DECEMBER 1988 LIDS-P-1841

A GEOMETRIC PROJECTION-SPACE RECONSTRUCTION ALGORITHM *

Jerry L. Prince
Alan S. Willsky

Laboratory for Information and Decision Systems
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology, Cambridge, MA 02139

December 13, 1988

Abstract

We present a method to reconstruct images from finite sets of noisy projections that may be available only over limited or sparse angles. The algorithm calculates the maximum *a posteriori* (MAP) estimate of the full sinogram (which is an image of the 2-D Radon transform of the object) from the available data. It is implemented using a primal-dual constrained optimization procedure that solves a partial differential equation in the primal phase with an efficient local relaxation algorithm and uses a simple Lagrange multiplier update in the dual phase. The sinogram prior probability is given by a Markov random field (MRF) that includes information about the mass, center of mass, and convex hull of the object, and about the smoothness, fundamental constraints, and periodicity of the 2-D Radon transform. The object is reconstructed using convolution backprojection applied to the estimated sinogram. We show several reconstructed objects which are obtained from simulated limited-angle and sparse-angle data using the described algorithm, and compare these results to images obtained using convolution backprojection directly.

*This research was supported by the National Science Foundation grant ECS-87-00903 and the U.S. Army Research Office grant DAAL03-86-K-1071. In addition, the work of the first author was partially supported by a U.S. Army Research Office Fellowship.

1 INTRODUCTION

This paper addresses the problem of ***reconstruction from projections***, the theoretical and practical aspects of which have received much attention over the past two decades. Among the applications that use the currently available reconstruction techniques are medicine, optics, material science, astronomy, geophysics, and magnetic resonance imaging [1]. The most widely known application of this theory is the problem of medical transmission X-ray tomography [2]. In this discipline, "pencil beam" X-rays are fired from many angles through a single cross section of the body, effectively measuring ***line integrals*** of the 2-D X-ray density function corresponding to the various tissues in the cross section. A collection of line integrals obtained over lines with the same angle but different lateral positions forms a 1-D function called a ***projection.*** Given a set of projections taken from many different angles, an image of the density function may be reconstructed and used in diagnosis. For many medical conditions this tomographic approach to imaging of the body leads to greatly improved imagery over conventional (chest-type) X-ray images and has proven to be of great benefit in medical diagnosis [3].

Consider a function $f(x)$ defined on the plane as depicted in Fig. 1. We denote the integral of f along the line $L(t,\theta)$ by $g(t,\theta)$. The function g for all values of t and θ is called the ***2-D Radon transform*** of f, and an image of $g(t,\theta)$, with t in the vertical direction and θ in the horizontal direction, is called a ***sinogram.*** For a single value of θ, g is a function of t only, and is called the ***projection*** of f at angle θ. The Radon transform of $f(x)$ may be written as

$$g(t,\theta) = \int_{x\in\mathbb{R}^2} f(x)\delta(t - x\cdot\omega)\,dx \tag{1}$$

where $\omega = [\cos\theta \ \ \sin\theta]^{\mathrm{T}}$, and $\delta(\cdot)$ is the Dirac delta function. It turns out that only certain functions $g(t,\theta)$ are valid Radon transforms; there are inherent mathematical consistency conditions that constrain $g(t,\theta)$ to lie in a particular functional subspace defined on the cylinder $\mathbb{R}^1 \times S^1$.

The fact that (1) is invertible (for a wide class of functions) is well known. Deans [4] describes many of the known exact inversion formulas. Except under certain (usually impractical) circumstances, however, it is not possible to determine f exactly given the value of g for only a *finite* number of lines $L(t,\theta)$. It has been the primary concern of engineers and physicists in this field, over the last 20 years or so, to study approximate inversion algorithms given such a finite measurement set. The performance of any particular algorithm depends on the nature of the measurements – their number and arrangement, and their noise properties – and often on the nature of the object itself.

In this paper, we are concerned with the case of low signal to noise ratio (SNR) and limited-angle or sparse-angle measurement configurations with parallel-ray projections. In the medical CT problem, for example, a line integral measurement may be noisy if low energy X-rays are used. Data acquisition may be restricted to a limited angular range if there is an obstruction, for example, and may be sparsely sampled if there is a requirement to obtain the data extremely rapidly. In these situations, images reconstructed using conventional methods are degraded by a variety of artifacts [5], and alternate methods must be used.

Several investigators have developed algorithms to compensate for some of the data deficiencies described above. The modified transform methods of [6], [7], and [8] account for missing projections but do not explicitly address the the issue of noise in the data. Finite-series expansion methods use additional criteria such as minimum norm [9], minimum variance [10,11], and maximum entropy [12] to account for noise, or missing projections, or both. In many cases, streaking artifacts are still present in the reconstructed images [13] and, in all but certain very special imaging geometries, such as in [14], the computations are very time consuming and, hence, impractical.

Other researchers have developed methods to incorporate explicit geometric information about objects. The method of ***projection onto convex sets*** (POCS) [15,16] incorporates prior knowledge by sequentially projecting candidate estimates onto a collection of convex sets, where each set represents some prior knowledge. Noise in the data tends to cause the method to diverge, however, and even though it is possible to account for the noise using a smoothing operator [17], finite pixel error caused by iteration between Radon space and object space may still cause convergence to the wrong solution [14]. Other investigators avoid the latter problem by iterating entirely in projection space [18,19]. Another approach to geometric modeling is to parameterize the object directly in the class of interest, reducing the number of parameters that must be estimated. The work by Rossi and Willsky [20,21], Bresler and Macovski [22,23,24], Hanson [25], and more recently Soumekh [26], Horn [27], and Fishburn et. al. [28] are examples of this kind of modeling.

Our approach is to treat the noise, the physical imaging geometry, and prior probabilistic information as fundamental and explicitly modeled pieces of an overall inverse problem formulation. Our solution satisfies the maximum *a posteriori* (MAP) criterion and incorporates the following prior geometric information:

- The values of line integrals taken over lines close in either lateral displacement (with the same angle) or angle (with the same lateral displacement) tend to be similar in value.

- The Radon transform of any cross section obeys certain fundamental mathematical consistency conditions. In particular, these conditions prescribe constraints on the mass and center of mass of each projection.

- The convex support of the cross-section density function uniquely specifies a related region of support of the Radon transform.

Since both the primary processing and the introduction of prior knowledge take place in Radon space, our approach is a *projection-space* method. This takes advantage of the fact that the noise is well-modeled as white in this domain, so that the criterion of statistical optimality is easily specified, but has the disadvantage that prior information about the object is not conveniently incorporated in Radon space. For example, a local prior probabilistic model of the object is decidedly non-local when transformed into Radon space. We circumvent this problem by modeling directly in projection space, using a Markov random field (MRF) model of sinograms that incorporates the three properties listed above.

