol{A}$ is sparse, but $\boldsymbol{A}'\boldsymbol{A}$ is not.

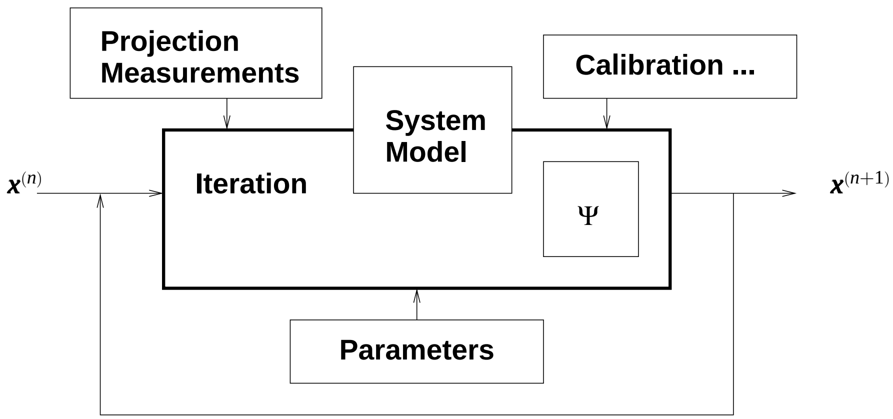
All algorithms are imperfect. No single best solution.

Singular value decomposition (*SVD*) techniques have been proposed for the OLS cost function as a method for reducing the computation problem, *e.g.*, [158–167].

The idea is that one could precompute the pseudo-inverse of $\boldsymbol{A}$ "once and for all." However $\boldsymbol{A}$ includes physical effects like attenuation, which change for every patient. And for data-weighted least squares, $\boldsymbol{W}$ changes for each scan too.

Image reconstruction never requires the matrix inverse $[\boldsymbol{A}'\boldsymbol{A}]^{-1}$; all that is required is a solution to the normal equations $[\boldsymbol{A}'\boldsymbol{A}]\hat{\boldsymbol{x}} = \boldsymbol{A}'\boldsymbol{y}$ which is easier, but still nontrivial.

General Iteration



Deterministic iterative mapping: $\boldsymbol{x}^{(n+1)} = \mathcal{M}(\boldsymbol{x}^{(n)})$

There are also stochastic iterative algorithms, such as simulated annealing [121] and the stochastic EM algorithm [168].

Ideal Algorithm

$$\boldsymbol{x}^\star \triangleq \arg\min_{\boldsymbol{x} \geq \mathbf{0}} \Psi(\boldsymbol{x}) \qquad \text{(global minimizer)}$$

Properties

stable and convergent	$\{\boldsymbol{x}^{(n)}\}$ converges to $\boldsymbol{x}^\star$ if run indefinitely
converges quickly	$\{\boldsymbol{x}^{(n)}\}$ gets "close" to $\boldsymbol{x}^\star$ in just a few iterations
globally convergent	$\lim_n \boldsymbol{x}^{(n)}$ independent of starting image $\boldsymbol{x}^{(0)}$
fast	requires minimal computation per iteration
robust	insensitive to finite numerical precision
user friendly	nothing to adjust (*e.g.*, acceleration factors)
parallelizable	(when necessary)
simple	easy to program and debug
flexible	accommodates any type of system model

(matrix stored by row or column, or factored, or projector/backprojector)

Choices: forgo one or more of the above

One might argue that the "ideal algorithm" would be the algorithm that produces $\boldsymbol{x}^{\text{true}}$. In the framework presented here, it is the job of the cost function to try to make $\boldsymbol{x}^\star \approx \boldsymbol{x}^{\text{true}}$, and the job of the algorithm to find $\boldsymbol{x}^\star$ by minimizing Ψ.

In fact, *nothing* in the above list really has to do with image quality. In the statistical framework, image quality is determined by Ψ, not by the algorithm.

Note on terminology: "algorithms" do not really converge, it is the *sequence* of estimates $\{\boldsymbol{x}^{(n)}\}$ that converges, but everyone abuses this all the time, so I will too.

Classic Algorithms

Non-gradient based
- Exhaustive search
- Nelder-Mead simplex (amoeba)

Converge very slowly, but work with nondifferentiable cost functions.

Gradient based
- Gradient descent
$$\boldsymbol{x}^{(n+1)} \triangleq \boldsymbol{x}^{(n)} - \alpha \nabla \Psi\left(\boldsymbol{x}^{(n)}\right)$$
Choosing α to ensure convergence is nontrivial.
- Steepest descent
$$\boldsymbol{x}^{(n+1)} \triangleq \boldsymbol{x}^{(n)} - \alpha_n \nabla \Psi\left(\boldsymbol{x}^{(n)}\right) \quad \text{where} \quad \alpha_n \triangleq \arg\min_{\alpha} \Psi\left(\boldsymbol{x}^{(n)} - \alpha \nabla \Psi\left(\boldsymbol{x}^{(n)}\right)\right)$$
Computing stepsize α_n can be expensive or inconvenient.

Limitations
- Converge slowly.
- Do not easily accommodate nonnegativity constraint.

Nice discussion of optimization algorithms in [169].

Row and column gradients:

$$\nabla \Psi(\boldsymbol{x}) = \left[\frac{\partial}{\partial x_1}\Psi, \frac{\partial}{\partial x_2}\Psi, \ldots, \frac{\partial}{\partial x_{n_p}}\Psi\right], \qquad \nabla = \nabla'$$

Using gradients excludes nondifferentiable penalty functions such as the Laplacian prior which involves $|x_j - x_k|$. See [170–172] for solutions to this problem.

Gradients & Nonnegativity - A Mixed Blessing

Unconstrained optimization of differentiable cost functions:

$$\nabla \Psi(\boldsymbol{x}) = \boldsymbol{0} \;\text{ when }\; \boldsymbol{x} = \boldsymbol{x}^\star$$

- A necessary condition always.
- A sufficient condition for strictly convex cost functions.
- Iterations search for zero of gradient.

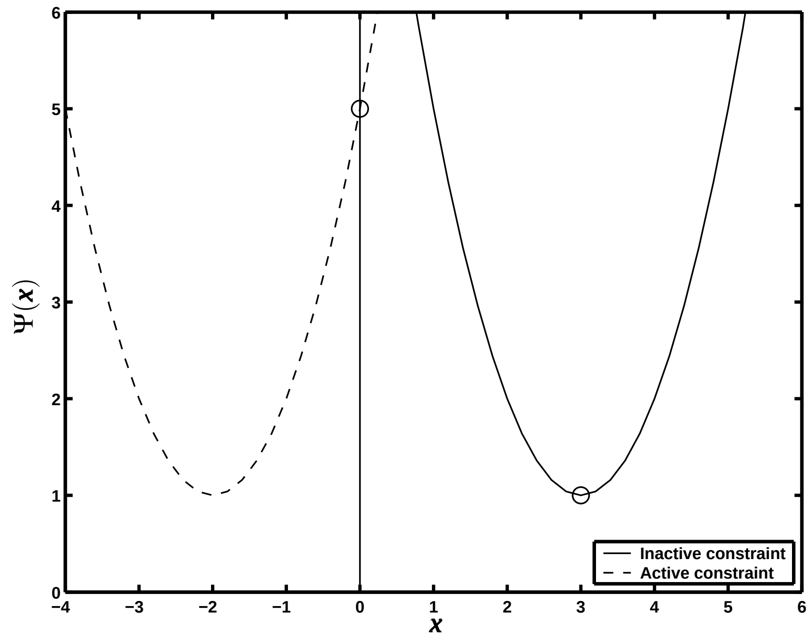
Nonnegativity-constrained minimization:

Karush-Kuhn-Tucker conditions

$$\left.\frac{\partial}{\partial x_j}\Psi(\boldsymbol{x})\right|_{\boldsymbol{x}=\boldsymbol{x}^\star} \;\text{is}\; \begin{cases} = 0, & x_j^\star > 0 \\ \geq 0, & x_j^\star = 0 \end{cases}$$

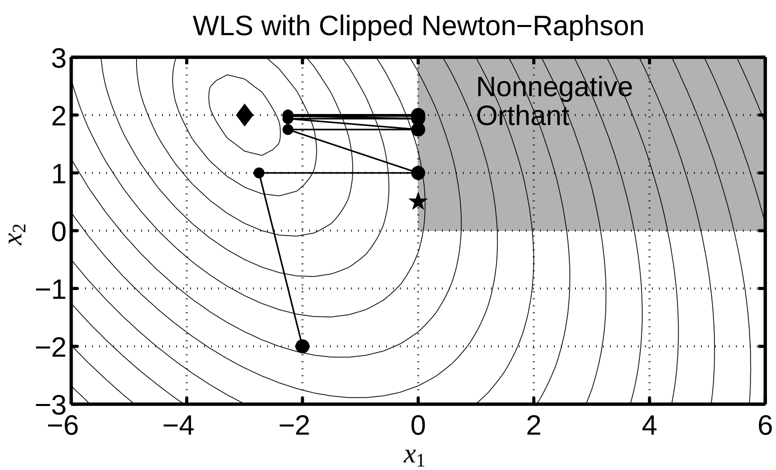
- A necessary condition always.
- A sufficient condition for strictly convex cost functions.
- Iterations search for ???
- $0 = x_j^\star \frac{\partial}{\partial x_j}\Psi(\boldsymbol{x}^\star)$ is a necessary condition, but never sufficient condition.

Karush-Kuhn-Tucker Illustrated



The usual condition $\frac{\partial}{\partial x_j}\Psi(\boldsymbol{x}) = 0$ only applies for pixels where the nonnegativity constraint is inactive.

Why Not Clip Negatives?



Newton-Raphson with negatives set to zero each iteration.
Fixed-point of iteration is not the constrained minimizer!

By clipped negatives, I mean you start with some nominal algorithm $\mathcal{M}_0(\boldsymbol{x})$ and modify it to be: $\boldsymbol{x}^{(n+1)} = \mathcal{M}(\boldsymbol{x}^{(n)})$ where $\mathcal{M}(\boldsymbol{x}) = [\mathcal{M}_0(\boldsymbol{x})]_+$ and the jth element of $[\boldsymbol{x}]_+$ is x_j if $x_j > 0$ or 0 if $x_j \leq 0$. Basically, you run your favorite iteration and then set any negatives to zero before proceeding to the next iteration.

Simple 2D quadratic problem. Curves show contours of equal value of the cost function Ψ.

Same problem arises with upper bounds too.

The above problem applies to many simultaneous update iterative methods. For sequential update methods, such as coordinate descent, clipping works fine.

There are some simultaneous update iterative methods where it will work though; projected gradient descent with a positive-definite diagonal preconditioner, for example.

Newton-Raphson Algorithm

$$\boxed{\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - [\nabla^2 \Psi(\boldsymbol{x}^{(n)})]^{-1} \nabla \Psi(\boldsymbol{x}^{(n)})}$$

Advantage:
- Super-linear convergence rate (if convergent)

Disadvantages:
- Requires twice-differentiable Ψ
- Not guaranteed to converge
- Not guaranteed to monotonically decrease Ψ
- Does not enforce nonnegativity constraint
- Computing Hessian $\nabla^2 \Psi$ often expensive
- Impractical for image recovery due to matrix inverse

General purpose remedy: bound-constrained Quasi-Newton algorithms

$\nabla^2 \Psi(\boldsymbol{x})$ is called the *Hessian matrix*. It is a $n_p \times n_p$ matrix (where n_p is the dimension of $\boldsymbol{x}$). The j,kth element of it is $\frac{\partial^2}{\partial x_j \partial x_k} \Psi(\boldsymbol{x})$.

A "matrix inverse" actually is not necessary. One can rewrite the above iteration as $\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \boldsymbol{d}^{(n)}$ where $\boldsymbol{d}^{(n)}$ is the solution to the system of equations: $\nabla^2 \Psi(\boldsymbol{x}^{(n)}) \boldsymbol{d}^{(n)} = \nabla \Psi(\boldsymbol{x}^{(n)})$. Unfortunately, this is a non-sparse $n_p \times n_p$ system of equations, requiring $O(n_p^3)$ flops to solve, which is expensive. Instead of solving the system exactly one could use approximate iterative techniques, but then it should probably be considered a preconditioned gradient method rather than Newton-Raphson.

Quasi-Newton algorithms [173–176] [177, p. 136] [178, p. 77] [179, p. 63].

bound-constrained Quasi-Newton algorithms (LBFGS) [175, 180–183].

Newton's Quadratic Approximation

2nd-order Taylor series:

$$\Psi(\boldsymbol{x}) \approx \phi(\boldsymbol{x}; \boldsymbol{x}^{(n)}) \triangleq \Psi(\boldsymbol{x}^{(n)}) + \nabla \Psi(\boldsymbol{x}^{(n)}) (\boldsymbol{x} - \boldsymbol{x}^{(n)}) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{x}^{(n)})^T \nabla^2 \Psi(\boldsymbol{x}^{(n)}) (\boldsymbol{x} - \boldsymbol{x}^{(n)})$$

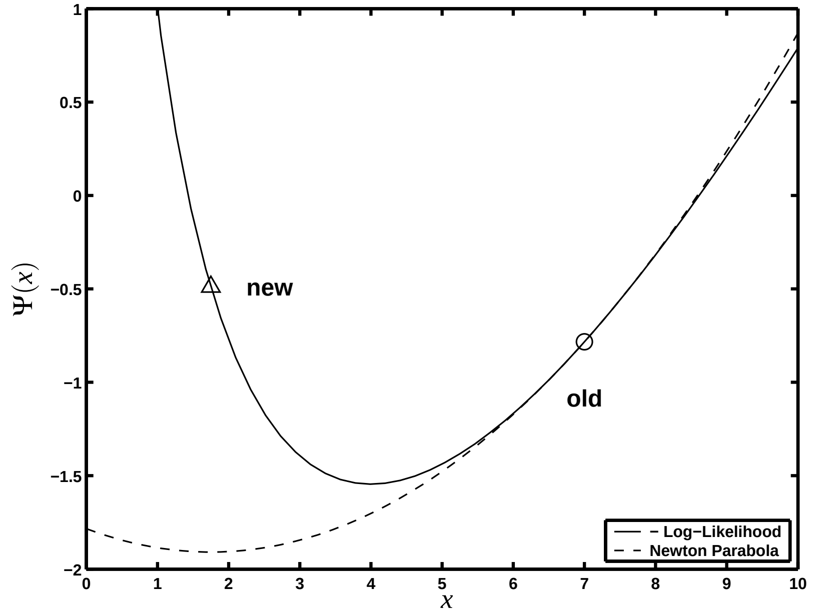
Set $\boldsymbol{x}^{(n+1)}$ to the ("easily" found) minimizer of this quadratic approximation:

$$\begin{aligned} \boldsymbol{x}^{(n+1)} &\triangleq \arg\min_{\boldsymbol{x}} \phi(\boldsymbol{x}; \boldsymbol{x}^{(n)}) \\ &= \boldsymbol{x}^{(n)} - [\nabla^2 \Psi(\boldsymbol{x}^{(n)})]^{-1} \nabla \Psi(\boldsymbol{x}^{(n)}) \end{aligned}$$

Can be nonmonotone for Poisson emission tomography log-likelihood,
even for a single pixel and single ray:

$$\Psi(x) = (x + r) - y \log(x + r).$$

Nonmonotonicity of Newton-Raphson



Consideration: Monotonicity

An algorithm is monotonic if

$$\Psi\left(\boldsymbol{x}^{(n+1)}\right) \leq \Psi\left(\boldsymbol{x}^{(n)}\right), \quad \forall \boldsymbol{x}^{(n)}.$$

Three categories of algorithms:
- Nonmonotonic (or unknown)
- Forced monotonic (*e.g.*, by line search)
- Intrinsically monotonic (by design, simplest to implement)

Forced monotonicity

Most nonmonotonic algorithms can be converted to forced monotonic algorithms by adding a line-search step:

$$\boldsymbol{x}^{\text{temp}} \triangleq \mathcal{M}\left(\boldsymbol{x}^{(n)}\right), \quad \boldsymbol{d} = \boldsymbol{x}^{\text{temp}} - \boldsymbol{x}^{(n)}$$

$$\boldsymbol{x}^{(n+1)} \triangleq \boldsymbol{x}^{(n)} - \alpha_n \boldsymbol{d}^{(n)} \text{ where } \alpha_n \triangleq \arg\min_{\alpha} \Psi\left(\boldsymbol{x}^{(n)} - \alpha \boldsymbol{d}^{(n)}\right)$$

Inconvenient, sometimes expensive, nonnegativity problematic.

Although monotonicity is not a necessary condition for an algorithm to converge globally to $\boldsymbol{x}^\star$, it is often the case that global convergence and monotonicity go hand in hand. In fact, for strictly convex Ψ, algorithms that monotonically decrease Ψ each iteration are guaranteed to converge under reasonable regularity conditions [184].

Any algorithm containing a line search step will have difficulties with nonnegativity. In principle one can address these problems using a “bent-line” search [185], but this can add considerable computation per iteration.

Conjugate Gradient (CG) Algorithm

Advantages:
- Fast converging (if suitably preconditioned) (in unconstrained case)
- Monotonic (forced by line search in nonquadratic case)
- Global convergence (unconstrained case)
- Flexible use of system matrix $\boldsymbol{A}$ and tricks
- Easy to implement in unconstrained quadratic case
- Highly parallelizable

Disadvantages:
- Nonnegativity constraint awkward (slows convergence?)
- Line-search somewhat awkward in nonquadratic cases
- Possible need to “restart” after many iterations

Highly recommended for unconstrained quadratic problems (*e.g.*, PWLS without nonnegativity). Useful (but perhaps not ideal) for Poisson case too.

CG is like steepest descent, but the search direction is modified each iteration to be conjugate to the previous search direction.

Preconditioners [186, 187]

Poisson case [91, 188, 189].

Efficient line-search for (nonquadratic) edge-preserving regularization described in [187].

Consideration: Parallelization

Simultaneous (fully parallelizable)
update all pixels simultaneously using all data
EM, Conjugate gradient, ISRA, OSL, SIRT, MART, ...

Block iterative (ordered subsets)
update (nearly) all pixels using one subset of the data at a time
OSEM, RBBI, ...

Row action
update many pixels using a single ray at a time
ART, RAMLA

Pixel grouped (multiple column action)
update some (but not all) pixels simultaneously a time, using all data
Grouped coordinate descent, multi-pixel SAGE
(Perhaps the most nontrivial to implement)

Sequential (column action)
update one pixel at a time, using all (relevant) data
Coordinate descent, SAGE

Sequential algorithms are the least parallelizable since one cannot update the second pixel until the first pixel has been updated (to preserve monotonicity and convergence properties).