Because of the particular form of the chosen MRF, we are able to formulate an analogous problem on the sinogram continuum (as opposed to the usual lattice system), which leads to a closed-form solution given by a partial differential equation (PDE) with constraints. We solve this constrained PDE using a primal-dual optimization approach that solves the PDE with assumed Lagrange multipliers in the primal phase and updates the Lagrange multipliers in the dual phase. The primal phase is fast (although iterative) and parallelizable, due to the local interactions of the MRF; the dual phase is fast and partially parallelizable, since it uses a simple update formula on each of the columns of the sinogram separately.

The paper is organized as follows. In Section 2 we present additional background related to the support and consistency of the 2-D Radon transform. Section 3 develops a Markov random field model of sinograms, and formally defines the maximum *a posteriori* solution. In Section 4, we present a fast iterative solution that solves this large-scale optimization problem and Section 5 presents some experimental results. Finally, we discuss these results and some outstanding problems in Section 6.

2 CONSISTENCY AND CONVEX SUPPORT

2.1 Consistency of the Radon Transform

An important fact that we exploit in this paper is that not all functions $g : \mathbb{R}^1 \times S^1 \rightarrow \mathbb{R}^1$ are valid 2-D Radon transforms. A valid Radon transform, that is, a function that is the Radon transform of some function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^1$, is constrained to lie in a particular functional subspace of the space of all real functions. This subspace is characterized by the property that g is even in t and ω and by the property that certain generalized Fourier coefficients of g must be zero. The precise mathematical conditions for the consistency are given by the following theorem due to Ludwig [31].

Theorem 1 (2-D Consistency Theorem) *In order for $g(t,\theta)$ to be the 2-D Radon transform of a function $f \in \mathcal{S}(\mathbb{R}^2)$, where $\mathcal{S}$ is the space of rapidly decreasing C^∞ functions, it is necessary and sufficient that*

(a) $g \in \mathcal{S}(\mathbb{R}^1 \times S^1)$,

(b) $g(t, \theta + \pi) = g(-t, \theta)$, *and*

(c) the integral

$$\int_{-\infty}^{\infty} g(t,\theta)t^k \, dt \tag{2}$$

be a homogeneous polynomial of degree k in $\cos\theta$ and $\sin\theta$ for all $k \geq 0$.

Proof *See [31], [30], or [32].* □

The two lowest order moments of $g(t,\theta)$ give the mass and center of mass constraints. The mass constraint tells us that the integral of any projection, which may be thought of as the mass of the projection, must have the same value for any θ, and that value is equal to the integral of $f(x)$. If, for example, a noisy measurement of a true Radon transform has any two projections that do not integrate to the same value *then the measurement is not a valid Radon transform*, and it follows that an inverse transform is theoretically undefined. The center of mass constraint tells us that the (1-D) center of mass of a given projection is equal to the projection of the (2-D) center of mass of the object onto the ω-axis. From this one can see that the collection of centers of mass of the projections for different θ must be a cosinusoidal function with period 2π. If that is not true for a given measurement then, again, the measurement is not a valid Radon transform. These two facts are easily shown using the Consistency Theorem, and may be stated as

$$m = \int_{-\infty}^{\infty} g(t,\theta)\, dt = \int_{x\in\mathbf{R}^2} f(x)\, dx \quad \forall\theta \tag{3}$$

and

$$c(\theta) = \frac{1}{m}\int_{-\infty}^{\infty} g(t,\theta)t\, dt = a\cos\theta + b\sin\theta\,, \tag{4}$$

for some real constants a and b. We refer to m as the *mass* of $f(x)$ and (3) as the *mass constraint* for the 2-D Radon transform; the quantity $c(\theta)$ is the *center of mass* of projection $g(t,\theta)$, and equation (4) is the *center of mass constraint.* It is also true that the center of mass of the projection $g(t,\theta)$ is the projection of the 2-D center of mass of $f(x)$ (see [32]) and, indeed, if (R,ϕ) denotes the polar coordinates of the center of mass of the object, then it can be shown that [20]

$$c(\theta) = R\cos(\theta - \phi)\ .$$

Given the above development, we see that if the mass and center of mass were known *a priori* then (3) and (4) should be imposed as constraints on the estimated sinogram. The center of mass constraint $c(\theta) = 0$, imposed in Section 3, implies that the object is centered at the origin. Given a known center of mass it is possible to adjust the observed sinogram (by shifting each of the projections in t) to make it appear as if the object were centered at the origin. This adjustment may always be accomplished provided one has a field of view large enough to encompass both the original object and the object shifted to the origin. We assume this to be the case.

2.2 Object Support and Radon Transform Support

The convex support of an object is smallest convex set that supports the function $f(x)$. In this section we develop a relationship between the convex support of an object and a particular region of support of its 2-D Radon transform. This relationship is a special case of the support theorem stated and proved by Lax and Phillips in [29] and also discussed by Helgason [30].

Suppose f is zero outside D_T, the disk of radius T centered at the origin. Then it is easy to see from the definition of the 2-D Radon transform in (1) that $g(t, \theta)$ must be zero when $t \notin [-T, T]$. Using the periodicity of the 2-D Radon transform established in Theorem 1, one can now conclude that $g(t, \theta)$ is completely determined by its values on the rectangle

$$\mathcal{Y}_T = \{(t, \theta) \mid -T \leq t \leq T, 0 \leq \theta \leq \pi\}. \qquad (5)$$

But this idea can be refined even further. Let $\mathcal{F}$ be the set of points in Y_T for which $f(x) \neq 0$. Now consider the Radon transform $g(t, \theta)$ of f, and the unit vector $\omega = [\cos\theta \;\; \sin\theta]^{\mathrm{T}}$. With reference to Fig. 2 and to (1), we see that for any given ω, the value of the Radon transform must be zero for $t \geq t_+$ and $t \leq t_-$. Here, t_+ is the lateral position of the line perpendicular to ω which is positioned as far as possible in the $+\omega$ direction so it just grazes the set $\mathcal{F}$; t_- is the lateral position of the line perpendicular to ω which is positioned as far a possible in the $-\omega$ direction so it just grazes the set $\mathcal{F}$. The quantities t_+ and t_- are called ***support values*** and the corresponding lines are called ***support lines*** of the set $\mathcal{F}$. Knowledge of both

t_+ and t_- for all θ in $[0, \pi)$ determines the convex hull of $\mathcal{F}$, denoted hul($\mathcal{F}$), which is, by definition, the smallest convex set containing $\mathcal{F}$. The set hul($\mathcal{F}$) is also the convex support of the function $f(x)$.

From the above discussion, we conclude that the 2-D Radon transform is completely determined by its values on the set

$$\mathcal{G} = \{(t, \theta) \in \mathcal{Y}_T \mid t_-(\theta) \leq t \leq t_+(\theta)\} \tag{6}$$

where, for clarity, we have explicitly indicated the functional dependence of t_+ and t_- on θ. An example of such a set is shown in Fig. 3. For a given object support set $\mathcal{F}$, we think of $\mathcal{G}$ as the matching region of support in Radon space. However, although $\mathcal{F}$ uniquely determines $\mathcal{G}$, it is clear that $\mathcal{G}$ determines only hul($\mathcal{F}$), not $\mathcal{F}$ itself. Furthermore, $\mathcal{G}$ is not necessarily the actual support of $g(t, \theta)$ since it is possible for $g(t, \theta)$ to be zero when $(t, \theta) \in \mathcal{G}$ if $\mathcal{F}$ is not connected. We are primarily concerned with the convex support of f, since this is what may be determined directly from knowledge of $\mathcal{G}$.

In Section 3 we assume that an estimate of $\mathcal{G}$ is available, and we define a prior probability on sinograms that gives a low probability to sinograms that have non-zero values outside of $\mathcal{G}$.

3 SINOGRAM MRF AND MAP ESTIMATION

We chose to represent prior probabilistic knowledge about sinograms using a Markov random field (MRF) on a discrete sinogram lattice. There are several reasons for this choice. First, the MRF is a convenient way to describe processes with local interaction in such a way that the joint probability over all sites is easily determined. Second, the constraints that arise from the 2-D Radon transform consistency conditions are easily incorporated in the MRF by limiting the space of allowable configurations. Third, knowledge of the convex support of the object, which is treated as a penalty rather than a constraint, may be incorporated into the MRF by adding an additional self-potential term (see below). Finally, this choice, along with the particular details described below, leads directly to a statistically optimum maximum *a posteriori* solution. We will see in the following section that the form of this estimate has an analogous variational formulation that, once solved, leads to a fast iterative solution.

3.1 A Sinogram MRF

This section develops a Markov Random field (MRF) on the sinogram lattice that includes the mass and center of mass constraints, and that includes the periodicity and smoothness of the 2-D Radon transform and the convex support of the object. The ingredients needed to define a MRF are [33]: 1) the lattice, 2) the potential functions, 3) the graph structure, and 4) the feasible configurations.

The Sinogram Lattice As discussed in Sections 1 and 2, a sinogram is an image of the Radon transform (or measured Radon transform) of an object over the truncated domain $\mathcal{Y}_T$ (see equation (5)), with brighter intensities corresponding to larger values of $g(t, \theta)$. In order to define a MRF, however, we require a finite lattice system, rather than the continuum of points in $\mathcal{Y}_T$. Therefore, we define the *sinogram lattice* to be a rectangularly sampled version of $\mathcal{Y}_T$ given by

$$\mathcal{Y}_S = \{(t,\theta) \mid t = \frac{2T}{n_d} i, i = -\frac{n_d - 1}{2}, \dots, \frac{n_d - 1}{2},\ \theta = \frac{\pi}{n_v} j, j = 0, \dots, n_v\} \tag{7}$$

where n_d is the number of sample points in t, and n_v is the number of sample points in θ.

For convenience we adopt the following notation for the remainder of this section. A *site* in the sinogram lattice is denoted $s = (i, j)$ and the set of all sites by S. A *site value* is denoted in several ways: $g_s = g_{ij} = g(t_i, \theta_j)$. The collection of all site values is called the *discrete sinogram* or just the *sinogram* when the meaning is clear from the context. Note, that a site in the sinogram lattice corresponds to a line in the plane passing through the disk D_T. In particular, the site (i, j) corresponds to the line $L(t_i, \theta_j) = \{(x, y) \in \mathbb{R}^2 \mid x \cos\theta_j + y \sin\theta_j = t_i\}$.

The Potential Functions The physics of this problem does not specify for us a prior probability on sinograms. We rely on experimentation and intuition to surmise what a reasonable form for a prior might be, given what we know about the types of objects under consideration and the transformation of those objects via the Radon transform. After much thought and experimentation, one key idea has driven us to implement what turns out to be the simplest kind of MRF. This idea is simply that sinograms tend to be locally smooth. A prior that produces such sample functions is an MRF with potential functions that prescribe an affinity between nearest neighbors — this is the so-called nearest-neighbor "blob" model [34]. Let $s, r \in S$ denote sites that are either vertical or horizontal nearest neighbors. To prescribe affinities between the sinogram values defined on these sites we define the vertical pair-potential function as

$$V_{(s,r)} = b_v(g_s - g_r)^2 \tag{8}$$

and the horizontal pair-potential function as

$$V_{\langle s,r \rangle} = b_h(g_s - g_r)^2 \tag{9}$$

where (s, r) and $\langle s, r \rangle$ represent pairs of adjacent sites in the vertical and horizontal directions, respectively. The positive constants b_v and b_h allow one to make the vertical and horizontal affinities of different strength, thus making this a non-isotropic random field [35]. In this paper, we choose the constants b_v and b_h *a priori*; however, it is possible to use the actual data to estimate these coefficients as part of a hierarchical estimation algorithm [36].

The self-potentials are defined using knowledge of the object's support. As developed in Section 2, the object's support $\mathcal{F}$ implies a matching region of support $\mathcal{G}$ within the Radon transform domain $\mathcal{Y}_T$. If either set were known exactly then we would insist that sinogram values in the region

$$\bar{\mathcal{G}} = \mathcal{Y}_T - \mathcal{G} \tag{10}$$

be exactly zero. However, we shall assume that only an *estimate* of the sinogram support is available, and that we have a measure of that estimate's reliability. Therefore, sinograms with non-zero values in $\bar{\mathcal{G}}$ should have low probability, but not as low if the support estimate is unreliable. To provide this effect, we define the support self-potential as

$$V_s = \begin{cases} \kappa g_s^2 \,, & (t_i, \theta_j) \in \bar{\mathcal{G}} \\ 0 \,, & \text{otherwise} \end{cases} \tag{11}$$

where g_s stands for the value of the sinogram at site s=(i,j), and κ is a positive constant which is used to reflect the support measurement's reliability.

The Graph Structure The form of the potential functions described above dictate the required neighborhood structure, and, in fact, only nearest neighbors are necessary. In this case, the most general form of the MRF energy function is [33]

$$U(g) = \sum_{s} V_s(g_s) + \sum_{(s,r)} V_{(s,r)}(g_s, g_r) + \sum_{\langle s,r \rangle} V_{\langle s,r \rangle}(g_s, g_r) \tag{12}$$

where the first summation is taken over all sites in the lattice, the second over all vertical nearest-neighbors, and the third over all horizontal nearest-neighbors.

Since all objects are known to be zero outside the disk of radius T, the boundary value above and below the sinogram must be zero as shown in Fig. 4a. The boundaries to the left and right of the sinogram must be treated differently, however. Here, we use the symmetry property of the consistency theorem stated in Section 2

$$g(t,\theta) = g(-t, \theta + \pi)$$

to conclude that the neighbors wrap around in a toroid that is twisted or flipped about the t-axis as shown in Fig. 4b. In other words, the sinogram is actually defined on a Mobius strip. Given these boundaries conditions, pair potentials involving one site outside of the lattice may now be computed. Sites above or below the sinogram are given the fixed value zero, whereas sites to the left or right are given the value of the sinogram on the opposite side of the sinogram and in the opposite t direction.