SAGE [190, 191]
Grouped coordinate descent [192]
Multi-pixel SAGE [193]
RAMLA [194]
OSEM [26]
RBBI [195–197]
ISRA [198–200]
OSL [201, 202]

Coordinate Descent Algorithm

aka Gauss-Siedel, successive over-relaxation (SOR), iterated conditional modes (ICM)

Update one pixel at a time, holding others fixed to their most recent values:

$$x_j^{\text{new}} = \arg\min_{x_j \geq 0} \Psi\left(x_1^{\text{new}}, \ldots, x_{j-1}^{\text{new}}, x_j, x_{j+1}^{\text{old}}, \ldots, x_{n_p}^{\text{old}}\right), \qquad j = 1, \ldots, n_p$$

Advantages:
- Intrinsically monotonic
- Fast converging (from good initial image)
- Global convergence
- Nonnegativity constraint trivial

Disadvantages:
- Requires column access of system matrix $\boldsymbol{A}$
- Cannot exploit some "tricks" for $\boldsymbol{A}$, *e.g.*, factorizations
- Expensive "arg min" for nonquadratic problems
- Poorly parallelizable

Fast convergence shown by Sauer and Bouman with clever frequency-domain analysis [203].

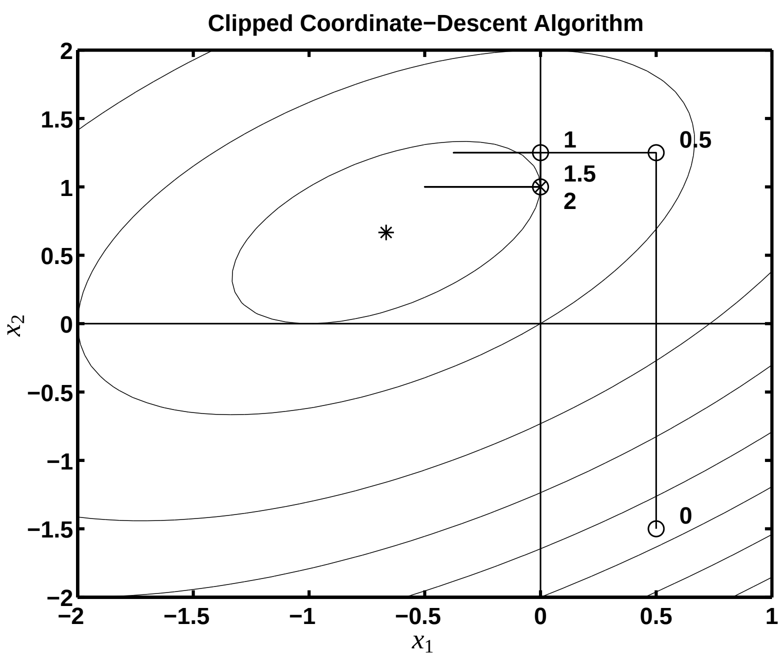
Any ordering can be used. Convergence rate may vary with ordering.

Global convergence even with negatives clipped [204].

One can replace the "arg min" with a one-dimensional Newton-Raphson step [192, 205–207]. However, this change then loses the guarantee of monotonicity for nonquadratic Ψ. Also, evaluating the second partial derivatives of Ψ with respect to x_j is expensive (costs an extra modified backprojection per iteration) [192].

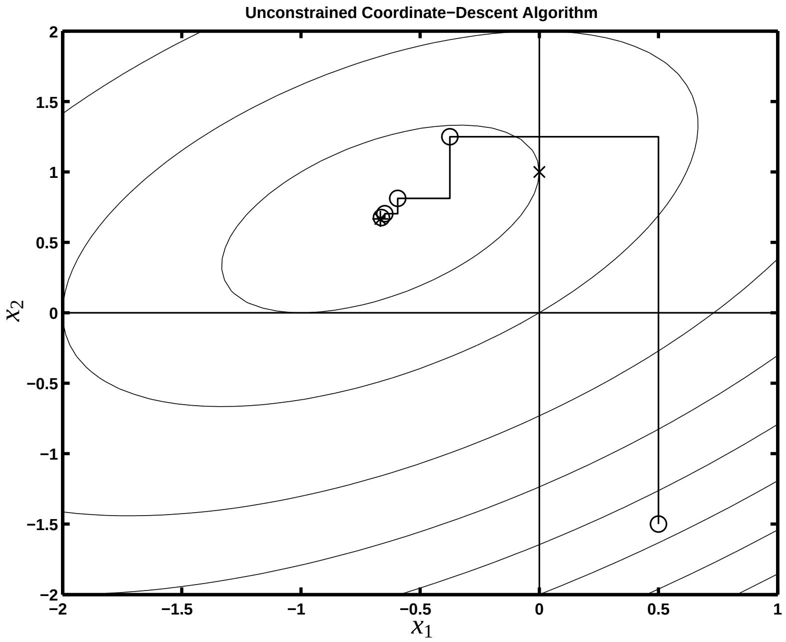
The paraboloidal surrogates coordinate descent (PSCD) algorithm circumvents these problems [208].

Constrained Coordinate Descent Illustrated



In this particular case, the nonnegativity constraint led to exact convergence in 1.5 iterations.

Coordinate Descent - Unconstrained



In general coordinate descent converges at a linear rate [84, 203].

Interestingly, for this particular problem the nonnegativity constraint accelerated convergence.

Coordinate-Descent Algorithm Summary

Recommended when all of the following apply:
- quadratic or nearly-quadratic convex cost function
- nonnegativity constraint desired
- precomputed and stored system matrix $\boldsymbol{A}$ with column access
- parallelization not needed (standard workstation)

Cautions:
- Good initialization (*e.g.*, properly scaled FBP) essential.
  (Uniform image or zero image cause slow initial convergence.)
- Must be programmed carefully to be efficient.
  (Standard Gauss-Siedel implementation is suboptimal.)
- Updates high-frequencies fastest $\Longrightarrow$ poorly suited to unregularized case

Used daily in UM clinic for 2D SPECT / PWLS / nonuniform attenuation

In saying “not good for the unregularized case” I am assuming one does not really wish to find the minimizer of Ψ in that case. If you really want the minimizer of Ψ in the unregularized case, then coordinate descent may still be useful.

Summary of General-Purpose Algorithms

Gradient-based
- Fully parallelizable
- Inconvenient line-searches for nonquadratic cost functions
- Fast converging in unconstrained case
- Nonnegativity constraint inconvenient

Coordinate-descent
- Very fast converging
- Nonnegativity constraint trivial
- Poorly parallelizable
- Requires precomputed/stored system matrix

CD is well-suited to moderate-sized 2D problem (*e.g.*, 2D PET),
but poorly suited to large 2D problems (X-ray CT) and fully 3D problems

Neither is ideal.

$\therefore$ need *special-purpose algorithms* for image reconstruction!

Interior-point methods for general-purpose constrained optimization have recently been applied to image reconstruction [209] and deserve further examination.

Data-Mismatch Functions Revisited

For fast converging, intrinsically monotone algorithms, consider the form of Ψ.

WLS:

$$Ł(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \frac{1}{2} w_i (y_i - [\boldsymbol{A}\boldsymbol{x}]_i)^2 = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i), \qquad \text{where } \mathsf{h}_i(l) \triangleq \frac{1}{2} w_i (y_i - l)^2.$$

Emission Poisson (negative) log-likelihood:

$$Ł(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} ([\boldsymbol{A}\boldsymbol{x}]_i + r_i) - y_i \log([\boldsymbol{A}\boldsymbol{x}]_i + r_i) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i)$$

$$\text{where } \mathsf{h}_i(l) \triangleq (l + r_i) - y_i \log(l + r_i)\,.$$

Transmission Poisson log-likelihood:

$$Ł(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \left(b_i \, \mathrm{e}^{-[\boldsymbol{A}\boldsymbol{x}]_i} + r_i\right) - y_i \log\left(b_i \, \mathrm{e}^{-[\boldsymbol{A}\boldsymbol{x}]_i} + r_i\right) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i)$$

$$\text{where } \mathsf{h}_i(l) \triangleq (b_i e^{-l} + r_i) - y_i \log\left(b_i e^{-l} + r_i\right).$$

MRI, polyenergetic X-ray CT, confocal microscopy, image restoration, ...
All have same *partially separable* form.

All the algorithms discussed this far are generic; they can be applied to any differentiable Ψ.

General Imaging Cost Function

General form for data-mismatch function:

$$\text{\L}(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i)$$

General form for regularizing penalty function:

$$R(\boldsymbol{x}) = \sum_k \psi_k([\boldsymbol{C}\boldsymbol{x}]_k)$$

General form for cost function:

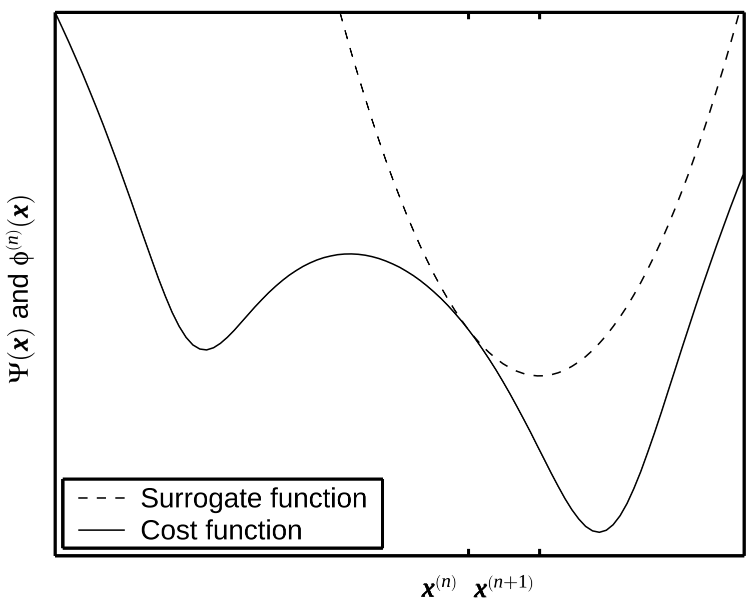
$$\boxed{\Psi(\boldsymbol{x}) = \text{\L}(\boldsymbol{x}) + \beta R(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i) + \beta \sum_k \psi_k([\boldsymbol{C}\boldsymbol{x}]_k)}$$

Properties of Ψ we can exploit:

- summation form (due to independence of measurements)
- convexity of h_i functions (usually)
- summation argument (inner product of $\boldsymbol{x}$ with ith row of $\boldsymbol{A}$)

Most methods that use these properties are forms of *optimization transfer*.

Optimization Transfer Illustrated



This figure does not do justice to the problem. A one-dimensional Ψ is usually easy to minimize. The problem is in multiple dimensions.

Optimization Transfer

General iteration:

$$\boxed{\boldsymbol{x}^{(n+1)} = \arg\min_{\boldsymbol{x}\geq \boldsymbol{0}} \phi\left(\boldsymbol{x}; \boldsymbol{x}^{(n)}\right)}$$

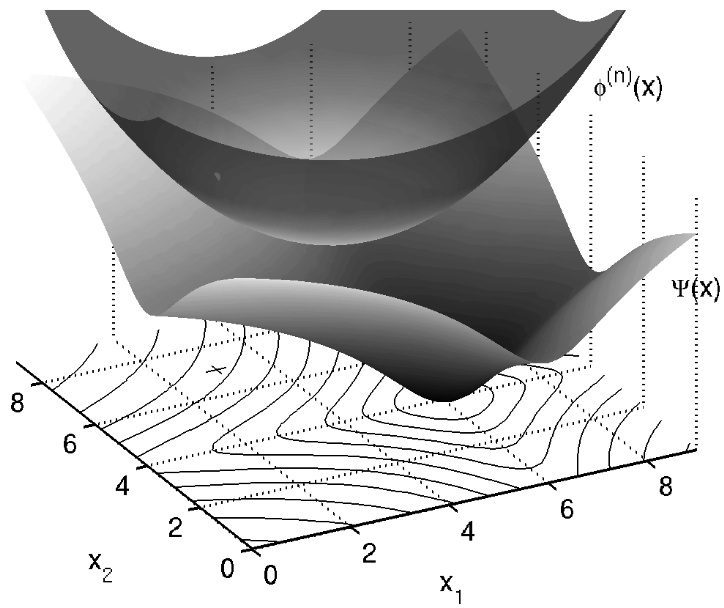
Monotonicity conditions (cost function Ψ decreases provided these hold):

- $\phi(\boldsymbol{x}^{(n)}; \boldsymbol{x}^{(n)}) = \Psi(\boldsymbol{x}^{(n)})$ (matched current value)
- $\nabla_{\boldsymbol{x}} \phi(\boldsymbol{x}; \boldsymbol{x}^{(n)}) \Big|_{\boldsymbol{x}=\boldsymbol{x}^{(n)}} = \nabla \Psi(\boldsymbol{x}) \Big|_{\boldsymbol{x}=\boldsymbol{x}^{(n)}}$ (matched gradient)
- $\phi(\boldsymbol{x}; \boldsymbol{x}^{(n)}) \geq \Psi(\boldsymbol{x}) \quad \forall \boldsymbol{x} \geq \boldsymbol{0}$ (lies above)

These 3 (sufficient) conditions are satisfied by the *Q* function of the EM algorithm (and its relatives like SAGE).

The 3rd condition is *not* satisfied by the Newton-Raphson quadratic approximation, which leads to its nonmonotonicity.

Optimization Transfer in 2d



Optimization Transfer cf EM Algorithm

E-step: choose surrogate function $\phi(\boldsymbol{x};\boldsymbol{x}^{(n)})$

M-step: minimize surrogate function

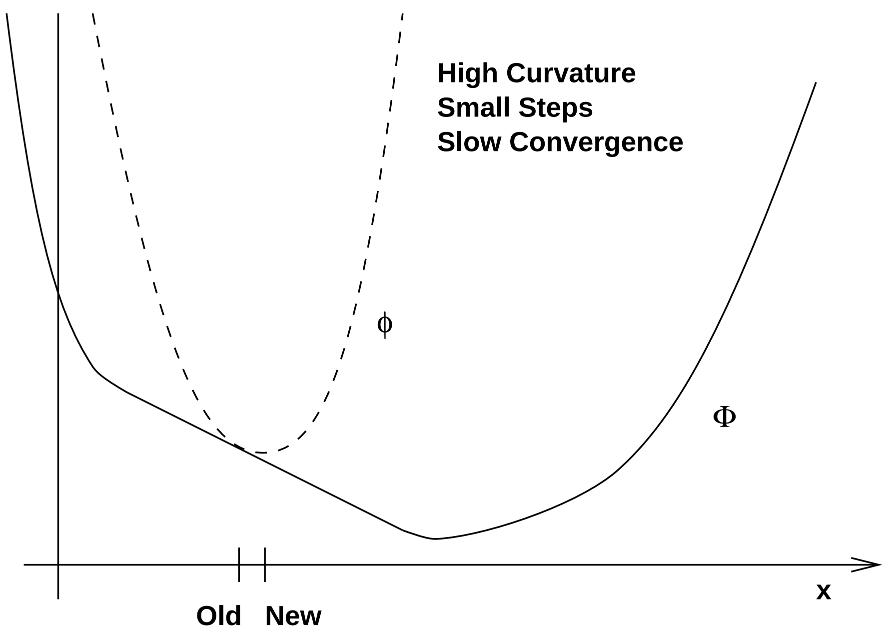
$$\boldsymbol{x}^{(n+1)} = \arg\min_{\boldsymbol{x} \geq \boldsymbol{0}} \phi\left(\boldsymbol{x};\boldsymbol{x}^{(n)}\right)$$

Designing surrogate functions
- Easy to "compute"
- Easy to minimize
- Fast convergence rate

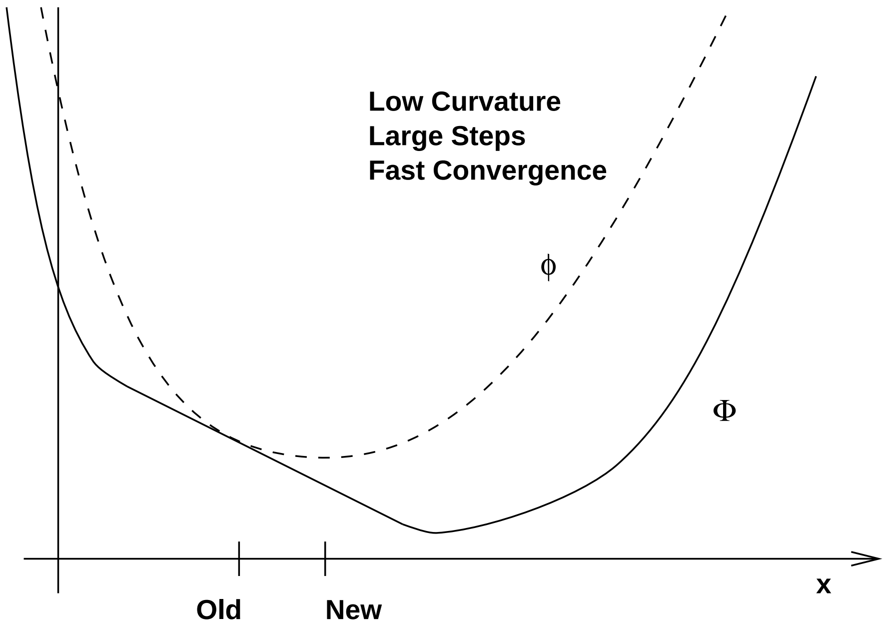
Often mutually incompatible goals $\therefore$ compromises

From the point of view of "per iteration convergence rate," the optimal "surrogate function" would be just Ψ itself. However, then the M-step is very difficult (in fact it is the original optimization problem). Such an "algorithm" would converge in one very expensive "iteration."

Convergence Rate: Slow

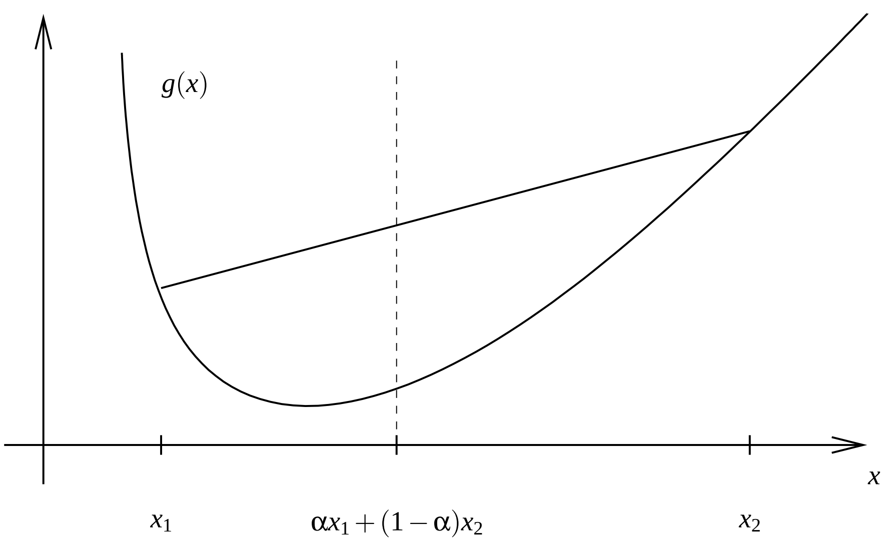


Convergence Rate: Fast



Tradeoff between curvature and ease of M-step... Can we beat this tradeoff?

Tool: Convexity Inequality



g convex $\Longrightarrow g(\alpha x_1 + (1-\alpha)x_2) \le \alpha g(x_1) + (1-\alpha) g(x_2)$ for $\alpha \in [0,1]$

More generally: $\alpha_k \ge 0$ and $\sum_k \alpha_k = 1 \Longrightarrow g(\sum_k \alpha_k x_k) \; \le \; \sum_k \alpha_k g(x_k)$. Sum outside!