The Feasible Configurations The mass and center of mass constraints used for the MRF are discrete approximations to the integral expressions of (3) and (4). Thus, letting m denote the object's mass we have

$$\frac{2T}{n_d} \sum_{i=1}^{n_d} g_{ij} = m \quad \forall j, \ 1 \le j \le n_v \tag{13}$$

and

$$\frac{1}{m} \frac{2T}{n_d} \sum_{i=1}^{n_d} t_i g_{ij} = 0 \quad \forall j, \ 1 \le j \le n_v \tag{14}$$

where

$$t_i = \frac{2T}{n_d}(i - \frac{n_d + 1}{2}) \tag{15}$$

is the lateral position of the ith line.

The presence of constraints, even linear equality constraints such as these, makes the computation of the MAP estimate more difficult since the algorithm must be a constrained optimization method [37]. In fact, the solution must be an element of a set of feasible discrete sinograms Ω_g, which contains all real matrices of dimension n_d by n_v that satisfy (13) and (14).

3.2 The MAP Formulation

The Gibbs Prior Having now defined all the elements of the MRF, the joint probability density for the discrete sinogram prior is simply given by the Gibbs density

$$p(g) = \tfrac{1}{Z}e^{-U(g)} \quad g \in \Omega_g \tag{16}$$

where g denotes the vector of sinogram site values and Z is given by

$$Z = \int_{g \in \Omega_g} e^{U(g)}\, dg$$

so that $p(g)$ integrates to one. The function $U(g)$ is the energy function defined in (12).

The Observations We assume that noisy observations of the true site values are available over a (possibly) limited-angle or sparse-angle subset $\mathcal{Y}_O$ of $\mathcal{Y}_T$ and that the observations are given by

$$y_{ij} = g_{ij} + n_{ij} \quad (t_i, \theta_j) \in \mathcal{Y}_O \quad , \tag{17}$$

where the n_{ij} are independent zero-mean Gaussian random variables with variance σ^2. Letting g denote the vector of true sinogram site values and letting y and n denote the vector of observations and noise samples, respectively, we may write the observation equation in vector form as

$$y = Sg + n \tag{18}$$

where S is a matrix that *selects* the observations as follows. In the measurement configuration we consider, a column of the matrix given by $[g_{ij}]$ is either observed completely (in additive noise) or not observed at all. Suppose, for the purposes of this discussion only, we form g by stacking the columns of $[g_{ij}]$, stacking all the *observed* columns first, from the top proceeding downwards, and the remaining columns following in any order. Then, denoting the number of observed columns by n_o, we see that S is given by

$$S = [\, I \,|\, 0 \,]$$

where I is the $n_o n_d \times n_o n_d$ identity matrix.

The Sinogram MAP Estimate Now, with the observation equation given by (18) and the prior probability given by (16), we may now derive the form of the MAP estimate $\hat{g}_{map}$. Denoting the noise covariance matrix by K_n we may use (18) to write the conditional measurement density (zero mean, jointly Gaussian) as

$$p(y|g) = |2\pi K_n|^{-1/2} \exp\left(-\frac{1}{2}(y - Sg)^{\mathrm{T}} K_n^{-1}(y - Sg)\right). \tag{19}$$

Using the definition of conditional probability twice, and the prior probability given by (16), the posterior distribution is found to be

$$\begin{aligned} p(g|y) &= p(y|g)p(g)/p(y) & (20)\\ &= \frac{|2\pi K_n|^{-1/2}}{Z\,p(y)} \exp(-(\frac{1}{2}(y - Sg)^{\mathrm{T}} K_n^{-1}(y - Sg) + U(g))) \end{aligned}$$

The MAP estimate, which maximizes (20) with the true observations Y substituted into the expression, is given by

$$\hat{g}_{map} = \operatorname*{argmin}_{g \in \Omega_g} \frac{1}{2\sigma^2}(Y - Sg)^{\mathrm{T}}(Y - Sg) + U(g) \tag{21}$$

where we have used the fact that $K_n = \sigma^2 I$.

The posterior distribution of (20) is a Gibbs density, and since the observation equation is not convolutional, its graph structure is identical to that of the prior. The only significant difference between the two MRF's is the form of the energy function, which, in the posterior density, contains a self potential term that couples the observations y to the sinogram g. The identification of the posterior density as a Gibbs density serves as the basis for the *stochastic relaxation* and *simulated annealing* methods of Geman and Geman [33] and others, algorithms that have been the focus of much research in recent years. These methods, besides being generally very slow, are inappropriate for this application for two reasons. First, the stochastic methods do not conveniently incorporate constraints [32]. Second, and most importantly, the minimization problem of (21) requires minimizing a quadratic function with linear constraints, which when taken advantage of as we do in the next section, leads to a much faster algorithm.

4 Primal-Dual MAP Algorithm

This section develops the theory and implementation of a fast iterative algorithm for computing $\hat{g}_{map}$. The key step in the development this method is to write the vector minimization problem of (21) as a constrained minimization problem involving an unknown function $g(t, \theta)$ over the *continuous* domain $\mathcal{Y}_T$. The solution to this variational formulation is a partial differential equation (PDE) with three unknown functions: the sinogram and two Lagrange multiplier functions. We use a generic primal-dual method to find the solution to this PDE, incorporating a fast iterative local relaxation algorithm in the primal stage and simple Lagrange multiplier update formulas in the dual stage.

4.1 Variational MAP Formulation and Solution

The Minimization Problem Consider the problem, which we refer to as (V), of minimizing

$$I = \iint_{\mathcal{Y}_O} \frac{1}{2\sigma^2}(y-g)^2 \, dt \, d\theta + \iint_{\bar{\mathcal{G}}} \kappa g^2 \, dt \, d\theta + \iint_{\mathcal{Y}_T} \left[\beta \left(\frac{\partial g}{\partial t}\right)^2 + \gamma \left(\frac{\partial g}{\partial \theta}\right)^2\right] dt \, d\theta \qquad (22)$$

subject to the equality constraints

$$\begin{aligned} J_1 &= m = & \int_{-\mathrm{T}}^{\mathrm{T}} g(t, \theta) \, dt \\ J_2 &= 0 = & \frac{1}{m} \int_{-\mathrm{T}}^{\mathrm{T}} t g(t, \theta) \, dt \end{aligned} \qquad (23)$$

and boundary conditions

$$\begin{aligned} g(\mathrm{T},\theta) &= g(-\mathrm{T},\theta)=0 \\ g(t,0) &= g(-t,\pi) \end{aligned} \tag{24}$$

where κ, β, and γ are positive constants. This problem is a continuous formulation of the sinogram MAP problem of (21). The first term in I is analogous to the first term in (21) — both represent a penalty that seeks to keep the estimate close to the observations. The second two terms are analogous to the terms of $U(g)$, given in (12). The first term comprises the support information, which matches the summation over single sites in $U(g)$. The second term integrates the sum of the squares of the two partial derivatives of g, which corresponds to the two summations of pair-potentials in $U(g)$. The two integral constraints in (23) are exactly the mass and center of mass constraints. Finally, the boundary conditions, which include the twisted boundary, are stated in (24).

To simplify notation we define the following indicator function notations:

$$\bar{\chi}_G(t,\theta)=\begin{cases} 1 & (t,\theta)\in\bar{\mathcal{G}} \\ 0 & \text{otherwise} \end{cases} \quad , \tag{25}$$

which indicates the complement of the region of support in the sinogram domain, and

$$\chi_Y(t,\theta)=\begin{cases} 1 & (t,\theta)\in\mathcal{Y}_O \\ 0 & \text{otherwise} \end{cases} \quad , \tag{26}$$

which indicates the region in the sinogram over which observations are available. Using this notation we may write I as

$$I=\iint_{\mathcal{Y}_T}\kappa\bar{\chi}_G g^2+\beta\left(\frac{\partial g}{\partial t}\right)^2+\gamma\left(\frac{\partial g}{\partial \theta}\right)^2+\frac{1}{2\sigma^2}\chi_Y(y-g)^2\,dt\,d\theta \ . \tag{27}$$

The problem is now in the form of a classical variational problem which may be solved using standard calculus of variations techniques (see [38], for example).

Partial Differential Equation A necessary condition for $g(t,\theta)$ to be a solution to (V) is that it satisfy the following second order partial differential equation (PDE) [32]

$$\left(2\kappa\bar{\chi}_G+\frac{1}{\sigma^2}\chi_Y\right)g-2\beta\frac{\partial^2 g}{\partial t^2}-2\gamma\frac{\partial^2 g}{\partial \theta^2}=\frac{1}{\sigma^2}\chi_Y y-\lambda_1(\theta)-\lambda_2(\theta)t \tag{28}$$

and the additional boundary condition

$$\frac{\partial g(t,0)}{\partial t} = \frac{\partial g(-t,\pi)}{\partial t} \ . \tag{29}$$

In addition, $g(t,\theta)$ must satisfy the original constraints and boundary conditions. Note that (28) contains *three* unknown functions: $g(t,\theta)$, and two Lagrange multiplier functions $\lambda_1(\theta)$ and $\lambda_2(\theta)$ (one for each constraint). To simplify the expressions in the remainder of this section we use the notation g_t and g_{tt} to stand for the first and second partial derivatives of $g(t,\theta)$ with respect to t, respectively, and g_θ and $g_{\theta\theta}$ to stand for the first and second partial derivatives of $g(t,\theta)$ with respect to θ, respectively.

To solve (28) for $g(t,\theta)$, λ_1 and λ_2 must first be determined. An analytic expression for λ_1 may be found by integrating both sides of (28) and simplifying; and an analogous expression for λ_2 may be found by multiplying both sides of (28) by t, then integrating and simplifying. The results are [32]

$$\begin{aligned}
\lambda_1(\theta) &= \frac{-1}{2T}\left[\int_{-T}^{T} 2\kappa\bar{\chi}_G g\, dt - 2\beta g_t|_{-T}^{T} + \frac{m}{\sigma^2}\chi_Y - \frac{1}{\sigma^2}\int_{-T}^{T} \chi_Y y\, dt\right] &(30)\\
\lambda_2(\theta) &= \frac{-3}{2T^3}\left[\int_{-T}^{T} 2t\kappa\bar{\chi}_G g\, dt - 2\beta t g_t|_{-T}^{T} - \frac{1}{\sigma^2}\int_{-T}^{T} t\chi_Y y\, dt\right] , &(31)
\end{aligned}$$

These equations may be substituted into (28) to give an integro-differential equation in a single unknown function $g(t,\theta)$.

Primal-Dual Optimization Method Unfortunately, although the resulting integro-differential equation could be discretized and solved numerically, the problem is very large and computationally intractable. The approach we take instead is to use the generic primal-dual method described by Bertsekas in [39] to solve the PDE directly. In outline, the method requires us to solve (28) numerically given *estimated* Lagrange multipliers, update the Lagrange multipliers if the solution doesn't meet the required constraints, and repeat until a jointly optimum trio of $\hat{g}$, $\hat{\lambda}_1$, and $\hat{\lambda}_2$ is found.

To find an initial estimate of the Lagrange multipliers we make several approximations which often hold true at or near the solution. First, near the solution we expect that

$g(t,\theta) \approx 0$ for $(t,\theta) \in \bar{\mathcal{G}}$, especially when κ is large. Hence, we may make the approximations

$$\int_{-T}^{T} 2\kappa\bar{\chi}_G g \, dt \approx 0 \quad \text{and} \quad \int_{-T}^{T} 2t\kappa\bar{\chi}_G g dt \approx 0.$$

Second, the terms in (30) and (31) involving the partial derivative g_t evaluated at $\pm T$ may often be close to be zero. These approximations, applied to (30) and (31), yield the following initial Lagrange multiplier estimates

$$\begin{aligned} \lambda_1^0(\theta) &= \frac{-1}{2T}\left[\frac{m}{\sigma^2}\chi_Y - \frac{1}{\sigma^2}\int_{-T}^{T} \chi_Y y \, dt\right] & (32) \\ \lambda_2^0(\theta) &= \frac{-3}{2T^3}\left[-\frac{1}{\sigma^2}\int_{-T}^{T} t\chi_Y y \, dt\right], & (33) \end{aligned}$$

each of which may be evaluated given only the data. Substituting these functions for the true Lagrange multipliers in (28) yields the PDE

$$\begin{aligned} &\left(2\kappa\bar{\chi}_G + \frac{1}{\sigma^2}\chi_Y\right) g - 2\beta g_{tt} - 2\gamma g_{\theta\theta} = \\ &\quad \frac{1}{\sigma^2}\chi_Y y - \frac{1}{2T\sigma^2}\int_{-T}^{T} \chi_Y y \, dt + \frac{m}{2T\sigma^2}\chi_Y - \frac{3t}{2T^3\sigma^2}\int_{-T}^{T} t\chi_Y y \, dt, \qquad (34) \end{aligned}$$

which, unlike the original, has a single unknown $g(t,\theta)$, and may be solved numerically using any of several techniques for solving elliptic PDEs as discussed below.

If g does not meet the constraints after solving (34), the first primal stage, then we conclude that the approximations used to derive the approximate PDE of (34) were not accurate. This situation will in fact occur if β is large or if the observations are missing entire projections, such as in the limited-angle and sparse-angle problems. Therefore, the dual stage updates the Lagrange multiplier functions to move them closer to their optimum values using the following formulas [39]

$$\begin{aligned} \lambda_1^{k+1}(\theta) &= \lambda_1^k(\theta) + \alpha\left(m - \int_{-T}^{T} g(t,\theta) \, dt\right) & (35) \\ \lambda_2^{k+1}(\theta) &= \lambda_2^k(\theta) + \alpha\left(0 - \frac{1}{m}\int_{-T}^{T} tg(t,\theta) \, dt\right) & (36) \end{aligned}$$

where α is a positive constant, and k is an iteration counter. Note that the update calculations can be done independently, and therefore in parallel, for each projection in the sinogram. After each update, the new Lagrange multiplier functions are substituted into

(28), which is solved numerically for a new $\hat{g}$. When $\hat{g}$ meets the constraints to within a specified tolerance, then the three functions are jointly optimal and $\hat{g}$ is the desired sinogram estimate.