The emission Poisson ray log-likelihood h_i is strictly convex on $(-r_i, \infty)$. This turns out to be adequate for the derivation.

Example 1: Classical ML-EM Algorithm

Negative Poisson log-likelihood cost function (unregularized):

$$\Psi(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i), \qquad \mathsf{h}_i(l) = (l + r_i) - y_i \log(l + r_i).$$

Intractable to minimize directly due to summation within logarithm.

Clever trick due to De Pierro (let $\bar{y}_i^{(n)} = [\boldsymbol{A}\boldsymbol{x}^{(n)}]_i + r_i$):

$$[\boldsymbol{A}\boldsymbol{x}]_i = \sum_{j=1}^{n_\mathrm{p}} a_{ij} x_j = \sum_{j=1}^{n_\mathrm{p}} \left[\frac{a_{ij} x_j^{(n)}}{\bar{y}_i^{(n)}}\right] \left(\frac{x_j}{x_j^{(n)}} \bar{y}_i^{(n)}\right).$$

Since the h_i's are *convex* in Poisson emission model:

$$\mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i) = \mathsf{h}_i\left(\sum_{j=1}^{n_\mathrm{p}} \left[\frac{a_{ij} x_j^{(n)}}{\bar{y}_i^{(n)}}\right] \left(\frac{x_j}{x_j^{(n)}} \bar{y}_i^{(n)}\right)\right) \le \sum_{j=1}^{n_\mathrm{p}} \left[\frac{a_{ij} x_j^{(n)}}{\bar{y}_i^{(n)}}\right] \mathsf{h}_i\left(\frac{x_j}{x_j^{(n)}} \bar{y}_i^{(n)}\right)$$

$$\Psi(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i) \le \phi(\boldsymbol{x}; \boldsymbol{x}^{(n)}) \triangleq \sum_{i=1}^{n_\mathrm{d}} \sum_{j=1}^{n_\mathrm{p}} \left[\frac{a_{ij} x_j^{(n)}}{\bar{y}_i^{(n)}}\right] \mathsf{h}_i\left(\frac{x_j}{x_j^{(n)}} \bar{y}_i^{(n)}\right)$$

Replace convex cost function $\Psi(\boldsymbol{x})$ with *separable* surrogate function $\phi(\boldsymbol{x}; \boldsymbol{x}^{(n)})$.

The clever (multiplicative) trick in the first equation is due to Alvaro De Pierro [200].

Note that the bracketed terms sum over j to unity.

I believe that this is the shortest and simplest possible derivation of the ML-EM algorithm, out of five distinct derivations I have seen.

This derivation is complete only for the case $r_i = 0$. It is easily generalized to $r_i \neq 0$.

"ML-EM Algorithm" M-step

E-step gave separable surrogate function:

$$\phi(\boldsymbol{x}; \boldsymbol{x}^{(n)}) = \sum_{j=1}^{n_\mathrm{p}} \phi_j(x_j; \boldsymbol{x}^{(n)}), \text{ where } \phi_j(x_j; \boldsymbol{x}^{(n)}) \triangleq \sum_{i=1}^{n_\mathrm{d}} \left[\frac{a_{ij} x_j^{(n)}}{\bar{y}_i^{(n)}}\right] \mathsf{h}_i\left(\frac{x_j}{x_j^{(n)}} \bar{y}_i^{(n)}\right).$$

M-step separates:

$$\boldsymbol{x}^{(n+1)} = \arg\min_{\boldsymbol{x} \ge \boldsymbol{0}} \phi(\boldsymbol{x}; \boldsymbol{x}^{(n)}) \Longrightarrow x_j^{(n+1)} = \arg\min_{x_j \ge 0} \phi_j(x_j; \boldsymbol{x}^{(n)}), \qquad j = 1, \ldots, n_\mathrm{p}$$

Minimizing:

$$\frac{\partial}{\partial x_j} \phi_j(x_j; \boldsymbol{x}^{(n)}) = \sum_{i=1}^{n_\mathrm{d}} a_{ij} \dot{\mathsf{h}}_i\left(\bar{y}_i^{(n)} x_j / x_j^{(n)}\right) = \sum_{i=1}^{n_\mathrm{d}} a_{ij} \left[1 - \frac{y_i}{\bar{y}_i^{(n)} x_j / x_j^{(n)}}\right] \Bigg|_{x_j = x_j^{(n+1)}} = 0.$$

Solving (in case $r_i = 0$):

$$\boxed{x_j^{(n+1)} = x_j^{(n)} \left[\sum_{i=1}^{n_\mathrm{d}} a_{ij} \frac{y_i}{[\boldsymbol{A}\boldsymbol{x}^{(n)}]_i}\right] \Big/ \left(\sum_{i=1}^{n_\mathrm{d}} a_{ij}\right), \qquad j = 1, \ldots, n_\mathrm{p}}$$

- Derived without any statistical considerations, unlike classical EM formulation.
- Uses only convexity and algebra.
- Guaranteed monotonic: surrogate function ϕ satisfies the 3 required properties.
- M-step trivial due to *separable surrogate*.

When $r_i = 0$, $\dot{\mathsf{h}}_i(l) \triangleq \frac{d}{dl} \mathsf{h}_i(l) = 1 - y_i / l$.

Case where $r_i \neq 0$ can also be handled with more algebra. Just replace final $[\boldsymbol{A}\boldsymbol{x}^{(n)}]_i$ with $\bar{y}_i^{(n)} = [\boldsymbol{A}\boldsymbol{x}^{(n)}]_i + r_i$.

To be rigorous, we should check that the Karush-Kuhn-Tucker condition holds for our minimizer of $\phi_j(\cdot; \boldsymbol{x}^{(n)})$. It does, provided $\boldsymbol{x}^{(n)} \ge \boldsymbol{0}$.

I prefer this derivation over the statistical EM derivation, even though we are doing statistical image reconstruction. Statistics greatly affect the design of Ψ, but minimizing Ψ is really just a numerical problem, and statistics need not have any role in that.

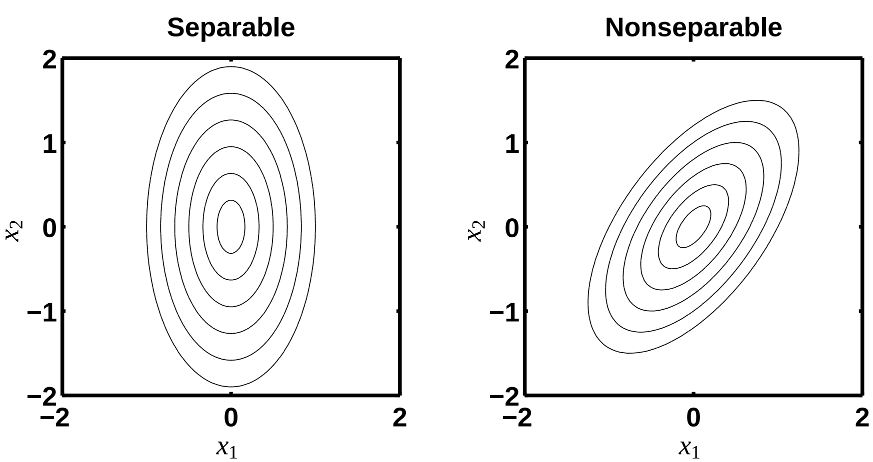
ML-EM is Scaled Gradient Descent

$$
\begin{aligned}
x_j^{(n+1)} &= x_j^{(n)} \left[\sum_{i=1}^{n_{\mathrm{d}}} a_{ij} \frac{y_i}{\bar{y}_i^{(n)}} \right] \Big/ \left(\sum_{i=1}^{n_{\mathrm{d}}} a_{ij} \right) \\
&= x_j^{(n)} + x_j^{(n)} \left[\sum_{i=1}^{n_{\mathrm{d}}} a_{ij} \left(\frac{y_i}{\bar{y}_i^{(n)}} - 1 \right) \right] \Big/ \left(\sum_{i=1}^{n_{\mathrm{d}}} a_{ij} \right) \\
&= \boxed{ x_j^{(n)} - \left(\frac{x_j^{(n)}}{\sum_{i=1}^{n_{\mathrm{d}}} a_{ij}} \right) \frac{\partial}{\partial x_j} \Psi\!\left(\boldsymbol{x}^{(n)}\right), } \qquad j = 1, \ldots, n_{\mathrm{p}}
\end{aligned}
$$

$$
\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} + \boldsymbol{D}(\boldsymbol{x}^{(n)}) \, \nabla \Psi\!\left(\boldsymbol{x}^{(n)}\right)
$$

This particular diagonal scaling matrix remarkably
- ensures monotonicity,
- ensures nonnegativity.

Consideration: Separable vs Nonseparable



Contour plots: loci of equal function values.

Uncoupled vs coupled minimization.

To find the minimizer of a separable function, one can minimize separately with respect to each argument. To find the minimizer of a nonseparable function, one must consider the variables together. In this sense the minimization problem “couples” together the unknown parameters.

Separable Surrogate Functions (Easy M-step)

The preceding EM derivation structure applies to *any* cost function of the form

$$\Psi(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i).$$

cf ISRA (for nonnegative LS), “convex algorithm” for transmission reconstruction

Derivation yields a separable surrogate function

$$\Psi(\boldsymbol{x}) \le \phi\left(\boldsymbol{x};\boldsymbol{x}^{(n)}\right), \quad \text{where} \quad \phi\left(\boldsymbol{x};\boldsymbol{x}^{(n)}\right) = \sum_{j=1}^{n_\mathrm{p}} \phi_j\left(x_j;\boldsymbol{x}^{(n)}\right)$$

M-step separates into 1D minimization problems (fully parallelizable):

$$\boldsymbol{x}^{(n+1)} = \arg\min_{\boldsymbol{x}\ge\boldsymbol{0}} \phi\left(\boldsymbol{x};\boldsymbol{x}^{(n)}\right) \Longrightarrow x_j^{(n+1)} = \arg\min_{x_j\ge 0} \phi_j\left(x_j;\boldsymbol{x}^{(n)}\right), \qquad j = 1,\ldots,n_\mathrm{p}$$

Why do EM / ISRA / convex-algorithm / etc. converge so slowly?

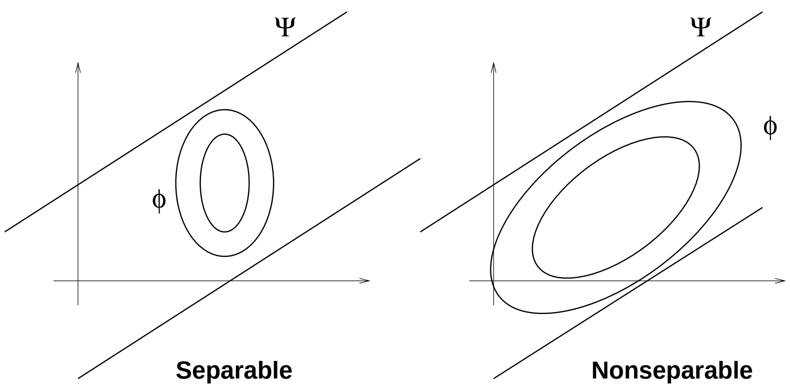
Unfortunately, choosing additively separable surrogate functions generally leads to very high curvature surrogates, which gives very slow convergence rates. EM is the classic example.

The classic EM algorithm is simple to implement precisely because it uses separable surrogate functions.

The derivation of the “convex algorithm” for the Poisson transmission problem [210] and the convergence proof of the ISRA algorithm [200] use a very similar derivation.

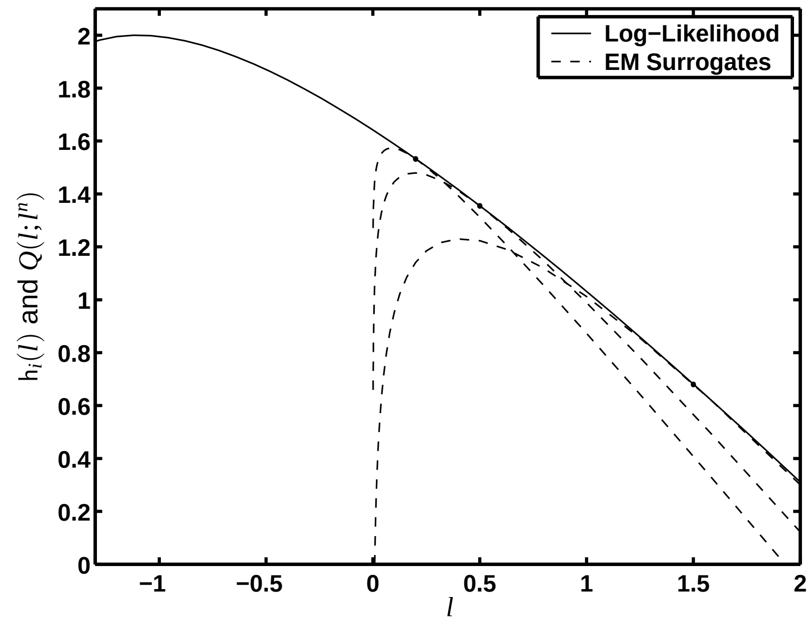
Clarify: the self-similar surrogate function is easy to minimize because it is separable. So even though L and Q are composed of the same ray-log likelihood functions, the latter is easier to minimize because it is separable.

Separable vs Nonseparable



Separable surrogates (*e.g.*, EM) have high curvature ∴ slow convergence.
Nonseparable surrogates can have lower curvature ∴ faster convergence.
Harder to minimize? Use paraboloids (quadratic surrogates).

High Curvature of EM Surrogate



Sublinear convergence rate of EM shown in [211].

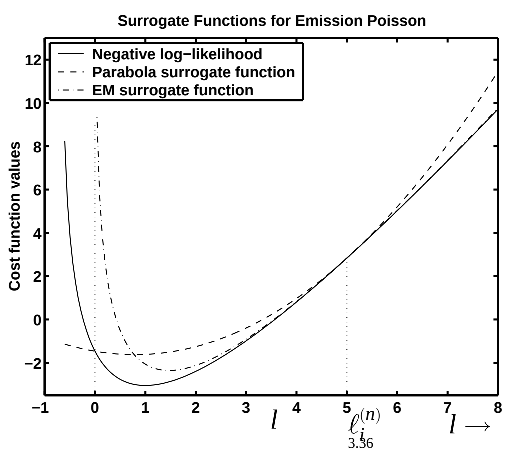
1D Parabola Surrogate Function

Find parabola $q_i^{(n)}(l)$ of the form:

$$q_i^{(n)}(l) = \mathsf{h}_i\left(\ell_i^{(n)}\right) + \dot{\mathsf{h}}_i\left(\ell_i^{(n)}\right)(l - \ell_i^{(n)}) + c_i^{(n)} \frac{1}{2}(l - \ell_i^{(n)})^2, \text{ where } \ell_i^{(n)} \triangleq [\boldsymbol{A}\boldsymbol{x}^{(n)}]_i$$

Satisfies tangent condition. Choose curvature to ensure "lies above" condition:

$$c_i^{(n)} \triangleq \min\left\{ c \geq 0 : q_i^{(n)}(l) \geq \mathsf{h}_i(l), \quad \forall l \geq 0 \right\}.$$



Lower curvature!

Paraboloidal Surrogate

Combining 1D parabola surrogates yields *paraboloidal surrogate*:

$$\Psi(\boldsymbol{x}) = \sum_{i=1}^{n_\mathrm{d}} \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i) \leq \phi\left(\boldsymbol{x}; \boldsymbol{x}^{(n)}\right) = \sum_{i=1}^{n_\mathrm{d}} q_i^{(n)}([\boldsymbol{A}\boldsymbol{x}]_i)$$

Rewriting: $\phi\left(\boldsymbol{\delta}+\boldsymbol{x}^{(n)}; \boldsymbol{x}^{(n)}\right) = \Psi\left(\boldsymbol{x}^{(n)}\right) + \nabla\Psi\left(\boldsymbol{x}^{(n)}\right)\boldsymbol{\delta} + \frac{1}{2}\boldsymbol{\delta}'\boldsymbol{A}'\,\mathrm{diag}\left\{c_i^{(n)}\right\}\boldsymbol{A}\boldsymbol{\delta}$

Advantages

- Surrogate $\phi(\boldsymbol{x}; \boldsymbol{x}^{(n)})$ is *quadratic*, unlike Poisson log-likelihood
  $\Longrightarrow$ easier to minimize
- Not separable (unlike EM surrogate)
- Not self-similar (unlike EM surrogate)
- Small curvatures $\Longrightarrow$ fast convergence
- Intrinsically monotone global convergence
- Fairly simple to derive / implement

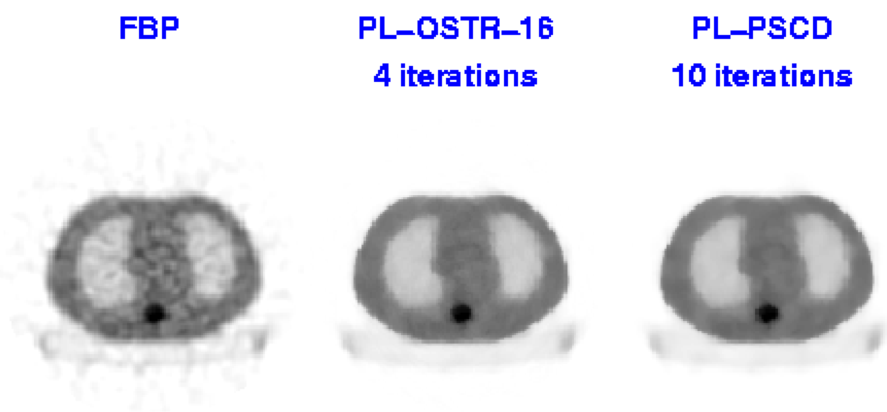
Quadratic minimization

- Coordinate descent
  - \+ fast converging
  - \+ Nonnegativity easy
  - \- precomputed column-stored system matrix
- Gradient-based quadratic minimization methods
  - \- Nonnegativity inconvenient

Instead of coordinate descent, one could also apply nonnegativity-constrained conjugate gradient.

PSCD recommended for 2D emission Poisson likelihood when system matrix precomputed and stored by columns.