The constant α, which appears in (35) and (36), is chosen large enough so that convergence to the correct Lagrange multipliers (and, hence, the correct solution to (V)) occurs quickly, yet not so large that the sequence will not converge. Bertsekas [39] describes the the selection of α and relates this generic primal-dual method to the method of multipliers, about which a great deal of theory is known. In our experiments, α was chosen empirically to yield a good rate of convergence for our problem.

4.2 Numerical Methods

The sinogram is approximated on a rectilinear grid with vertical and horizontal sample spacings given by $\Delta_t = 2T/n_d$ and $\Delta_\theta = \pi/n_v$, respectively. The PDE of (28) may then be approximated at an interior point by the finite difference equation [40]

$$d_{i,j}g_{i,j} - r_{i,j}g_{i+1,j} - l_{i,j}g_{i-1,j} - t_{i,j}g_{i,j+1} - b_{i,j}g_{i,j-1} = s_{i,j} \tag{37}$$

where

$$\begin{aligned} l_{i,j} &= 2\hat{\beta} \\ r_{i,j} &= 2\hat{\beta} \\ b_{i,j} &= 2\hat{\gamma} \\ t_{i,j} &= 2\hat{\gamma} \\ d_{i,j} &= 4\hat{\beta} + 4\hat{\gamma} + \left(2\kappa\bar{\chi}_G + \frac{1}{\sigma^2}\chi_Y\right)\bigg|_{(t_i,\theta_j)} \\ s_{i,j} &= \left(\frac{1}{\sigma^2}\chi_Y y - \lambda_1(\theta) - \lambda_2(\theta)t\right)\bigg|_{(t_i,\theta_j)} \quad , \end{aligned} \tag{38}$$

and $\hat{\beta} = \beta/\Delta_t^2$ and $\hat{\gamma} = \gamma/\Delta_\theta^2$. Sinogram values in (37) that correspond to points outside the lattice must be evaluated according to the boundary conditions developed in Section 3.

The set of equations given by (37) for all j, $j = 1, \ldots, n_d$ and i, $i = 1, \ldots, n_v$ may be organized and written as a vector equation

$$Ag = s \ , \tag{39}$$

and, although this is a very high dimensional problem, the local interactions that result from the nearest-neighbor construction allow for efficient iterative solutions. Several traditional methods (cf. [41]) including Jacobi, simultaneous over-relaxation (SOR), and Chebyshev semi-iterative relaxation methods may be employed to solve (39). However, we have chosen to implement a relatively new method due to Kuo, Levy, and Musicus [40] which has been shown to have very favorable convergence properties, and is relatively easy to implement. This method, in addition, has been shown to be ideally suited for parallel implementation.

Our implementation of Kuo's local relaxation algorithm (KLR) is a special case of the more general implementation described in [40]. We assume the PDE to be of the form

$$-p\frac{\partial^2 u}{\partial x_1^2} - q\frac{\partial^2 u}{\partial x_2^2} + \varsigma(x_1, x_2)u = f(x_1, x_2),$$

where $(x_1, x_2) \in [0,1] \times [0,1]$, and to satisfy the conditions given in [40]. Then the PDE is approximated by the 5-point stencil

$$d_{i,j}u_{i,j} - ru_{i+1,j} - lu_{i-1,j} - tu_{i,j+1} - bu_{i,j-1} = s_{i,j},$$

with

$$l = p, \quad r = p, \quad b = q, \quad t = q,$$

$$d_{i,j} = 2p + 2q + \varsigma_{i,j}h^2, \quad s_{i,j} = f_{i,j}h^2$$

where h is the grid spacing and $\varsigma_{i,j}$ is defined as $\varsigma(ih, jh)$. Each grid point is assigned a color, either red or black, according to an alternating pattern as on a checkerboard. Then the local relaxation procedure can be written as:

red points ($i + j$ is even):

$$u_{i,j}^{(n+1)} = (1 - \omega_{i,j})u_{i,j}^{(n)} + \omega_{i,j}d_{i,j}^{-1}\left(lu_{i-1,j}^{(n)} + ru_{i+1,j}^{(n)} + bu_{i,j-1}^{(n)} + tu_{i,j+1}^{(n)} + s_{i,j}\right),$$

black points ($i + j$ is odd):

$$u_{i,j}^{(n+1)} = (1 - \omega_{i,j})u_{i,j}^{(n)} + \omega_{i,j}d_{i,j}^{-1}\left(lu_{i-1,j}^{(n+1)} + ru_{i+1,j}^{(n+1)} + bu_{i,j-1}^{(n+1)} + tu_{i,j+1}^{(n+1)} + s_{i,j}\right),$$

where $\omega_{i,j}$ is called the *local relaxation parameter* and is given by

$$\omega_{i,j} = \frac{2}{1 + \sqrt{1 - \rho_{i,j}^2}},$$

where

$$\rho_{i,j} = \frac{2}{d_{i,j}} \left(p \cos \frac{\pi}{M_1 + 1} + q \cos \frac{\pi}{M_2 + 1} \right).$$

One point related to convergence of KLR is worthy of comment. In the initialization phase, KLR calculates an array of *local relaxation parameters*, ω_{ij}, one per site, which are theoretically optimum for a particular boundary condition which our problem does not satisfy (because of the twisted boundary property). Therefore, KLR still converges to the correct solution, but it may do so more slowly than the predicted convergence rates. However, we have found in our experiments that no slow-down is evident; the rate, in practice, is still of order $\sqrt{N}$, where N is the total number of points in the grid.

5 Experimental Results

5.1 Overview

In this section, we present results from several simulation studies, each designed to demonstrate a different aspect of the sinogram MAP algorithms described in Section 4. The object that is used for all of the simulations in this section is an ellipse with the letters M I T in its interior, as shown in Fig. 5. The ellipse is centered at the origin and rotated 45 degrees in the clockwise direction, and has two values: 0 outside of the ellipse and 1 within the body of the ellipse, except within the letters, where the value is 0. The noise-free sinogram shown in Fig. 6a is calculated using approximate strip integrals (see [42]) from analytic expressions for the ellipse and characters in the interior. The sinogram has 81 rows and 60 columns, approximating $g(t, \theta)$ over the angular range $[\pi/2, 3\pi/2)]$ and the lateral range $[-T, T]$. The 81x81 pixel image in Fig. 6b is a reconstruction from the noise-free data of Fig. 6a using convolution backprojection (CBP) with a ramp filter (see [42]).

The MIT ellipse was chosen as a test object for this experiment because the loss of angular information has strikingly different effects depending on where the missing angles occur. For example, if the missing projections have lines that integrate along the long axis of the ellipse, then the narrowness of the ellipse cannot be observed, but the detail of the letters in the interior is quite apparent from the observed projections. If, however, the missing angles occur along the short axis of the ellipse, then the letters cannot be observed well, but the narrowness *is* apparent. The first case is where information about the boundary of the object has a striking effect on the reconstruction; the smoothing effect helps in both cases.

One noisy, two limited-angle, and two sparse-angle cases were derived from the noise-free sinogram and used as simulated observations. Fig. 6c shows a noisy sinogram, created by adding independent samples of zero-mean Gaussian noise with variance σ^2 to each element of the true sinogram of Fig. 6a. The signal to noise ratio (SNR) of this sinogram is 3.0dB, using the following definition of SNR:

$$\mathrm{SNR} = 10 \log \frac{\frac{\pi}{n_v} \frac{2T}{n_d} \sum_{j=1}^{n_v} \sum_{i=1}^{n_d} g^2(t_i, \theta_j)}{\sigma^2}, \tag{40}$$

where $g(t_i, \theta_j)$ is the true sinogram. Fig. 6d shows a reconstruction of Fig. 6c using CBP. Figs. 7a and 7b show sparse-angle reconstructions from, respectively, 15 and 10 evenly spaced projections of a 10.0dB noisy observed sinogram of the ellipse (not shown). This corresponds to projections taken 12.0 and 18.0 degrees apart, respectively. Figs. 7c and 7d are reconstructions from the first (left) 40 projections and the last (right) 40 projections, respectively, of the same 10.0dB noisy sinogram. The first limited-angle arrangement lacks projections with information about the narrow dimension of the ellipse, while the second arrangement lacks detailed information about the letters within the ellipse. Each of these four reconstructions was made using CBP, assigning zero to the missing projections.

Given an observed sinogram (perhaps noisy or only partially observed), the first processing step is to estimate the object mass using

$$\hat{m} = \frac{2T}{J n_d} \sum_{j \in J} \sum_{i=1}^{n_d} y(t_i, \theta_j), \tag{41}$$

where J is the set of observed projections and J is the number of elements in J. Each sinogram is then normalized by dividing each observation by $\hat{m}$ so that the normalized sinogram corresponds to an object with unit mass. This normalization is necessary so that the coefficients β, γ, and κ have the same qualitative effect on low mass sinograms as on high mass sinograms. Using the normalized sinogram, the noise variance is estimated from the top and bottom rows, and this estimate is used as the true variance in subsequent computations.

In general, the center of mass must also be estimated, perhaps using methods described in [32] or [21], so that the observed projections can be shifted to correspond to an object centered at the origin. In these simulations, however, we assume that the object is already centered at the origin — which is very nearly true for the M I T ellipse. In order to study the effect of correct and incorrect convex hull estimates, the convex hulls used in these studies are fixed and known, although not always correct. Experiments that show the full hierarchical procedure that first estimates the convex hull are described in [32].

5.2 Effect of Smoothing Coefficients

The coefficient γ has the effect of smoothing or blurring the sinogram in the horizontal direction; the coefficient β has a similar smoothing effect in the vertical direction. Fig. 8 shows sinogram MAP estimates resulting from the full-view observations of Fig. 6c, using no support information. Fig. 8a corresponds to $\gamma = 0.05$ and $\beta = 0.01$, Fig. 8b to $\gamma = 0.5$ and $\beta = 0.01$, Fig. 8c to $\gamma = 0.05$ and $\beta = 0.1$, and Fig. 8d to $\gamma = 0.005$ and $\beta = 0.001$. Images reconstructed from these sinograms using CBP are shown in the respective panels of Fig. 9. The reconstruction in Fig. 9a — which used what has empirically shown to be good smoothing coefficients — should be compared to the unprocessed CBP reconstruction of Fig. 6d.

It should be noted from Fig. 9b that excessive smoothing of the sinogram in the horizontal direction results in *circular* blurring of the reconstructed image. Similarly, the haziness of the image in Fig. 9c results from excess smoothing of the sinogram in the vertical direction, which effectively produces a low-pass filtering effect on each projection. There is noticeable improvement in both reconstructions shown in Figs. 9a and 9b over that in Fig. 6d; however, there are important differences. For one, the contrast between the ellipse body and the background is better for the larger smoothing coefficients of Fig. 9a. However, that enhancement also accompanies a decreased definition of the ellipse boundary. The legibility of the internal letters, however, appears to be best in the highest contrast image shown in Fig. 9a.

5.3 Effect of Known Support

Fig. 10 shows the effect of varying κ for known (correct) support. The different values of κ are given by (a) $\kappa = 0.0$, (b) $\kappa = 5.0$, (c) $\kappa = 10.0$, and (d) $\kappa = 10,000$. In each case, the full-view observations of Fig. 6c were used, and $\gamma = 0.05$ and $\beta = 0.01$. The object reconstructions were made using full-view CBP, and should be compared to those of Figs. 6d and 9a,b,c,d.

We see from the set of experiments shown in Fig. 10 that known support sharpens the boundary of the ellipse considerably. However, in the image with the sharpest boundary (Fig. 10d), the letters in the ellipse are not as legible as the images in the other panels — the contrast of the letters does not appear to be as large. This is likely to be due to the mass constraint, which, for κ large, must produce an estimate that has all its mass (for a given projection) between the two support values. But, in addition there is a smoothness requirement which is attempting to reduce abrupt variations within the projections. This may have the overall effect of increasing the magnitude of (normally small) values of line integrals that pass through the internal letters.

5.4 Effect of Incorrect Support

In this set of experiments we examine the effect of using support information which is *incorrect.* Fig. 11 shows results where the support corresponds to an ellipse which has been rotated 90 degrees from the correct orientation. The observed sinogram is that of Fig. 6c, and the algorithm used the smoothing coefficients $\gamma = 0.05$ and $\beta = 0.01$. The different reconstructions in Fig. 11 correspond to setting κ to (a) $\kappa = 0.0$, (b) $\kappa = 5.0$, (c) $\kappa = 10.0$, and (d) $\kappa = 10,000$.

This set of experiments shows that as κ grows larger, the image values outside the assumed region of support grow smaller. Eventually, this effect overwhelms the evidence of the observations and virtually obliterates the parts of the true ellipse that lie outside of the incorrect support. But the effect of the mass constraint and the smoothing coefficients also affect the appearance of the final image. Since each projection has mass m, when the support width is incorrectly narrowed, and κ is too large, then the sinogram values must be very large within the region of support *just to accommodate the required mass*, and the values will typically be very much larger than the observations. As mentioned previously, this will have the effect of reducing the contrast of the inner details of these projections, and the effect on the image is to eliminate contrast within even the intersection of the correct support and the incorrect support. On those projections that have support values that are much too wide, it is the smoothing terms that dominate. In order to lower the overall energy of the sinogram (that is, the energy term in the Markov random field), the vertical pair-potentials or equivalently, the vertical derivatives should be small. Therefore, these projections tend to become as smooth as possible over the prescribed support and contribute to the image a "shadow" ellipse which corresponds to the incorrect support.