Example: PSCD for PET Transmission Scans



- square-pixel basis
- strip-integral system model
- shifted-Poisson statistical model
- edge-preserving convex regularization (Huber)
- nonnegativity constraint
- inscribed circle support constraint
- paraboloidal surrogate coordinate descent (PSCD) algorithm

Separable Paraboloidal Surrogate

To derive a parallelizable algorithm apply another De Pierro trick:

$$[\boldsymbol{A}\boldsymbol{x}]_i = \sum_{j=1}^{n_p} \pi_{ij} \left[\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(n)}) + \ell_i^{(n)} \right], \qquad \ell_i^{(n)} = [\boldsymbol{A}\boldsymbol{x}^{(n)}]_i.$$

Provided $\pi_{ij} \geq 0$ and $\sum_{j=1}^{n_p} \pi_{ij} = 1$, since parabola q_i is convex:

$$q_i^{(n)}([\boldsymbol{A}\boldsymbol{x}]_i) = q_i^{(n)}\left(\sum_{j=1}^{n_p} \pi_{ij} \left[\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(n)}) + \ell_i^{(n)} \right] \right) \leq \sum_{j=1}^{n_p} \pi_{ij}\, q_i^{(n)}\left(\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(n)}) + \ell_i^{(n)} \right)$$

$$\therefore \phi(\boldsymbol{x};\boldsymbol{x}^{(n)}) = \sum_{i=1}^{n_d} q_i^{(n)}([\boldsymbol{A}\boldsymbol{x}]_i) \leq \tilde{\phi}(\boldsymbol{x};\boldsymbol{x}^{(n)}) \triangleq \sum_{i=1}^{n_d} \sum_{j=1}^{n_p} \pi_{ij}\, q_i^{(n)}\left(\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(n)}) + \ell_i^{(n)} \right)$$

Separable Paraboloidal Surrogate:

$$\tilde{\phi}(\boldsymbol{x};\boldsymbol{x}^{(n)}) = \sum_{j=1}^{n_p} \phi_j(x_j;\boldsymbol{x}^{(n)}), \qquad \phi_j(x_j;\boldsymbol{x}^{(n)}) \triangleq \sum_{i=1}^{n_d} \pi_{ij}\, q_i^{(n)}\left(\frac{a_{ij}}{\pi_{ij}} (x_j - x_j^{(n)}) + \ell_i^{(n)} \right)$$

Parallelizable M-step (cf gradient descent!):

$$x_j^{(n+1)} = \arg\min_{x_j \geq 0} \phi_j(x_j;\boldsymbol{x}^{(n)}) = \left[x_j^{(n)} - \frac{1}{d_j^{(n)}} \frac{\partial}{\partial x_j} \Psi(\boldsymbol{x}^{(n)}) \right]_+, \qquad d_j^{(n)} = \sum_{i=1}^{n_d} \frac{a_{ij}^2}{\pi_{ij}} c_i^{(n)}$$

Natural choice is $\pi_{ij} = |a_{ij}|/|a|_i$, $|a|_i = \sum_{j=1}^{n_p} |a_{ij}|$

De Pierro's "additive trick" in [117].

For the natural choice $\pi_{ij} = |a_{ij}|/|a|_i$, we have

$$d_j^{(n)} = \sum_{i=1}^{n_d} |a_{ij}|\, |a|_i\, c_i^{(n)}$$

Example: Poisson ML Transmission Problem

Transmission negative log-likelihood (for ith ray):

$$\mathsf{h}_i(l) = (b_i e^{-l} + r_i) - y_i \log(b_i e^{-l} + r_i).$$

Optimal (smallest) parabola surrogate curvature (Erdoğan, T-MI, Sep. 1999):

$$c_i^{(n)} = c(\ell_i^{(n)}, \mathsf{h}_i), \qquad c(l,h) = \begin{cases} \left[2\dfrac{h(0) - h(l) + \dot{h}(l)l}{l^2} \right]_+, & l > 0 \\ \left[\ddot{h}(l) \right]_+, & l = 0. \end{cases}$$

Separable Paraboloidal Surrogate (SPS) Algorithm:

Precompute $|a|_i = \sum_{j=1}^{n_p} a_{ij}, \qquad i = 1, \ldots, n_d$

$$\begin{aligned} \ell_i^{(n)} &= [\boldsymbol{A}\boldsymbol{x}^{(n)}]_i, && \text{(forward projection)} \\ \bar{y}_i^{(n)} &= b_i e^{-\ell_i^{(n)}} + r_i && \text{(predicted means)} \\ \dot{\mathsf{h}}_i^{(n)} &= 1 - y_i/\bar{y}_i^{(n)} && \text{(slopes)} \\ c_i^{(n)} &= c(\ell_i^{(n)}, \mathsf{h}_i) && \text{(curvatures)} \end{aligned}$$

$$x_j^{(n+1)} = \left[x_j^{(n)} - \frac{1}{d_j^{(n)}} \frac{\partial}{\partial x_j} \Psi(\boldsymbol{x}^{(n)}) \right]_+ = \left[x_j^{(n)} - \frac{\sum_{i=1}^{n_d} a_{ij} \dot{\mathsf{h}}_i^{(n)}}{\sum_{i=1}^{n_d} a_{ij} |a|_i c_i^{(n)}} \right]_+, \qquad j = 1, \ldots, n_p$$

Monotonically decreases cost function each iteration. No logarithm!

Note that this algorithm never takes the logarithm of the transmission data, since it is based directly on a statistical model for the raw measurements. This is a significant part of the reason why it works well for low-count measurements.

Optimal parabola surrogate curvature for transmission problem [208]. Emission problem [212].

A Matlab m-file for this algorithm is available from
`http://www.eecs.umich.edu/~fessler /code`
as `transmission/tml_sps.m`

Related m-files also of interest include `transmission/tpl_osps.m`

The MAP-EM M-step "Problem"

Add a penalty function to our surrogate for the negative log-likelihood:

$$\Psi(\boldsymbol{x}) = \text{Ł}(\boldsymbol{x}) + \beta R(\boldsymbol{x})$$

$$\phi\left(\boldsymbol{x};\boldsymbol{x}^{(n)}\right) = \sum_{j=1}^{n_p} \phi_j\left(x_j;\boldsymbol{x}^{(n)}\right) + \beta R(\boldsymbol{x})$$

M-step: $\boldsymbol{x}^{(n+1)} = \arg\min_{\boldsymbol{x}\geq\boldsymbol{0}} \phi\left(\boldsymbol{x};\boldsymbol{x}^{(n)}\right) = \arg\min_{\boldsymbol{x}\geq\boldsymbol{0}} \sum_{j=1}^{n_p} \phi_j\left(x_j;\boldsymbol{x}^{(n)}\right) + \beta R(\boldsymbol{x}) = ?$

For nonseparable penalty functions, the M-step is coupled ∴ difficult.

Suboptimal solutions

- Generalized EM (GEM) algorithm (coordinate descent on ϕ)
  Monotonic, but inherits slow convergence of EM.
- One-step late (OSL) algorithm (use outdated gradients) (Green, T-MI, 1990)

$$\frac{\partial}{\partial x_j}\phi(\boldsymbol{x};\boldsymbol{x}^{(n)}) = \frac{\partial}{\partial x_j}\phi_j(x_j;\boldsymbol{x}^{(n)}) + \beta\frac{\partial}{\partial x_j}R(\boldsymbol{x}) \overset{?}{\approx} \frac{\partial}{\partial x_j}\phi_j(x_j;\boldsymbol{x}^{(n)}) + \beta\frac{\partial}{\partial x_j}R(\boldsymbol{x}^{(n)})$$

  Nonmonotonic. Known to diverge, depending on β.
  Temptingly simple, but *avoid!*

Contemporary solution

- Use separable surrogate for penalty function too (De Pierro, T-MI, Dec. 1995)
  Ensures monotonicity. Obviates all reasons for using OSL!

OSL [201, 202]
GEM [213–215]

De Pierro's separable penalty derived in [117].

De Pierro's MAP-EM Algorithm

Apply separable paraboloidal surrogates to penalty function:

$$R(\boldsymbol{x}) \leq R_{\text{SPS}}(\boldsymbol{x};\boldsymbol{x}^{(n)}) = \sum_{j=1}^{n_p} R_j(x_j;\boldsymbol{x}^{(n)})$$

Overall separable surrogate: $\phi\left(\boldsymbol{x};\boldsymbol{x}^{(n)}\right) = \sum_{j=1}^{n_p} \phi_j\left(x_j;\boldsymbol{x}^{(n)}\right) + \beta \sum_{j=1}^{n_p} R_j(x_j;\boldsymbol{x}^{(n)})$

The M-step becomes fully parallelizable:

$$x_j^{(n+1)} = \arg\min_{x_j\geq 0} \phi_j\left(x_j;\boldsymbol{x}^{(n)}\right) - \beta R_j(x_j;\boldsymbol{x}^{(n)}), \qquad j = 1,\ldots,n_p.$$

Consider quadratic penalty $R(\boldsymbol{x}) = \sum_k \psi([\boldsymbol{C}\boldsymbol{x}]_k)$, where $\psi(t) = t^2/2$.
If $\gamma_{kj} \geq 0$ and $\sum_{j=1}^{n_p} \gamma_{kj} = 1$ then

$$[\boldsymbol{C}\boldsymbol{x}]_k = \sum_{j=1}^{n_p} \gamma_{kj} \left[\frac{c_{kj}}{\gamma_{kj}}(x_j - x_j^{(n)}) + [\boldsymbol{C}\boldsymbol{x}^{(n)}]_k\right].$$

Since ψ is convex:

$$\begin{aligned}\psi([\boldsymbol{C}\boldsymbol{x}]_k) &= \psi\left(\sum_{j=1}^{n_p} \gamma_{kj} \left[\frac{c_{kj}}{\gamma_{kj}}(x_j - x_j^{(n)}) + [\boldsymbol{C}\boldsymbol{x}^{(n)}]_k\right]\right) \\ &\leq \sum_{j=1}^{n_p} \gamma_{kj}\, \psi\left(\frac{c_{kj}}{\gamma_{kj}}(x_j - x_j^{(n)}) + [\boldsymbol{C}\boldsymbol{x}^{(n)}]_k\right)\end{aligned}$$

Often we just choose

$$\gamma_{kj} = \begin{cases} \frac{1}{\text{number of nonzero } c_{kj}\text{'s in } k\text{th row of } \boldsymbol{C}}, & c_{kj} \neq 0 \\ 0, & \text{otherwise,}\end{cases}$$

which satisfies the two conditions $\gamma_{kj} \geq 0$ and $\sum_{j=1}^{n_p} \gamma_{kj} = 1$, *e.g.*

$$\boldsymbol{C} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 1 \end{bmatrix}, \quad \{\gamma_{kj}\} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} \end{bmatrix}.$$

Alternatively we use the choice

$$\gamma_{kj} = \frac{|c_{kj}|}{\sum_{j'=1}^{n_p} |c_{kj'}|},$$

which happens to yield the same result when the elements of $\boldsymbol{C}$ are just ± 1 as in the above example. For non-unity c_{kj}'s, the latter ratio seems to be preferable in terms of convergence rate.

De Pierro's Algorithm Continued

So $R(\boldsymbol{x}) \leq R(\boldsymbol{x};\boldsymbol{x}^{(n)}) \triangleq \sum_{j=1}^{n_p} R_j(x_j;\boldsymbol{x}^{(n)})$ where

$$R_j(x_j;\boldsymbol{x}^{(n)}) \triangleq \sum_k \gamma_{kj}\,\psi\left(\frac{c_{kj}}{\gamma_{kj}}(x_j - x_j^{(n)}) + [\boldsymbol{C}\boldsymbol{x}^{(n)}]_k\right)$$

M-step: Minimizing $\phi_j(x_j;\boldsymbol{x}^{(n)}) + \beta R_j(x_j;\boldsymbol{x}^{(n)})$ yields the iteration:

$$x_j^{(n+1)} = \frac{x_j^{(n)} \sum_{i=1}^{n_d} a_{ij} y_i / \bar{y}_i^{(n)}}{B_j + \sqrt{B_j^2 + \left(x_j^{(n)} \sum_{i=1}^{n_d} a_{ij} y_i / \bar{y}_i^{(n)}\right)\left(\beta \sum_k c_{kj}^2 / \gamma_{kj}\right)}}$$

where $B_j \triangleq \frac{1}{2}\left[\sum_{i=1}^{n_d} a_{ij} + \beta \sum_k \left(c_{kj}[\boldsymbol{C}\boldsymbol{x}^{(n)}]_k - \frac{c_{kj}^2}{\gamma_{kj}} x_j^{(n)}\right)\right], \qquad j = 1,\ldots,n_p$

and $\bar{y}_i^{(n)} = [\boldsymbol{A}\boldsymbol{x}^{(n)}]_i + r_i$.

Advantages: Intrinsically monotone, nonnegativity, fully parallelizable.
Requires only a couple % more computation per iteration than ML-EM

Disadvantages: Slow convergence (like EM) due to separable surrogate

---

As a concrete example, consider $R(\boldsymbol{x}) = \sum_{j=1}^{n_p} \frac{1}{2} \sum_{k\in\mathcal{N}_j} \frac{1}{2}(x_j - x_k)^2$ with $\mathcal{N}_j$ corresponding to the $|\mathcal{N}_j|$ nearest neighbors to the jth pixel. For this penalty with the choice $\gamma_{kj} = |c_{kj}|/c_k$ where $c_k = \sum_{j=1}^{n_p} |c_{kj}| = |\mathcal{N}_j|$, the separable surrogate is [117]:

$$R_j(x_j;\boldsymbol{x}^{(n)}) = \sum_{k\in\mathcal{N}_j} \frac{1}{2}\frac{1}{|\mathcal{N}_j|}\left(|\mathcal{N}_j|(x_j - x_j^{(n)}) + x_j^{(n)} - x_j^{(k)}\right)^2.$$

Matlab m-file available from `http://www.eecs.umich.edu/~fessler /code` as `emission/eql_emdp.m`

Caution: use stable quadratic roots [169] (slightly more complicated than above).

One can make an ordered-subsets version of De Pierro's MAP-EM easily. Such an approach is preferable to the OSL version of OS-EM mentioned by Hudson and Larkin [26].

One can do multiple M-step subiterations for minimal additional computation with some improvement in convergence rate.

For a tomography problem with a 64×64 image, 64×80 sinogram, and strip-area system matrix, De Pierro's MAP-EM algorithm requires 4% more flops per iteration than classic ML-EM.

Ordered Subsets Algorithms

aka *block iterative* or *incremental gradient* algorithms

The gradient appears in essentially every algorithm:

$$Ł(\boldsymbol{x}) = \sum_{i=1}^{n_d} h_i([\boldsymbol{A}\boldsymbol{x}]_i) \Longrightarrow \frac{\partial}{\partial x_j} Ł(\boldsymbol{x}) = \sum_{i=1}^{n_d} a_{ij}\,\dot{h}_i([\boldsymbol{A}\boldsymbol{x}]_i).$$

This is a *backprojection* of a sinogram of the derivatives $\left\{\dot{h}_i([\boldsymbol{A}\boldsymbol{x}]_i)\right\}$.

Intuition: with half the angular sampling, this backprojection would be fairly similar

$$\frac{1}{n_d}\sum_{i=1}^{n_d} a_{ij}\,\dot{h}_i(\cdot) \approx \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} a_{ij}\,\dot{h}_i(\cdot),$$

where $\mathcal{S}$ is a subset of the rays.

To "OS-ize" an algorithm, replace all backprojections with partial sums.

Recall typical iteration:

$$\boldsymbol{x}^{(n+1)} = \boldsymbol{x}^{(n)} - \boldsymbol{D}(\boldsymbol{x}^{(n)})\,\nabla\Psi\left(\boldsymbol{x}^{(n)}\right).$$

The dramatic improvements in apparent "convergence rate" of OSEM over classic ML-EM are due largely to the fact that the latter converges so slowly.

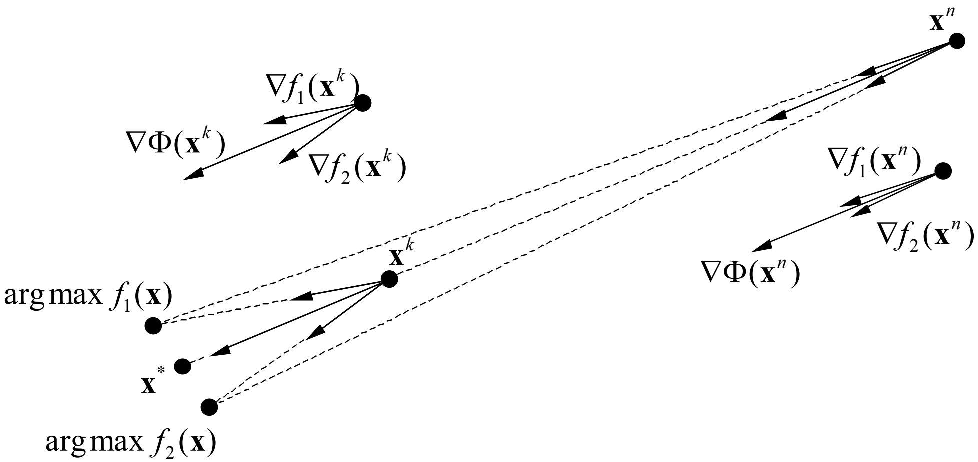
Modern, faster converging algorithms may benefit much less from OS modifications.

Richard Larkin (personal communication) has described the development of OSEM as something of a fortuitous programming "accident." In the course of developing software to implement the E-ML-EM algorithm, he first implemented a version that updated the image immediately after the reprojection of each view. Later he implemented the classical E-ML-EM algorithm but found it to give worse images (in the early iterations). (Due of course to its slow convergence.) The "immediate update" version turns out to be OSEM with 1 view per subset.

Several publications hinted at the use of subsets of projection views for acceleration, *e.g.*, [216–219], and D. Politte's 1983 dissertation. But it was the paper by Larking and Hudson that incited widespread use of OSEM [26].

In the general optimization literature, such algorithms are called *incremental gradient* methods [220–224], and they date back to the 1970's [225].

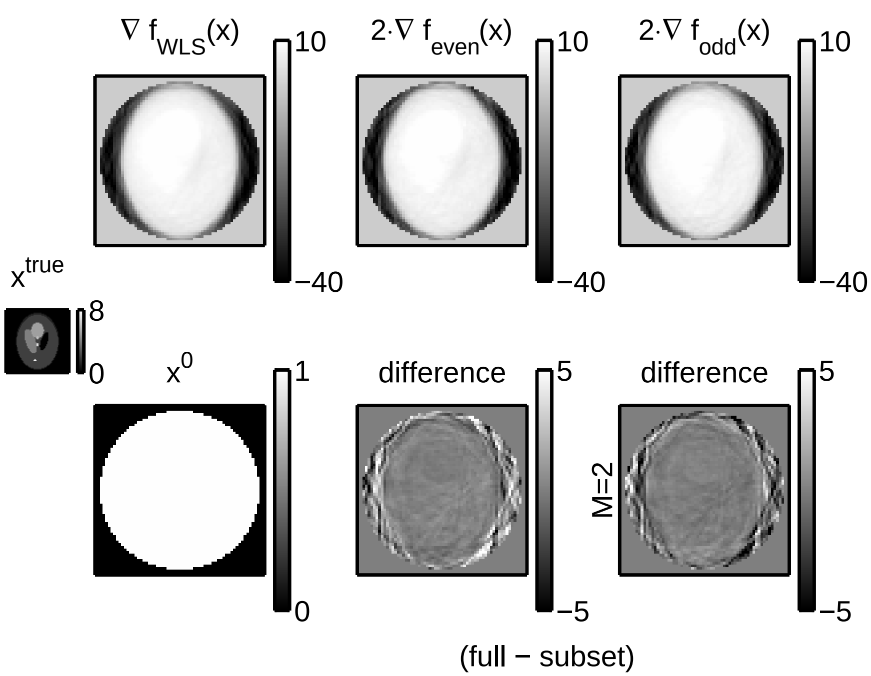
Geometric View of Ordered Subsets



Two subset case: $\Psi(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})$ (*e.g.*, odd and even projection views).

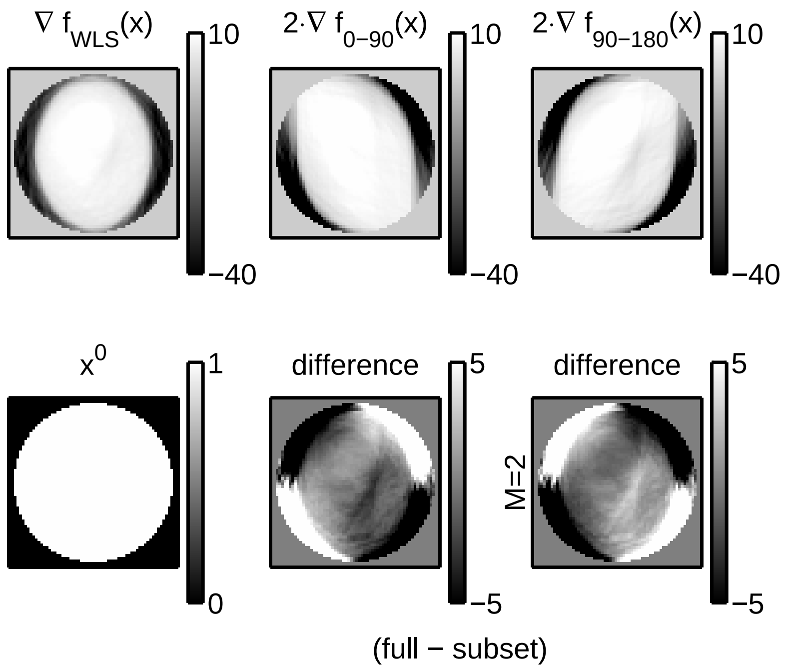
For $\boldsymbol{x}^{(n)}$ far from $\boldsymbol{x}^\star$, even partial gradients should point roughly towards $\boldsymbol{x}^\star$.
For $\boldsymbol{x}^{(n)}$ near $\boldsymbol{x}^\star$, however, $\nabla\Psi(\boldsymbol{x}) \approx \mathbf{0}$, so $\nabla f_1(\boldsymbol{x}) \approx -\nabla f_2(\boldsymbol{x}) \Longrightarrow$ cycles!
Issues. "Subset gradient balance": $\nabla\Psi(\boldsymbol{x}) \approx M \nabla f_k(\boldsymbol{x})$. Choice of ordering.

Incremental Gradients (WLS, 2 Subsets)



Here the initial image $\boldsymbol{x}^{(0)}$ is far from the solution so the incremental gradients, *i.e.*, the gradients computed from just the even or odd angles, agree well with the full gradient.

Subset Gradient Imbalance



Here the first subset was angles 0-90°, and the second subset was angles 90-180°, roughly speaking. Now the incremental gradients do not agree as well with the full gradient. (Of course the *sum* of the two incremental gradients would still equal the full gradient.) This imbalance is expected to slow “convergence.”

Problems with OS-EM

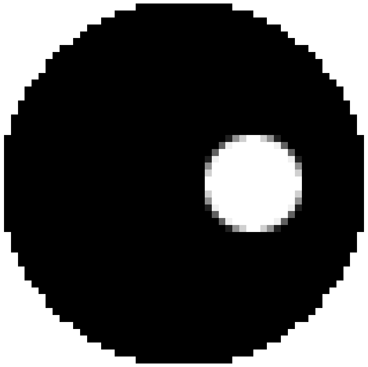
- Non-monotone
- Does not converge (may cycle)
- Byrne’s “rescaled block iterative” (RBI) approach converges only for consistent (noiseless) data
- ∴ unpredictable
  - What resolution after n iterations?
    Object-dependent, spatially nonuniform
  - What variance after n iterations?
  - ROI variance? (*e.g.*, for Huesman’s WLS kinetics)

RBI (rescaled block iterative) [197].

Soares and Glick *et al.* [41] [40] have extended the work of Barrett *et al.* [33] to the OSEM case.

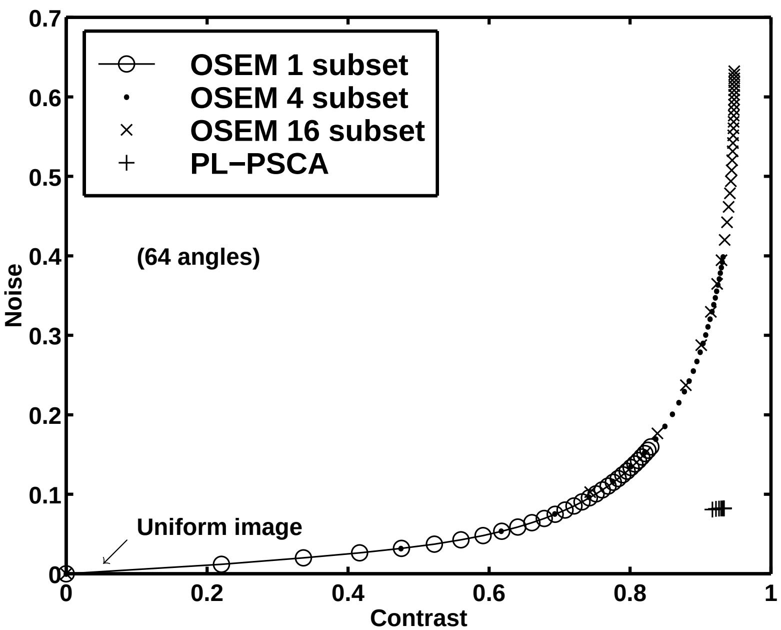
Wang *et al.* have extended it to the penalized case, for the OSL algorithm [39].

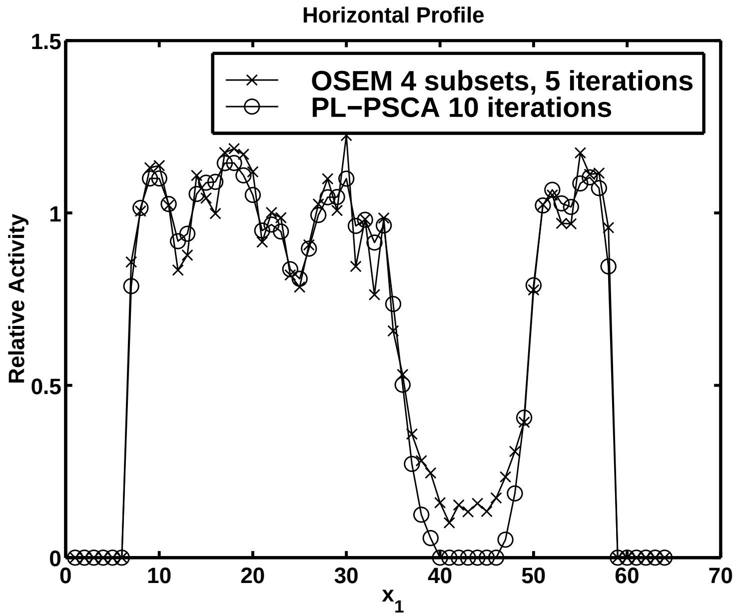
OSEM vs Penalized Likelihood



- 64×62 image
- 66×60 sinogram
- 10^6 counts
- 15% randoms/scatter
- uniform attenuation
- contrast in cold region
- within-region σ opposite side

Contrast-Noise Results





Making OS Methods Converge

- Relaxation
- Incrementalism

Relaxed block-iterative methods

$$\Psi(\boldsymbol{x}) = \sum_{m=1}^{M} \Psi_m(\boldsymbol{x})$$

$$\boldsymbol{x}^{(n+(m+1)/M)} = \boldsymbol{x}^{(n+m/M)} - \alpha_n D(\boldsymbol{x}^{(n+m/M)}) \nabla \Psi_m\left(\boldsymbol{x}^{(n+m/M)}\right), \qquad m = 0, \ldots, M-1$$

Relaxation of step sizes:

$$\alpha_n \to 0 \text{ as } n \to \infty, \qquad \sum_n \alpha_n = \infty, \qquad \sum_n \alpha_n^2 < \infty$$

- ART
- RAMLA, BSREM (De Pierro, T-MI, 1997, 2001)
- Ahn and Fessler, NSS/MIC 2001, T-MI 2003

Considerations

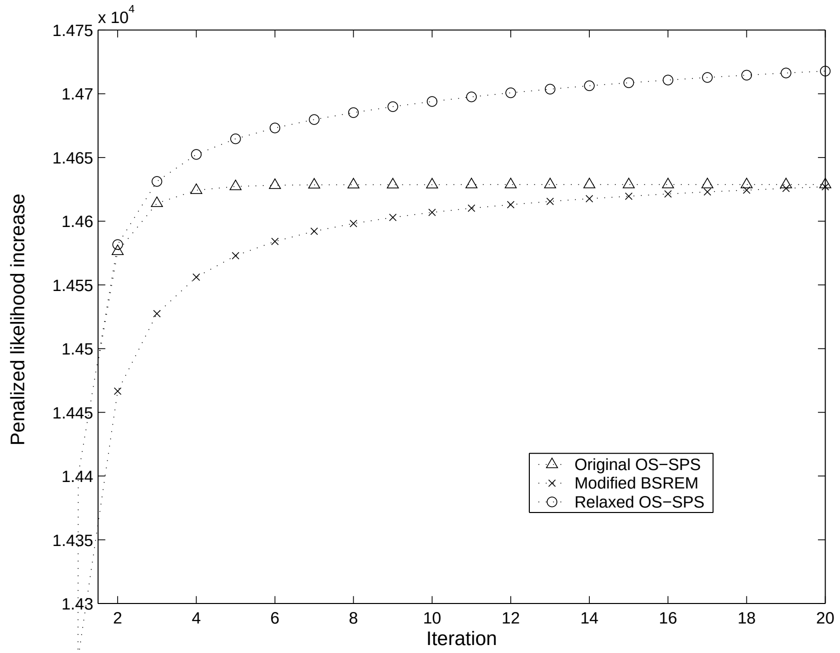
- Proper relaxation can induce convergence, *but* still lacks monotonicity.
- Choice of relaxation schedule requires experimentation.
- $\Psi_m(\boldsymbol{x}) = Ł_m(\boldsymbol{x}) + \frac{1}{M} R(\boldsymbol{x})$, so each Ψ_m includes part of the likelihood yet all of R

RAMLA [194] (for ML only)

Kudo [226] does not give convergence proof in English...

BSREM [227] convergence proof requires some “a posteriori” assumptions. These have been eliminated in [228].

Relaxed OS-SPS



[229]

Incremental Methods

Incremental EM applied to emission tomography by Hsiao *et al.* as C-OSEM

Incremental optimization transfer (Ahn & Fessler, MIC 2004)

Find majorizing surrogate for each sub-objective function:

$$\begin{aligned} \phi_m(\boldsymbol{x};\boldsymbol{x}) &= \Psi_m(\boldsymbol{x}), \qquad \forall \boldsymbol{x} \\ \phi_m(\boldsymbol{x};\bar{\boldsymbol{x}}) &\geq \Psi_m(\boldsymbol{x}), \qquad \forall \boldsymbol{x}, \bar{\boldsymbol{x}} \end{aligned}$$

Define the following augmented cost function: $F(\boldsymbol{x};\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_M) = \sum_{m=1}^M \phi_m(\boldsymbol{x};\bar{\boldsymbol{x}}_m)$.
Fact: by construction $\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}) = \arg\min_{\boldsymbol{x}} \min_{\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_M} F(\boldsymbol{x};\bar{\boldsymbol{x}}_1,\ldots,\bar{\boldsymbol{x}}_M)$.

Alternating minimization: `for` $m = 1,\ldots,M$:

$$\begin{aligned} \boldsymbol{x}^{\text{new}} &= \arg\min_{\boldsymbol{x}} F\left(\boldsymbol{x};\bar{\boldsymbol{x}}_1^{(n+1)},\ldots,\bar{\boldsymbol{x}}_{m-1}^{(n+1)},\bar{\boldsymbol{x}}_m^{(n)},\bar{\boldsymbol{x}}_{m+1}^{(n)},\ldots\bar{\boldsymbol{x}}_M^{(n)}\right) \\ \bar{\boldsymbol{x}}_m^{(n+1)} &= \arg\min_{\bar{\boldsymbol{x}}_m} F\left(\boldsymbol{x}^{\text{new}};\bar{\boldsymbol{x}}_1^{(n+1)},\ldots,\bar{\boldsymbol{x}}_{m-1}^{(n+1)},\bar{\boldsymbol{x}}_m,\bar{\boldsymbol{x}}_{m+1}^{(n)},\ldots\bar{\boldsymbol{x}}_M^{(n)}\right) = \boldsymbol{x}^{\text{new}}. \end{aligned}$$

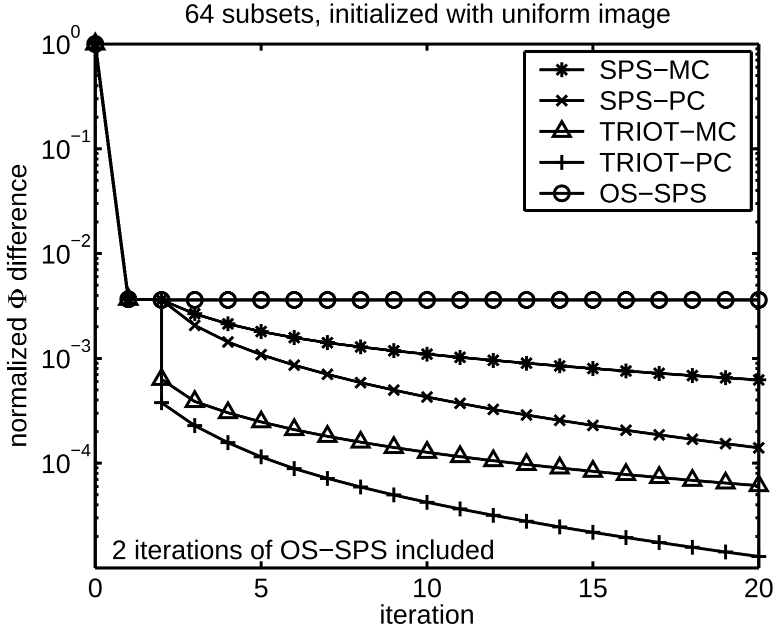
- *Use all current information, but increment the surrogate for only one subset.*
- Monotone in F, converges under reasonable assumptions on Ψ
- In constrast, OS-EM uses the information from *only* one subset at a time

C-OSEM [230]

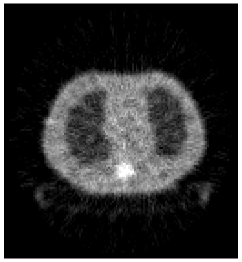
Incremental optimization transfer [231, 232]

TRIOT Example: Convergence Rate

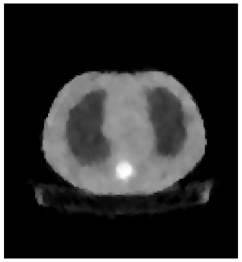
Transmission incremental optimization transfer (TRIOT)



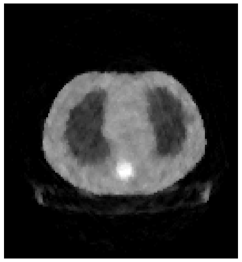
TRIOT Example: Attenuation Map Images



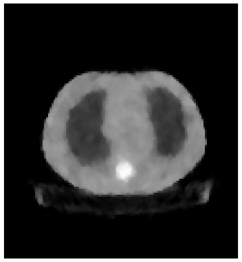
FBP



PL optimal image



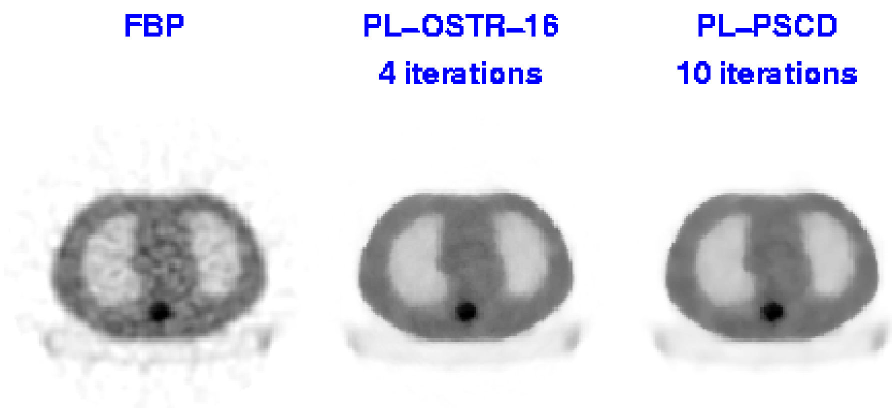
OS-SPS



TRIOT-PC

OS-SPS: 64 subsets, 20 iterations, one point of the limit cycle
TRIOT-PC: 64 subsets, 20 iterations, after 2 iterations of OS-SPS)

OSTR aka Transmission OS-SPS



Ordered subsets version of separable paraboloidal surrogates
for PET transmission problem with nonquadratic convex *regularization*

Matlab m-file `http://www.eecs.umich.edu/~fessler /code/transmission/tpl_osps.m`

Ordered subsets for PET/SPECT transmission scans [233], and for X-ray CT [46].

Precomputed curvatures for OS-SPS

Separable Paraboloidal Surrogate (SPS) Algorithm:

$$x_j^{(n+1)} = \left[x_j^{(n)} - \frac{\sum_{i=1}^{n_\mathrm{d}} a_{ij} \dot{\mathsf{h}}_i([\boldsymbol{A}\boldsymbol{x}^{(n)}]_i)}{\sum_{i=1}^{n_\mathrm{d}} a_{ij} |a|_i c_i^{(n)}} \right]_+, \qquad j = 1, \ldots, n_\mathrm{p}$$

Ordered-subsets abandons monotonicity, so why use optimal curvatures $c_i^{(n)}$?

Precomputed curvature:

$$c_i = \ddot{\mathsf{h}}_i\big(\hat{l}_i\big), \qquad \hat{l}_i = \arg\min_l \mathsf{h}_i(l)$$

Precomputed denominator (saves one backprojection each iteration!):

$$d_j = \sum_{i=1}^{n_\mathrm{d}} a_{ij} |a|_i c_i, \qquad j = 1, \ldots, n_\mathrm{p}.$$

OS-SPS algorithm with M subsets:

$$x_j^{(n+1)} = \left[x_j^{(n)} - \frac{\sum_{i \in \mathcal{S}^{(n)}} a_{ij} \dot{\mathsf{h}}_i([\boldsymbol{A}\boldsymbol{x}^{(n)}]_i)}{d_j / M} \right]_+, \qquad j = 1, \ldots, n_\mathrm{p}$$

Precomputed parabola surrogate curvature for transmission problem and ordered subsets [208, 233].

For emission problem, $c_i \approx 1/y_i$.

For transmission problem, $c_i \approx y_i$.

Precomputed curvatures combined with suitable relaxation yields guaranteed convergence for convex problems [228].

Summary of Algorithms

- General-purpose optimization algorithms
- Optimization transfer for image reconstruction algorithms
- Separable surrogates $\Longrightarrow$ high curvatures $\Longrightarrow$ slow convergence
- Ordered subsets accelerate *initial* convergence
  require relaxation or incrementalism for true convergence
- Principles apply to emission and transmission reconstruction
- Still work to be done...

Matlab/Freemat "image reconstruction toolbox" online:
`http://www.eecs.umich.edu/~fessler /code`

An Open Problem

Still no algorithm with all of the following properties:
- Nonnegativity easy
- Fast converging
- Intrinsically monotone global convergence
- Accepts any type of system matrix
- Parallelizable

Until an "ideal" algorithm is developed, OSEM will probably remain very popular...

Part 4. Performance Characteristics

- Spatial resolution properties
- Noise properties
- Detection properties

Spatial Resolution Properties

Choosing β can be painful, so ...

For true minimization methods:

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x})$$

the *local impulse response* is approximately (Fessler and Rogers, T-MI, 1996):

$$\boldsymbol{l}_j(\boldsymbol{x}) = \lim_{\delta \to 0} \frac{\mathsf{E}[\hat{\boldsymbol{x}}|\boldsymbol{x} + \delta \boldsymbol{e}_j] - \mathsf{E}[\hat{\boldsymbol{x}}|\boldsymbol{x}]}{\delta} \approx \left[-\nabla^{20}\Psi\right]^{-1} \nabla^{11}\Psi \frac{\partial}{\partial x_j} \bar{\boldsymbol{y}}(\boldsymbol{x}).$$

Depends only on chosen cost function and statistical model.
Independent of optimization algorithm (if iterated “to convergence”).

- Enables prediction of resolution properties
(provided Ψ is minimized)
- Useful for designing regularization penalty functions
with desired resolution properties. For penalized likelihood:

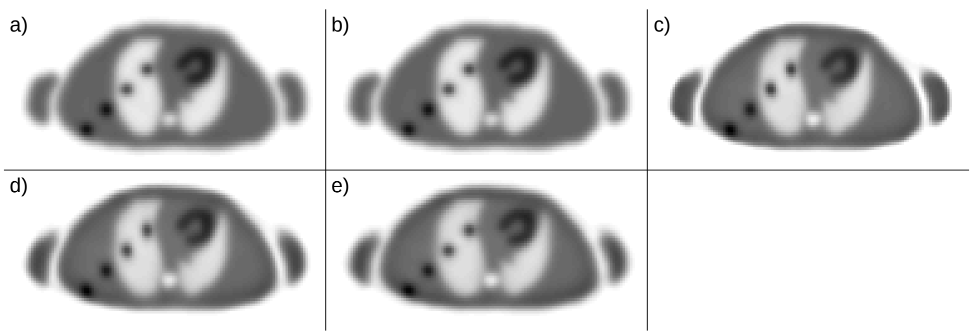
$$\boldsymbol{l}_j(\boldsymbol{x}) \approx [\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \beta \boldsymbol{R}]^{-1} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}\boldsymbol{x}^{\text{true}}.$$

- Helps choose β for desired spatial resolution

[36, 139, 234]

A commonly cited disadvantage of regularized methods is the need to select the regularization parameter β. One must also select the cutoff frequency for FBP, but at least that value is intuitive and works the same (resolution-wise) for all patients. Not so for stopping rules. The analysis in [36, 139, 234] brings some of the consistency of FBP-like resolution selection to statistical methods.

Modified Penalty Example, PET

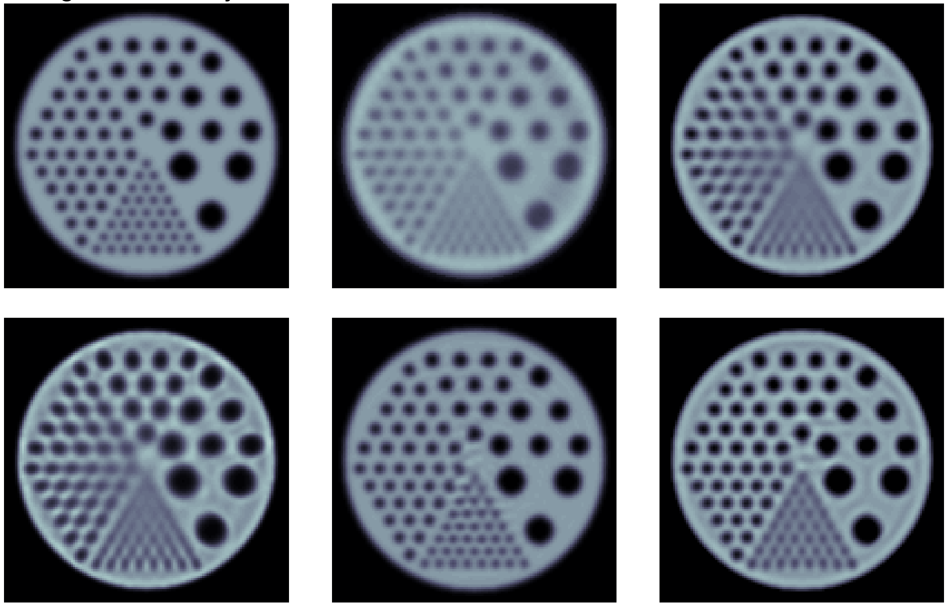


a) filtered backprojection
b) Penalized unweighted least-squares
c) PWLS with conventional regularization
d) PWLS with certainty-based penalty [36]
e) PWLS with modified penalty [139]

Figure from [139].

Modified Penalty Example, SPECT - Noiseless

Target filtered object FBP Conventional PWLS

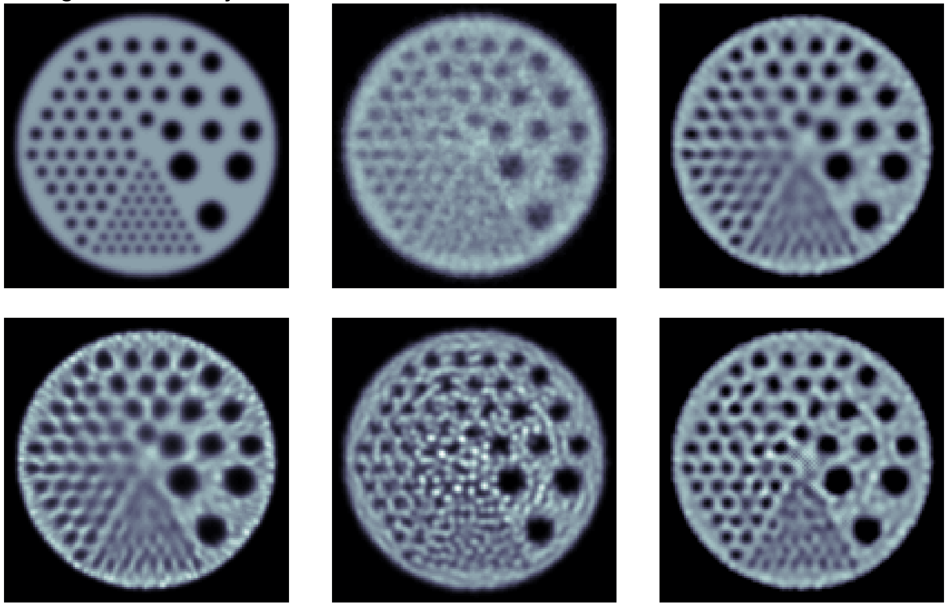


Truncated EM Post-filtered EM Modified Regularization

Figure from [140].

Modified Penalty Example, SPECT - Noisy

Target filtered object FBP Conventional PWLS



Truncated EM Post-filtered EM Modified Regularization

Figure from [140].

These were preliminary results, and they initially casted doubt on the claim sometimes made that post-filtered EM (or OSEM) is equivalent to truly regularized image reconstruction.

However, it turns out we had not matched spatial resolution as carefully as needed...

See [141, 235].

Regularized vs Post-filtered, with Matched PSF

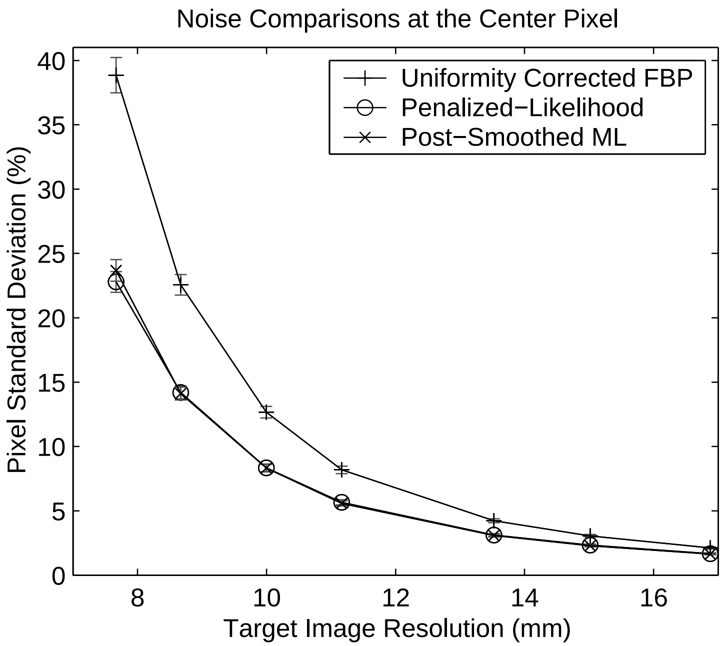


Figure from [141].

Reconstruction Noise Properties

For unconstrained (converged) minimization methods, the estimator is *implicit*:

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}, \boldsymbol{y}).$$

What is $\mathsf{Cov}\{\hat{\boldsymbol{x}}\}$? New simpler derivation.

Denote the column gradient by $g(\boldsymbol{x}, \boldsymbol{y}) \triangleq \nabla_{\boldsymbol{x}} \Psi(\boldsymbol{x}, \boldsymbol{y})$.
Ignoring constraints, the gradient is zero at the minimizer: $g(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) = \mathbf{0}$.
First-order Taylor series expansion:

$$\begin{aligned} g(\hat{\boldsymbol{x}}, \boldsymbol{y}) &\approx g(\boldsymbol{x}^{\text{true}}, \boldsymbol{y}) + \nabla_{\boldsymbol{x}} g(\boldsymbol{x}^{\text{true}}, \boldsymbol{y})(\hat{\boldsymbol{x}} - \boldsymbol{x}^{\text{true}}) \\ &= g(\boldsymbol{x}^{\text{true}}, \boldsymbol{y}) + \nabla_{\boldsymbol{x}}^2 \Psi\left(\boldsymbol{x}^{\text{true}}, \boldsymbol{y}\right)(\hat{\boldsymbol{x}} - \boldsymbol{x}^{\text{true}}). \end{aligned}$$

Equating to zero:

$$\hat{\boldsymbol{x}} \approx \boldsymbol{x}^{\text{true}} - \left[\nabla_{\boldsymbol{x}}^2 \Psi\left(\boldsymbol{x}^{\text{true}}, \boldsymbol{y}\right)\right]^{-1} \nabla_{\boldsymbol{x}} \Psi\left(\boldsymbol{x}^{\text{true}}, \boldsymbol{y}\right).$$

If the Hessian $\nabla^2 \Psi$ is weakly dependent on $\boldsymbol{y}$, then

$$\boxed{\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\nabla_{\boldsymbol{x}}^2 \Psi\left(\boldsymbol{x}^{\text{true}}, \bar{\boldsymbol{y}}\right)\right]^{-1} \mathsf{Cov}\left\{\nabla_{\boldsymbol{x}} \Psi\left(\boldsymbol{x}^{\text{true}}, \boldsymbol{y}\right)\right\} \left[\nabla_{\boldsymbol{x}}^2 \Psi\left(\boldsymbol{x}^{\text{true}}, \bar{\boldsymbol{y}}\right)\right]^{-1}.}$$

If we further linearize w.r.t. the data: $g(\boldsymbol{x}, \boldsymbol{y}) \approx g(\boldsymbol{x}, \bar{\boldsymbol{y}}) + \nabla_{\boldsymbol{y}} g(\boldsymbol{x}, \bar{\boldsymbol{y}})(\boldsymbol{y} - \bar{\boldsymbol{y}})$, then

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\nabla_{\boldsymbol{x}}^2 \Psi\right]^{-1} (\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{y}} \Psi) \; \mathsf{Cov}\{\boldsymbol{y}\} \; (\nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{y}} \Psi)' \left[\nabla_{\boldsymbol{x}}^2 \Psi\right]^{-1}.$$

The latter approximation was derived in [35].

Covariance Continued

Covariance approximation:

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\nabla_{\boldsymbol{x}}^2 \Psi\left(\boldsymbol{x}^{\text{true}}, \bar{\boldsymbol{y}}\right)\right]^{-1} \mathsf{Cov}\left\{\nabla_{\boldsymbol{x}} \Psi\left(\boldsymbol{x}^{\text{true}}, \boldsymbol{y}\right)\right\} \left[\nabla_{\boldsymbol{x}}^2 \Psi\left(\boldsymbol{x}^{\text{true}}, \bar{\boldsymbol{y}}\right)\right]^{-1}$$

Depends only on chosen cost function and statistical model.
Independent of optimization algorithm.

- Enables prediction of noise properties
- Can make variance images
- Useful for computing ROI variance (*e.g.*, for weighted kinetic fitting)
- Good variance prediction for quadratic regularization in nonzero regions
- Inaccurate for nonquadratic penalties, or in nearly-zero regions

Qi has developed an approximation that may help with the nonnegativity constraint [236].

Qi and Huesman's Detection Analysis

SNR of MAP reconstruction $>$ SNR of FBP reconstruction (T-MI, Aug. 2001)

quadratic regularization
SKE/BKE task
prewhitened observer
non-prewhitened observer

Open issues

Choice of regularizer to optimize detectability?
Active work in several groups.
(*e.g.*, 2004 MIC poster by Yendiki & Fessler.)

Qi's theoretical predictions are consistent with some empirical results, *e.g.*, [237].

See recent detection analyses papers, *e.g.*, [42, 69, 238–242].

References

[1] S. Webb. *From the watching of shadows: the origins of radiological tomography*. A. Hilger, Bristol, 1990.

[2] H. H. Barrett and K. J. Myers. *Foundations of image science*. Wiley, New York, 2003.

[3] J. Kay. The EM algorithm in medical imaging. *Stat. Meth. Med. Res.*, 6(1):55–75, January 1997.

[4] J. A. Fessler. Statistical image reconstruction methods for transmission tomography. In M. Sonka and J. Michael Fitzpatrick, editors, *Handbook of Medical Imaging, Volume 2. Medical Image Processing and Analysis*, pages 1–70. SPIE, Bellingham, 2000.

[5] R. M. Leahy and J. Qi. Statistical approaches in quantitative positron emission tomography. *Statistics and Computing*, 10(2):147–65, April 2000.

[6] M. Defrise. A short reader's guide to 3D tomographic reconstruction. *Computerized Medical Imaging and Graphics*, 25(2):113–6, March 2001.

[7] S. Vandenberghe, Y. D'Asseler, R. V. . Walle, T. Kauppinen, M. Koole, L. Bouwens, K. V. Laere, I. Lemahieu, and R. A. Dierckx. Iterative reconstruction algorithms in nuclear medicine. *Computerized Medical Imaging and Graphics*, 25(2):105–11, March 2001.

[8] G. L. Zeng. Image reconstruction, a tutorial. *Computerized Medical Imaging and Graphics*, 25(2):97–103, March 2001.

[9] R. M. Lewitt and S. Matej. Overview of methods for image reconstruction from projections in emission computed tomography. *Proc. IEEE*, 91(10):1588–611, October 2003.

[10] R. N. Bracewell. Strip integration in radio astronomy. *Aust. J. Phys.*, 9:198–217, 1956.

[11] G. Hounsfield. A method of apparatus for examination of a body by radiation such as x-ray or gamma radiation, 1972. US Patent 1283915. British patent 1283915, London. Issued to EMI Ltd. Filed Aug. 1968. See [1, Ch. 5].

[12] G. Muehllehner and R. A. Wetzel. Section imaging by computer calculation. *J. Nuc. Med.*, 12(2):76–85, 1971.

[13] D. E. Kuhl, R. Q. Edwards, A. R. Ricci, and M. Reivich. Quantitative section scanning using orthogonal tangent correction. *J. Nuc. Med.*, 14(4):196–200, April 1973.

[14] R. Gordon, R. Bender, and G. T. Herman. Algebraic reconstruction techniques (ART) for the three-dimensional electron microscopy and X-ray photography. *J. Theor. Biol.*, 29:471–81, 1970.

[15] R. Gordon and G. T. Herman. Reconstruction of pictures from their projections. *Comm. ACM*, 14:759–68, 1971.

[16] G. T. Herman, A. Lent, and S. W. Rowland. ART: mathematics and applications (a report on the mathematical foundations and on the applicability to real data of the algebraic reconstruction techniques). *J. Theor. Biol.*, 42:1–32, 1973.

[17] R. Gordon. A tutorial on ART (algebraic reconstruction techniques). *IEEE Trans. Nuc. Sci.*, 21(3):78–93, June 1974.

[18] R. Richardson. Bayesian-based iterative method of image restoration. *J. Opt. Soc. Am.*, 62(1):55–9, January 1972.

[19] L. Lucy. An iterative technique for the rectification of observed distributions. *The Astronomical J.*, 79(6):745–54, June 1974.

[20] A. J. Rockmore and A. Macovski. A maximum likelihood approach to emission image reconstruction from projections. *IEEE Trans. Nuc. Sci.*, 23:1428–32, 1976.

[21] A. J. Rockmore and A. Macovski. A maximum likelihood approach to transmission image reconstruction from projections. *IEEE Trans. Nuc. Sci.*, 24(3):1929–35, June 1977.

[22] 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*, 39(1):1–38, 1977.

[23] L. A. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imag.*, 1(2):113–22, October 1982.

[24] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomo.*, 8(2):306–16, April 1984.

[25] S. Geman and D. E. McClure. Bayesian image analysis: an application to single photon emission tomography. In *Proc. of Stat. Comp. Sect. of Amer. Stat. Assoc.*, pages 12–8, 1985.

[26] H. M. Hudson and R. S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans. Med. Imag.*, 13(4):601–9, December 1994.

[27] M. Goitein. Three-dimensional density reconstruction from a series of two-dimensional projections. *Nucl. Instr. Meth.*, 101(3):509–18, June 1972.

[28] T. F. Budinger and G. T. Gullberg. Three dimensional reconstruction in nuclear medicine emission imaging. *IEEE Trans. Nuc. Sci.*, 21(3):2–20, June 1974.

[29] R. H. Huesman, G. T. Gullberg, W. L. Greenberg, and T. F. Budinger. *RECLBL library users manual*. Lawrence Berkeley Laboratory, Berkeley, CA, 1977.

[30] R. H. Huesman. A new fast algorithm for the evaluation of regions of interest and statistical uncertainty in computed tomography. *Phys. Med. Biol.*, 29(5):543–52, May 1984.

[31] D. W. Wilson and B. M. W. Tsui. Noise properties of filtered-backprojection and ML-EM reconstructed emission tomographic images. *IEEE Trans. Nuc. Sci.*, 40(4):1198–1203, August 1993.

[32] D. W. Wilson and B. M. W. Tsui. Spatial resolution properties of FB and ML-EM reconstruction methods. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 1189–93, 1993.

[33] H. H. Barrett, D. W. Wilson, and B. M. W. Tsui. Noise properties of the EM algorithm: I. Theory. *Phys. Med. Biol.*, 39(5):833–46, May 1994.

[34] D. W. Wilson, B. M. W. Tsui, and H. H. Barrett. Noise properties of the EM algorithm: II. Monte Carlo simulations. *Phys. Med. Biol.*, 39(5):847–72, May 1994.

[35] J. A. Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography. *IEEE Trans. Im. Proc.*, 5(3):493–506, March 1996.

[36] J. A. Fessler and W. L. Rogers. Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs. *IEEE Trans. Im. Proc.*, 5(9):1346–58, September 1996.

[37] W. Wang and G. Gindi. Noise analysis of regularized EM SPECT reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1933–7, 1996.

[38] C. K. Abbey, E. Clarkson, H. H. Barrett, S. P. Mueller, and F. J. Rybicki. Approximate distributions for maximum likelihood and maximum a posteriori estimates under a Gaussian noise model. In J. Duncan and G. Gindi, editors, *Information Processing in Medical Im.*, pages 167–75. Springer-Verlag, Berlin, 1997.

[39] W. Wang and G. Gindi. Noise analysis of MAP-EM algorithms for emission tomography. *Phys. Med. Biol.*, 42(11):2215–32, November 1997.

[40] S. J. Glick and E. J. Soares. Noise characteristics of SPECT iterative reconstruction with a mis-matched projector-backprojector pair. *IEEE Trans. Nuc. Sci.*, 45(4):2183–8, August 1998.

[41] E. J. Soares, C. L. Byrne, T-S. Pan, S. J. Glick, and M. A. King. Modeling the population covariance matrices of block-iterative expectation-maximization reconstructed images. In *Proc. SPIE 3034, Med. Im. 1997: Im. Proc.*, pages 415–25, 1997.

[42] J. Qi and R. H. Huesman. Theoretical study of lesion detectability of MAP reconstruction using computer observers. *IEEE Trans. Med. Imag.*, 20(8):815–22, August 2001.

[43] D. Brasse, P. E. Kinahan, R. Clackdoyle, M. Defrise, C. Comtat, and D. W. Townsend. Fast fully 3-D image reconstruction in PET using planograms. *IEEE Trans. Med. Imag.*, 23(4):413–25, April 2004.

[44] J. A. Fessler, I. Elbakri, P. Sukovic, and N. H. Clinthorne. Maximum-likelihood dual-energy tomographic image reconstruction. In *Proc. SPIE 4684, Medical Imaging 2002: Image Proc.*, volume 1, pages 38–49, 2002.

[45] B. De Man, J. Nuyts, P. Dupont, G. Marchal, and P. Suetens. An iterative maximum-likelihood polychromatic algorithm for CT. *IEEE Trans. Med. Imag.*, 20(10):999–1008, October 2001.

[46] I. A. Elbakri and J. A. Fessler. Statistical image reconstruction for polyenergetic X-ray computed tomography. *IEEE Trans. Med. Imag.*, 21(2):89–99, February 2002.

[47] I. A. Elbakri and J. A. Fessler. Segmentation-free statistical image reconstruction for polyenergetic X-ray computed tomography with experimental validation. *Phys. Med. Biol.*, 48(15):2543–78, August 2003.

[48] P. E. Kinahan, J. A. Fessler, and J. S. Karp. Statistical image reconstruction in PET with compensation for missing data. *IEEE Trans. Nuc. Sci.*, 44(4):1552–7, August 1997.

[49] J. A. Fessler and B. P. Sutton. Nonuniform fast Fourier transforms using min-max interpolation. *IEEE Trans. Sig. Proc.*, 51(2):560–74, February 2003.

[50] B. P. Sutton, D. C. Noll, and J. A. Fessler. Fast, iterative image reconstruction for MRI in the presence of field inhomogeneities. *IEEE Trans. Med. Imag.*, 22(2):178–88, February 2003.

[51] B. P. Sutton, D. C. Noll, and J. A. Fessler. Dynamic field map estimation using a spiral-in / spiral-out acquisition. *Mag. Res. Med.*, 51(6):1194–204, June 2004.

[52] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Royal Stat. Soc. Ser. B*, 47(1):1–52, 1985.

[53] R. Van de Walle, H. H. Barrett, K. J. Myers, M. I. Altbach, B. Desplanques, A. F. Gmitro, J. Cornelis, and I. Lemahieu. Reconstruction of MR images from data acquired on a general non-regular grid by pseudoinverse calculation. *IEEE Trans. Med. Imag.*, 19(12):1160–7, December 2000.

[54] M. Bertero, C. De Mol, and E. R. Pike. Linear inverse problems with discrete data, I: General formulation and singular system analysis. *Inverse Prob.*, 1(4):301–30, November 1985.

[55] K. M. Hanson and G. W. Wecksung. Local basis-function approach to computed tomography. *Appl. Optics*, 24(23):4028–39, December 1985.

[56] E. J. Mazur and R. Gordon. Interpolative algebraic reconstruction techniques without beam partitioning for computed tomography. *Med. Biol. Eng. Comput.*, 33(1):82–6, January 1995.

[57] Y. Censor. Finite series expansion reconstruction methods. *Proc. IEEE*, 71(3):409–19, March 1983.

[58] D. L. Snyder. Utilizing side information in emission tomography. *IEEE Trans. Nuc. Sci.*, 31(1):533–7, February 1984.

[59] R. E. Carson, M. V. Green, and S. M. Larson. A maximum likelihood method for calculation of tomographic region-of-interest (ROI) values. *J. Nuc. Med.*, 26:P20, 1985.

[60] R. E. Carson and K. Lange. The EM parametric image reconstruction algorithm. *J. Am. Stat. Ass.*, 80(389):20–2, March 1985.

[61] R. E. Carson. A maximum likelihood method for region-of-interest evaluation in emission tomography. *J. Comp. Assisted Tomo.*, 10(4):654–63, July 1986.

[62] A. R. Formiconi. Least squares algorithm for region-of-interest evaluation in emission tomography. *IEEE Trans. Med. Imag.*, 12(1):90–100, March 1993.

[63] B. W. Reutter, G. T. Gullberg, and R. H. Huesman. Kinetic parameter estimation from attenuated SPECT projection measurements. *IEEE Trans. Nuc. Sci.*, 45(6):3007–13, December 1998.

[64] 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 Trans. Acoust. Sp. Sig. Proc.*, 32(4):886–906, August 1984.

[65] 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.*, 31(10):1693–701, October 1990.

[66] 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 Trans. Med. Imag.*, 13(2):217–26, June 1994.

[67] Z. P. Liang, F. E. Boada, R. T. Constable, E. M. Haacke, P. C. Lauterbur, and M. R. Smith. Constrained reconstruction methods in MR imaging. *Reviews of Magnetic Resonance in Medicine*, 4:67–185, 1992.

[68] G. S. Cunningham and A. Lehovich. 4D reconstructions from low-count SPECT data using deformable models with smooth interior intensity variations. In *Proc. SPIE 3979: Medical Imaging 2000: Image Proc.*, 2000.

[69] A. Yendiki and J. A. Fessler. A comparison of rotation- and blob-based system models for 3D SPECT with depth-dependent detector response. *Phys. Med. Biol.*, 49(11):2157–68, June 2004.

[70] R. M. Lewitt. Multidimensional digital image representations using generalized Kaiser-Bessel window functions. *J. Opt. Soc. Am. A*, 7(10):1834–46, October 1990.

[71] R. M. Lewitt. Alternatives to voxels for image representation in iterative reconstruction algorithms. *Phys. Med. Biol.*, 37(3):705–16, March 1992.

[72] G. L. Zeng and G. T. Gullberg. Unmatched projector/backprojector pairs in an iterative reconstruction algorithm. *IEEE Trans. Med. Imag.*, 19(5):548–55, May 2000.

[73] C. Kamphuis, F. J. Beekman, P. P. van Rijk, and M. A. Viergever. Dual matrix ordered subsets reconstruction for accelerated 3D scatter compensation in single-photon emission tomography. *Eur. J. Nuc. Med.*, 25(1):8–18, January 1998.

[74] F. J. Beekman, H. W. A. M. de Jong, and S. van Geloven. Efficient fully 3D iterative SPECT reconstruction with Monte Carlo based scatter compensation. *IEEE Trans. Med. Imag.*, 21(8):867–77, August 2002.

[75] R. Griesse and A. Walther. Evaluating gradients in optimal control: continuous adjoints versus automatic differentiation. *J. Optim. Theory Appl.*, 122(1):63–86, July 2004.

[76] C. W. Stearns and J. A. Fessler. 3D PET reconstruction with FORE and WLS-OS-EM. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 912–5, 2002.

[77] M. Yavuz and J. A. Fessler. Objective functions for tomographic reconstruction from randoms-precorrected PET scans. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 1067–71, 1996.

[78] M. Yavuz and J. A. Fessler. New statistical models for randoms-precorrected PET scans. In J Duncan and G Gindi, editors, *Information Processing in Medical Im.*, volume 1230 of *Lecture Notes in Computer Science*, pages 190–203. Springer-Verlag, Berlin, 1997.

[79] M. Yavuz and J. A. Fessler. Statistical image reconstruction methods for randoms-precorrected PET scans. *Med. Im. Anal.*, 2(4):369–78, December 1998.

[80] M. Yavuz and J. A. Fessler. Penalized-likelihood estimators and noise analysis for randoms-precorrected PET transmission scans. *IEEE Trans. Med. Imag.*, 18(8):665–74, August 1999.

[81] D. L. Snyder, A. M. Hammoud, and R. L. White. Image recovery from data acquired with a charge-coupled-device camera. *J. Opt. Soc. Am. A*, 10(5):1014–23, May 1993.

[82] D. L. Snyder, C. W. Helstrom, A. D. Lanterman, M. Faisal, and R. L. White. Compensation for readout noise in CCD images. *J. Opt. Soc. Am. A*, 12(2):272–83, February 1995.

[83] U. Engeland, T. Striker, and H. Luig. Count-rate statistics of the gamma camera. *Phys. Med. Biol.*, 43(10):2939–47, October 1998.

[84] J. A. Fessler. Penalized weighted least-squares image reconstruction for positron emission tomography. *IEEE Trans. Med. Imag.*, 13(2):290–300, June 1994.

[85] C. Comtat, P. E. Kinahan, M. Defrise, C. Michel, and D. W. Townsend. Fast reconstruction of 3D PET data with accurate statistical modeling. *IEEE Trans. Nuc. Sci.*, 45(3):1083–9, June 1998.

[86] B. R. Whiting. Signal statistics in x-ray computed tomography. In *Proc. SPIE 4682, Medical Imaging 2002: Med. Phys.*, pages 53–60, 2002.

[87] B. R. Whiting, L. J. Montagnino, and D. G. Politte. Modeling X-ray computed tomography sinograms, 2001. submitted to mp.

[88] I. A. Elbakri and J. A. Fessler. Efficient and accurate likelihood for iterative image reconstruction in X-ray computed tomography. In *Proc. SPIE 5032, Medical Imaging 2003: Image Proc.*, pages 1839–50, 2003.

[89] A. O. Hero, J. A. Fessler, and M. Usman. Exploring estimator bias-variance tradeoffs using the uniform CR bound. *IEEE Trans. Sig. Proc.*, 44(8):2026–41, August 1996.

[90] Y. C. Eldar. Minimum variance in biased estimation: bounds and asymptotically optimal estimators. *IEEE Trans. Sig. Proc.*, 52(7):1915–30, July 2004.

[91] E. Mumcuoglu, R. Leahy, S. Cherry, and E. Hoffman. Accurate geometric and physical response modeling for statistical image reconstruction in high resolution PET scanners. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1569–73, 1996.

[92] J. Qi, R. M. Leahy, E. U. Mumcuoglu, S. R. Cherry, A. Chatziioannou, and T. H. Farquhar. High resolution 3D Bayesian image reconstruction for microPET. In *Proc. Intl. Mtg. on Fully 3D Image Recon. in Rad. and Nuc. Med*, 1997.

[93] G. Christ. Exact treatment of the dual-energy method in CT using polyenergetic X-ray spectra. *Phys. Med. Biol.*, 29(12):1511–25, December 1984.

[94] Z. Liang. Compensation for attenuation, scatter, and detector response in SPECT reconstruction via iterative FBP methods. *Med. Phys.*, 20(4):1097–106, July 1993.

[95] X. L. Xu, J. S. Liow, and S. C. Strother. Iterative algebraic reconstruction algorithms for emission computed tomography: a unified framework and its application to positron emission tomography. *Med. Phys.*, 20(6):1675–84, November 1993.

[96] J. W. Wallis and T. R. Miller. Rapidly converging iterative reconstruction algorithms in single-photon emission computed tomography. *J. Nuc. Med.*, 34(10):1793–1800, October 1993.

[97] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. Royal Stat. Soc. Ser. B*, 46(2):149–92, 1984.

[98] 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.*, volume 2, pages 1292–6, 1995.

[99] J. M. M. Anderson, B. A. Mair, M. Rao, and C-H. Wu. Weighted least-squares reconstruction methods for positron emission tomography. *IEEE Trans. Med. Imag.*, 16(2):159–65, April 1997.

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

[101] C. Bouman and K. Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. Im. Proc.*, 2(3):296–310, July 1993.

[102] E. Tanaka. Improved iterative image reconstruction with automatic noise artifact suppression. *IEEE Trans. Med. Imag.*, 11(1):21–7, March 1992.

[103] B. W. Silverman, M. C. Jones, J. D. Wilson, and D. W. Nychka. A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *J. Royal Stat. Soc. Ser. B*, 52(2):271–324, 1990.

[104] T. R. Miller, J. W. Wallis, C. S. Butler, M. I. Miller, and D. L. Snyder. Improved brain SPECT by maximum-likelihood reconstruction. *J. Nuc. Med. (Abs. Book)*, 33(5):964, May 1992.

[105] F. J. Beekman, E. T. P. Slijpen, and W. J. Niessen. Selection of task-dependent diffusion filters for the post-processing of SPECT images. *Phys. Med. Biol.*, 43(6):1713–30, June 1998.

[106] D. S. Lalush and B. M. W. Tsui. Performance of ordered subset reconstruction algorithms under conditions of extreme attenuation and truncation in myocardial SPECT. *J. Nuc. Med.*, 41(4):737–44, April 2000.

[107] E. T. P. Slijpen and F. J. Beekman. Comparison of post-filtering and filtering between iterations for SPECT reconstruction. *IEEE Trans. Nuc. Sci.*, 46(6):2233–8, December 1999.

[108] R. H. Huesman. The effects of a finite number of projection angles and finite lateral sampling of projections on the propagation of statistical errors in transverse section reconstruction. *Phys. Med. Biol.*, 22(3):511–21, May 1977.

[109] D. L. Snyder and M. I. Miller. The use of sieves to stabilize images produced with the EM algorithm for emission tomography. *IEEE Trans. Nuc. Sci.*, 32(5):3864–71, October 1985.

[110] 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 Trans. Med. Imag.*, 6(3):228–38, September 1987.

[111] T. R. Miller and J. W. Wallis. Clinically important characteristics of maximum-likelihood reconstruction. *J. Nuc. Med.*, 33(9):1678–84, September 1992.

[112] A. Tikhonov and V. Arsenin. *Solution of ill-posed problems*. Wiley, New York, 1977.

[113] I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *Ann. Stat.*, 19(4):2032–66, 1991.

[114] D. L. Donoho, I. M. Johnstone, A. S. Stern, and JC. Hoch. Does the maximum entropy method improve sensitivity. *Proc. Natl. Acad. Sci.*, 87(13):5066–8, July 1990.

[115] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *J. Royal Stat. Soc. Ser. B*, 54(1):41–81, 1992.

[116] R. T. Constable and R. M. Henkelman. Why MEM does not work in MR image reconstruction. *Mag. Res. Med.*, 14(1):12–25, April 1990.
[117] A. R. De Pierro. A modified expectation maximization algorithm for penalized likelihood estimation in emission tomography. *IEEE Trans. Med. Imag.*, 14(1):132–137, March 1995.
[118] A. H. Delaney and Y. Bresler. A fast and accurate Fourier algorithm for iterative parallel-beam tomography. *IEEE Trans. Im. Proc.*, 5(5):740–53, May 1996.
[119] S. J. Lee, A. Rangarajan, and G. Gindi. A comparative study of the effects of using higher order mechanical priors in SPECT reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 4, pages 1696–1700, 1994.
[120] S-J. Lee, A. Rangarajan, and G. Gindi. Bayesian image reconstruction in SPECT using higher order mechanical models as priors. *IEEE Trans. Med. Imag.*, 14(4):669–80, December 1995.
[121] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images. *IEEE Trans. Patt. Anal. Mach. Int.*, 6(6):721–41, November 1984.
[122] B. W. Silverman, C. Jennison, J. Stander, and T. C. Brown. The specification of edge penalties for regular and irregular pixel images. *IEEE Trans. Patt. Anal. Mach. Int.*, 12(10):1017–24, October 1990.
[123] V. E. Johnson, W. H. Wong, X. Hu, and C. T. Chen. Image restoration using Gibbs priors: Boundary modeling, treatment of blurring, and selection of hyperparameter. *IEEE Trans. Patt. Anal. Mach. Int.*, 13(5):413–25, May 1991.
[124] V. E. Johnson. A model for segmentation and analysis of noisy images. *J. Am. Stat. Ass.*, 89(425):230–41, March 1994.
[125] S. Alenius and U. Ruotsalainen. Bayesian image reconstruction for emission tomography based on median root prior. *Eur. J. Nuc. Med.*, 24(3):258–65, 1997.
[126] S. Alenius, U. Ruotsalainen, and J. Astola. Using local median as the location of the prior distribution in iterative emission tomography image reconstruction. *IEEE Trans. Nuc. Sci.*, 45(6):3097–104, December 1998.
[127] W. Chlewicki, F. Hermansen, and S. B. Hansen. Noise reduction and convergence of Bayesian algorithms with blobs based on the huber function and median root prior. *Phys. Med. Biol.*, 49(20):4717–30, October 2004.
[128] V. Y. Panin, G. L. Zeng, and G. T. Gullberg. Total variation regulated EM algorithm. *IEEE Trans. Nuc. Sci.*, 46(6):2202–10, December 1999.
[129] P. Kisilev, M. Zibulevsky, and Y. Zeevi. Wavelet representation and total variation regularization in emission tomography. In *Proc. IEEE Intl. Conf. on Image Processing*, volume 1, pages 702–5, 2001.
[130] C. R. Vogel and M. E. Oman. Fast numerical methods for total variation minimization in image reconstruction. In *Proc. SPIE 2563, Adv. Signal Proc. Alg.*, pages 359–67, 1995.
[131] M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Prob.*, 20(5):1537–1564, October 2004.
[132] I-T. Hsiao, A. Rangarajan, and G. Gindi. A new convex edge-preserving median prior with applications to tomography. *IEEE Trans. Med. Imag.*, 22(5):580–5, May 2003.
[133] M. Nikolova. Thresholding implied by truncated quadratic regularization. *IEEE Trans. Sig. Proc.*, 48(12):3437–50, December 2000.
[134] A. Antoniadis and J. Fan. Regularization and wavelet approximations. *J. Am. Stat. Ass.*, 96(455):939–55, September 2001.
[135] D. F. Yu and J. A. Fessler. Edge-preserving tomographic reconstruction with nonlocal regularization. In *Proc. IEEE Intl. Conf. on Image Processing*, volume 1, pages 29–33, 1998.
[136] D. F. Yu and J. A. Fessler. Edge-preserving tomographic reconstruction with nonlocal regularization. *IEEE Trans. Med. Imag.*, 21(2):159–73, February 2002.
[137] J. Ye, Y. Bresler, and P. Moulin. A self-referencing level-set method for image reconstruction from sparse Fourier samples. In *Proc. IEEE Intl. Conf. on Image Processing*, volume 2, pages 33–6, 2001.
[138] M. J. Black and A. Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with applications in early vision. *Intl. J. Comp. Vision*, 19(1):57–91, July 1996.
[139] J. W. Stayman and J. A. Fessler. Regularization for uniform spatial resolution properties in penalized-likelihood image reconstruction. *IEEE Trans. Med. Imag.*, 19(6):601–15, June 2000.
[140] J. W. Stayman and J. A. Fessler. Nonnegative definite quadratic penalty design for penalized-likelihood reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 1060–3, 2001.
[141] J. W. Stayman and J. A. Fessler. Compensation for nonuniform resolution using penalized-likelihood reconstruction in space-variant imaging systems. *IEEE Trans. Med. Imag.*, 23(3):269–84, March 2004.
[142] C. T. Chen, X. Ouyang, W. H. Wong, and X. Hu. Improvement of PET image reconstruction using high-resolution anatomic images. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, page 2062, 1991. (Abstract.).
[143] R. Leahy and X. H. Yan. Statistical models and methods for PET image reconstruction. In *Proc. of Stat. Comp. Sect. of Amer. Stat. Assoc.*, pages 1–10, 1991.

[144] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers. Regularized emission image reconstruction using imperfect side information. *IEEE Trans. Nuc. Sci.*, 39(5):1464–71, October 1992.
[145] I. G. Zubal, M. Lee, A. Rangarajan, C. R. Harrell, and G. Gindi. Bayesian reconstruction of SPECT images using registered anatomical images as priors. *J. Nuc. Med. (Abs. Book)*, 33(5):963, May 1992.
[146] G. Gindi, M. Lee, A. Rangarajan, and I. G. Zubal. Bayesian reconstruction of functional images using anatomical information as priors. *IEEE Trans. Med. Imag.*, 12(4):670–80, December 1993.
[147] X. Ouyang, W. H. Wong, V. E. Johnson, X. Hu, and C-T. Chen. Incorporation of correlated structural images in PET image reconstruction. *IEEE Trans. Med. Imag.*, 13(4):627–40, December 1994.
[148] S. J. Lee, G. R. Gindi, I. G. Zubal, and A. Rangarajan. Using ground-truth data to design priors in Bayesian SPECT reconstruction. In Y. Bizais, C. Barillot, and R. D. Paola, editors, *Information Processing in Medical Im.* Kluwer, 1995.
[149] J. E. Bowsher, V. E. Johnson, T. G. Turkington, R. J. Jaszczak, C. E. Floyd, and R. E. Coleman. Bayesian reconstruction and use of anatomical a priori information for emission tomography. *IEEE Trans. Med. Imag.*, 15(5):673–86, October 1996.
[150] S. Sastry and R. E. Carson. Multimodality Bayesian algorithm for image reconstruction in positron emission tomography: a tissue composition model. *IEEE Trans. Med. Imag.*, 16(6):750–61, December 1997.
[151] R. Piramuthu and A. O. Hero. Side information averaging method for PML emission tomography. In *Proc. IEEE Intl. Conf. on Image Processing*, volume 2, pages 671–5, 1998.
[152] C. Comtat, P. E. Kinahan, J. A. Fessler, T. Beyer, D. W. Townsend, M. Defrise, and C. Michel. Reconstruction of 3d whole-body PET data using blurred anatomical labels. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 1651–5, 1998.
[153] A. O. Hero, R. Piramuthu, S. R. Titus, and J. A. Fessler. Minimax emission computed tomography using high resolution anatomical side information and B-spline models. *IEEE Trans. Info. Theory*, 45(3):920–38, April 1999.
[154] D. F. Yu and J. A. Fessler. Mean and variance of singles photon counting with deadtime. *Phys. Med. Biol.*, 45(7):2043–56, July 2000.
[155] D. F. Yu and J. A. Fessler. Mean and variance of coincidence photon counting with deadtime. *Nucl. Instr. Meth. Phys. Res. A.*, 488(1-2):362–74, August 2002.
[156] J. Qi and R. H. Huesman. Propagation of errors from the sensitivity image in list mode reconstruction. *IEEE Trans. Med. Imag.*, 23(9):1094–9, September 2004.
[157] J. Qi and R. H. Huesman. Effect of errors in the system matrix on iterative image reconstruction. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 5, pages 2854–8, 2004.
[158] Y. S. Shim and Z. H. Cho. SVD pseudoinversion image reconstruction. *IEEE Trans. Acoust. Sp. Sig. Proc.*, 29(4):904–9, August 1981.
[159] U. Raff, D. N. Stroud, and W. R. Hendee. Improvement of lesion detection in scintigraphic images by SVD techniques for resolution recovery. *IEEE Trans. Med. Imag.*, 5(1):35–44, March 1986.
[160] D. A. Fish, J. Grochmalicki, and E. R. Pike. Scanning SVD method for restoration of images with space-variant blur. *J. Opt. Soc. Am. A*, 13(3):464–9, March 1996.
[161] A. Caponnetto and M. Bertero. Tomography with a finite set of projections: singular value decomposition and re solution. *Inverse Prob.*, 13(5):1191–1205, October 1997.
[162] A. K. Louis. Incomplete data problems in x-ray computerized tomography. I. Singular value decomposition of the limited angle transform. *Numerische Mathematik*, 48(3):251–62, 1986.
[163] F. Natterer. Numerical treatment of ill-posed problems. In G Talenti, editor, *Inverse Prob.*, volume 1225, pages 142–67. Berlin, Springer, 1986. Lecture Notes in Math.
[164] R. C. Liu and L. D. Brown. Nonexistence of informative unbiased estimators in singular problems. *Ann. Stat.*, 21(1):1–13, March 1993.
[165] J. Ory and R. G. Pratt. Are our parameter estimators biased? The significance of finite-different regularization operators. *Inverse Prob.*, 11(2):397–424, April 1995.
[166] I. M. Johnstone. On singular value decompositions for the Radon Transform and smoothness classes of functions. Technical Report 310, Dept. of Statistics, Stanford, January 1989.
[167] M. F. Smith, C. E. Floyd, R. J. Jaszczak, and R. E. Coleman. Reconstruction of SPECT images using generalized matrix inverses. *IEEE Trans. Med. Imag.*, 11(2):165–75, June 1992.
[168] M. Lavielle. A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data. *Signal Processing*, 42(1):3–17, 1995.
[169] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. *Numerical recipes in C.* Cambridge Univ. Press, New York, 1988.
[170] K. Sauer and C. Bouman. Bayesian estimation of transmission tomograms using local optimization operations. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 3, pages 2089–93, 1991.
[171] K. Sauer and C. Bouman. Bayesian estimation of transmission tomograms using segmentation based optimization. *IEEE Trans. Nuc. Sci.*, 39(4):1144–52, August

1992.
[172] D. P. Bertsekas and S. K. Mitter. A descent numerical method for optimization problems with nondifferentiable cost functionals. *SIAM J. Control*, 1:637–52, 1973.
[173] W. C. Davidon. Variable metric methods for minimization. Technical Report ANL-5990, AEC Research and Development Report, Argonne National Laboratory, USA, 1959.
[174] H. F. Khalfan, R. H. Byrd, and R. B. Schnabel. A theoretical and experimental study of the symmetric rank-one update. *SIAM J. Optim.*, 3(1):1–24, 1993.
[175] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Software*, 23(4):550–60, December 1997.
[176] T. G. Kolda, D. P. O'Leary, and L. Nazareth. BFGS with update skipping and varying memory. *SIAM J. Optim.*, 8(4):1060–83, 1998.
[177] K. Lange. *Numerical analysis for statisticians*. Springer-Verlag, New York, 1999.
[178] B. T. Polyak. *Introduction to optimization*. Optimization Software Inc, New York, 1987.
[179] D. P. Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic-Press, New York, 1982.
[180] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comp.*, 16:1190–1208, 1995.
[181] L. Kaufman. Reduced storage, Quasi-Newton trust region approaches to function optimization. *SIAM J. Optim.*, 10(1):56–69, 1999.
[182] M. Hanke, J. G. Nagy, and C. Vogel. Quasi-Newton approach to nonnegative image restorations. *Linear Algebra and its Applications*, 316(1):223–36, September 2000.
[183] J. L. Morales and J. Nocedal. Automatic preconditioning by limited memory Quasi-Newton updating. *SIAM J. Optim.*, 10(4):1079–96, 2000.
[184] R. R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *J. Comput. System. Sci.*, 12(1):108–21, 1976.
[185] L. Kaufman. Implementing and accelerating the EM algorithm for positron emission tomography. *IEEE Trans. Med. Imag.*, 6(1):37–51, March 1987.
[186] N. H. Clinthorne, T. S. Pan, P. C. Chiao, W. L. Rogers, and J. A. Stamos. Preconditioning methods for improved convergence rates in iterative reconstructions. *IEEE Trans. Med. Imag.*, 12(1):78–83, March 1993.
[187] J. A. Fessler and S. D. Booth. Conjugate-gradient preconditioning methods for shift-variant PET image reconstruction. *IEEE Trans. Im. Proc.*, 8(5):688–99, May 1999.
[188] 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 Trans. Med. Imag.*, 13(3):687–701, December 1994.
[189] E. Ü Mumcuoğlu, R. M. Leahy, and S. R. Cherry. Bayesian reconstruction of PET images: methodology and performance analysis. *Phys. Med. Biol.*, 41(9):1777–1807, September 1996.
[190] J. A. Fessler and A. O. Hero. Space-alternating generalized expectation-maximization algorithm. *IEEE Trans. Sig. Proc.*, 42(10):2664–77, October 1994.
[191] J. A. Fessler and A. O. Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Trans. Im. Proc.*, 4(10):1417–29, October 1995.
[192] J. A. Fessler, E. P. Ficaro, N. H. Clinthorne, and K. Lange. Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction. *IEEE Trans. Med. Imag.*, 16(2):166–75, April 1997.
[193] J. A. Fessler and A. O. Hero. Space-alternating generalized EM algorithms for penalized maximum-likelihood image reconstruction. Technical Report 286, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, February 1994.
[194] J. A. Browne and A. R. De Pierro. A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography. *IEEE Trans. Med. Imag.*, 15(5):687–99, October 1996.
[195] C. L. Byrne. Block-iterative methods for image reconstruction from projections. *IEEE Trans. Im. Proc.*, 5(5):792–3, May 1996.
[196] C. L. Byrne. Convergent block-iterative algorithms for image reconstruction from inconsistent data. *IEEE Trans. Im. Proc.*, 6(9):1296–304, September 1997.
[197] C. L. Byrne. Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative methods. *IEEE Trans. Im. Proc.*, 7(1):100–9, January 1998.
[198] M. E. Daube-Witherspoon and G. Muehllehner. An iterative image space reconstruction algorithm suitable for volume ECT. *IEEE Trans. Med. Imag.*, 5(2):61–66, June 1986.
[199] J. M. Ollinger. Iterative reconstruction-reprojection and the expectation-maximization algorithm. *IEEE Trans. Med. Imag.*, 9(1):94–8, March 1990.
[200] A. R. De Pierro. On the relation between the ISRA and the EM algorithm for positron emission tomography. *IEEE Trans. Med. Imag.*, 12(2):328–33, June 1993.
[201] P. J. Green. Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Trans. Med. Imag.*, 9(1):84–93, March 1990.
[202] P. J. Green. On use of the EM algorithm for penalized likelihood estimation. *J. Royal Stat. Soc. Ser. B*, 52(3):443–452, 1990.
[203] K. Sauer and C. Bouman. A local update strategy for iterative reconstruction from projections. *IEEE Trans. Sig. Proc.*, 41(2):534–48, February 1993.

[204] K. G. Murty. *Linear complementarity, linear and nonlinear programming*. Helderman Verlag, Berlin, 1988.
[205] 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.*, volume 5, pages 2907–10, 1995.
[206] C. A. Bouman and K. Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Trans. Im. Proc.*, 5(3):480–92, March 1996.
[207] J. A. Fessler. Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans. *IEEE Trans. Im. Proc.*, 4(10):1439–50, October 1995.
[208] H. Erdoğan and J. A. Fessler. Monotonic algorithms for transmission tomography. *IEEE Trans. Med. Imag.*, 18(9):801–14, September 1999.
[209] C. A. Johnson, J. Seidel, and A. Sofer. Interior point methodology for 3-D PET reconstruction. *IEEE Trans. Med. Imag.*, 19(4):271–85, April 2000.
[210] K. Lange and J. A. Fessler. Globally convergent algorithms for maximum a posteriori transmission tomography. *IEEE Trans. Im. Proc.*, 4(10):1430–8, October 1995.
[211] J. A. Fessler, N. H. Clinthorne, and W. L. Rogers. On complete data spaces for PET reconstruction algorithms. *IEEE Trans. Nuc. Sci.*, 40(4):1055–61, August 1993.
[212] 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.*, volume 2, pages 1132–5, 1998.
[213] 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.*, pages 458–66, 1989.
[214] T. Hebert and R. Leahy. A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors. *IEEE Trans. Med. Imag.*, 8(2):194–202, June 1989.
[215] T. J. Hebert and R. Leahy. Statistic-based MAP image reconstruction from Poisson data using Gibbs priors. *IEEE Trans. Sig. Proc.*, 40(9):2290–303, September 1992.
[216] E. Tanaka. Utilization of non-negativity constraints in reconstruction of emission tomograms. In S L Bacharach, editor, *Information Processing in Medical Im.*, pages 379–93. Martinus-Nijhoff, Boston, 1985.
[217] R. M. Lewitt and G. Muehllehner. Accelerated iterative reconstruction for positron emission tomography based on the EM algorithm for maximum likelihood estimation. *IEEE Trans. Med. Imag.*, 5(1):16–22, March 1986.
[218] T. Hebert, R. Leahy, and M. Singh. Three-dimensional maximum-likelihood reconstruction for an electronically collimated single-photon-emission imaging system. *J. Opt. Soc. Am. A*, 7(7):1305–13, July 1990.
[219] S. Holte, P. Schmidlin, A. Lindén, G. Rosenqvist, and L. Eriksson. Iterative image reconstruction for emission tomography: A study of convergence and quantitation problems. *IEEE Trans. Nuc. Sci.*, 37(2):629–635, April 1990.
[220] D. P. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM J. Optim.*, 7(4):913–26, November 1997.
[221] R. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 255–68. Kluwer, Dordrencht, 1998.
[222] A. Nedic and D. Bertsekas. Convergence rate of incremental subgradient algorithms. In S. Uryasev and P. M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, pages 263–304. Kluwer, New York, 2000.
[223] A. Nedic, D. Bertsekas, and V. Borkar. Distributed asynchronous incremental subgradient methods. In D. Butnariu Y. Censor S. Reich, editor, *Inherently Parallel Algorithms in Feasibility and Optimization and Their Applications*. Elsevier, Amsterdam, 2000.
[224] A. Nedic and D. P. Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM J. Optim.*, 12(1):109–38, 2001.
[225] V. M. Kibardin. Decomposition into functions in the minimization problem. *Avtomatika i Telemekhanika*, 9:66–79, September 1979. Translation: p. 1311-23 in Plenum Publishing Co. "Adaptive Systems".
[226] H. Kudo, H. Nakazawa, and T. Saito. Convergent block-iterative method for general convex cost functions. In *Proc. Intl. Mtg. on Fully 3D Image Recon. in Rad. and Nuc. Med*, pages 247–250, 1999.
[227] A. R. De Pierro and M. E. B. Yamagishi. Fast EM-like methods for maximum 'a posteriori' estimates in emission tomography. *IEEE Trans. Med. Imag.*, 20(4):280–8, April 2001.
[228] S. Ahn and J. A. Fessler. Globally convergent image reconstruction for emission tomography using relaxed ordered subsets algorithms. *IEEE Trans. Med. Imag.*, 22(5):613–26, May 2003.
[229] S. Ahn and J. A. Fessler. Globally convergent ordered subsets algorithms: Application to tomography. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 2, pages 1064–8, 2001.

[230] P. Khurd, I-T. Hsiao, A. Rangarajan, and G. Gindi. A globally convergent regularized ordered-subset EM algorithm for list-mode reconstruction. *IEEE Trans. Nuc. Sci.*, 51(3):719–25, June 2004.
[231] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero. Convergent incremental optimization transfer algorithms: Application to tomography. *IEEE Trans. Med. Imag.*, 25(3):283–96, March 2006.
[232] S. Ahn, J. A. Fessler, D. Blatt, and A. O. Hero. Incremental surrogates algorithms: Application to transmission tomography. In *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, volume 5, pages 2835–9, 2004.
[233] H. Erdoğan and J. A. Fessler. Ordered subsets algorithms for transmission tomography. *Phys. Med. Biol.*, 44(11):2835–51, November 1999.
[234] 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*, volume 2, pages 685–9, 1998.
[235] J. Nuyts and J. A. Fessler. A penalized-likelihood image reconstruction method for emission tomography, compared to post-smoothed maximum-likelihood with matched spatial resolution. *IEEE Trans. Med. Imag.*, 22(9):1042–52, September 2003.
[236] J. Qi and R. M. Leahy. Resolution and noise properties of MAP reconstruction for fully 3D PET. In *Proc. Intl. Mtg. on Fully 3D Image Recon. in Rad. and Nuc. Med*, pages 35–9, 1999.
[237] T. H. Farquhar, J. Llacer, C. K. Hoh, J. Czernin, S. S. Gambhir, M. A. Seltzer, D. H. Silverman, J. Qi, C. Hsu, and E. J. Hoffman. ROC and localization ROC analyses of lesion detection in whole-body FDG PET: effects of acquisition mode, attenuation correciton and reconstruction algorithm. *J. Nuc. Med.*, 40(12):2043–52, December 1999.
[238] P. Bonetto, J. Qi, and R. M. Leahy. Covariance approximation for fast and accurate computation of channelized Hotelling observer statistics. *IEEE Trans. Nuc. Sci.*, 47(4):1567–72, August 2000.
[239] J. Qi. Theoretical evaluation of the detectability of random lesions in Bayesian emission reconstruction. In *Information Processing in Medical Im.*, pages 354–65, 2003.
[240] P. K. Khurd and G. R. Gindi. LROC model observers for emission tomographic reconstruction. In *Proc. SPIE 5372, Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment*, pages 509–20, 2004.
[241] J. Qi and R. H. Huesman. Fast approach to evaluate MAP reconstruction for lesion detection and localization. In *Proc. SPIE 5372, Medical Imaging 2004: Image Perception, Observer Performance, and Technology Assessment*, pages 273–82, 2004.
[242] J. Qi. Analysis of lesion detectability in Bayesian emission reconstruction with nonstationary object variability. *IEEE Trans. Med. Imag.*, 23(3):321–9, March 2004.

The literature on image reconstruction is enormous and growing. Many valuable publications are not included in this list, which is not intended to be comprehensive.

Slides and lecture notes available from:

`http://www.eecs.umich.edu/~fessler`