Fig. 12 shows a sequence of reconstructions that have kept κ to the constant 5.0, but vary the orientation of the assumed object support. In these reconstructions, we have used the support of an ellipse that has the same size and eccentricity at the true object support, but has been rotated in the counter-clockwise direction by (a) 0.0 degrees, (b) 15.0 degrees, (c) 30.0 degrees, and (d) 45.0 degrees.

This set of experiments shows that a modest choice of κ, together with a less severe support error will produce an image that retains many of the details of the true image with only a small "shadow" due to the incorrect support. However, it is clear that an incorrect support estimate can produce results much worse than having not introduced any support information whatsoever (compare these results to that of Fig. 11a).

In Fig. 13 we show a sequence of reconstructions that have used $\kappa = 5.0$, but with support which is the incorrect size. Figs. 13a and 13b show two cases where the support is too small, and Figs. 13c and 13d show two cases where the support is too large. Overall, the size of the support increases from Fig. 13a to Fig. 13d. The reconstruction using the correct support and $\kappa = 5.0$ may be seen in Fig. 12a.

We may conclude from this set of experiments that is is preferable to err on the side of using a support estimate that is too large than too small, in general. Although the boundaries are not as sharp when the support is too large, the loss of contrast in the interior and the effect of double-boundaries for small support is much more undesirable.

5.5 Sparse-Angle Studies

Fig. 14 shows the results of several sparse-angle experiments. The (a) and (b) images correspond to the 15-view and 10-view 10dB cases respectively, where $\gamma = 0.05$ and $\beta = 0.01$ and the support is known and $\kappa = 10,000$. The (c) and (d) images correspond to the 15-view and 10-view cases, respectively, with the same smoothing coefficients, but with $\kappa = 0.0$ — i.e., no known support information is used.

This experiment demonstrates nicely the potential of the algorithm. In either sparse-angle case, the contrast of the image is improved over those in Fig. 7 dramatically. And while the boundary is quite sharp as expected in the case of $\kappa = 10,000$, it is quite clear what the shape of the object is in case of $\kappa = 0.0$. The loss of contrast in the interior of the ellipse when $\kappa = 10,000$ remains evident in these experiments, however.

5.6 Limited-Angle Studies

Fig. 15 shows the results of several limited-angle studies. The (a) and (b) images are reconstructions obtained with known support (with $\kappa = 10,000$) from the two limited-angle cases. The experiment resulting in panel (a) uses the first 40 (left-most) projections, whereas panel (b) uses the last 40 (right-most) projections. Panels (c) and (d) correspond to the same observations as in (a) and (b), respectively, but in these cases no support information was used. As in the sparse-angle studies, the smoothing coefficients for all four studies were $\gamma = 0.05$ and $\beta = 0.01$.

These limited-angle studies show behavior which is similar to the sparse-angle studies. The boundary of the ellipse is quite sharp, as expected, in the case of $\kappa = 10,000$, and there is an accompanying loss of contrast in the interior. The images generated using $\kappa = 0.0$

have different problems, however. In particular, the image in Fig. 15c shows good contrast in the letters in the interior but is unable to provide any boundary definition on the long sides of the ellipse. This is because the leftmost 40 projections which are observed view the ellipse from the *broadside*, and as such do not contain information about the narrow ellipse dimension. The image in Fig. 15d suffers from the opposite problem. There is a loss of definition of the letters in the interior because many of the projections that would normally be obtained from the broadside of the ellipse are missing. It is in the first case that support knowledge can aid tremendously; unfortunately, when projections from the broadside of the ellipse are missing, there is little that our method can do to provide any additional clarity of the interior detail.

6 DISCUSSION

It is generally acknowledged in the computed tomography literature that in the case of noisy and limited-angle or sparse-angle data, prior knowledge is essential in order to obtain good reconstructions. We have focused on three types of prior knowledge:

- Line integrals close in either angle or lateral displacement tend to be similar in value,
- Radon transform functions are constrained to a certain functional subspace, and
- Knowledge of the convex hull of the object is equivalent to knowledge of a particular region of support of the object's Radon transform.

In Section 3, we developed a Markov random field (MRF) model of sinograms that contains prior information about the mass and center of mass of the unknown sinogram, the convex support of the object, and the expected similarity of line integrals which are close in either angle or lateral displacement. The primal-dual MAP estimation algorithm developed in Section 4 is based on a variational formulation that leads to an efficient solution of the original Markov random field MAP estimation problem. Even with the necessity of keeping and updating Lagrange multipliers, this method is fast and memory efficient, and is parallelizable in both the primal phase and the dual phase.

The simulations presented in Section 5 show the range of results that may be obtained using this approach. The improvement over the unprocessed CBP reconstructions is quite dramatic in all cases where the support was correct or nearly correct. In particular, the boundary of the ellipse is made much sharper, and the letters within the ellipse can be made out in all of the processed cases, and in none of the unprocessed cases.

In most cases the convex support of the object is not known *a priori*. In other research, we have reported several methods to estimate the convex hull of objects from the available data [43],[44],[32]. These methods require two steps: 1) estimation of the support values from observed projections and 2) estimation of a complete set of feasible support values for all projections (including the ones corresponding to missing observations). The feasible support values uniquely identify a convex polyhedron, which serves as our estimate of the convex support of the object. In the second step, different types of prior geometric knowledge about the shape of the object may be used to force estimates to be circular, elliptic, or have smooth boundaries. For example, knowledge that the MIT ellipse is an ellipse leads to nearly perfect support estimation from the noisy, sparse-angle, and limited-angle cases used in Section 5, even though the size, eccentricity, and orientation of the ellipse is not known *a priori*. Estimating the convex support, mass, and center of mass of the object is viewed as a part of a hierarchical reconstruction algorithm in [32]. With these steps in place, the only user inputs that are required are the values of the smoothing coefficients β and γ.

In this paper only two consistency constraints were imposed: mass and center of mass. A method that incorporates a much larger number of consistency constraints and that requires no prior geometric knowledge related to these constraints — e.g. mass and center of mass — is described in [32]. The two methods contrast in the following way. The approach described in this paper characterizes, through the mass and center of mass, a functional subspace in which the desired sinogram must lie, and forces this to happen using Lagrange multipliers. The alternate approach described in [32] characterizes the subspace orthogonal to the desired sinogram, and again uses Lagrange multipliers to achieve this goal. The alternate method is also a generic primal-dual optimization algorithm, however, the Lagrange multipliers are scalars rather than functions. The primal stage is almost identical to that given herein, but the dual stage generally requires more computation and has less potential for parallelism.

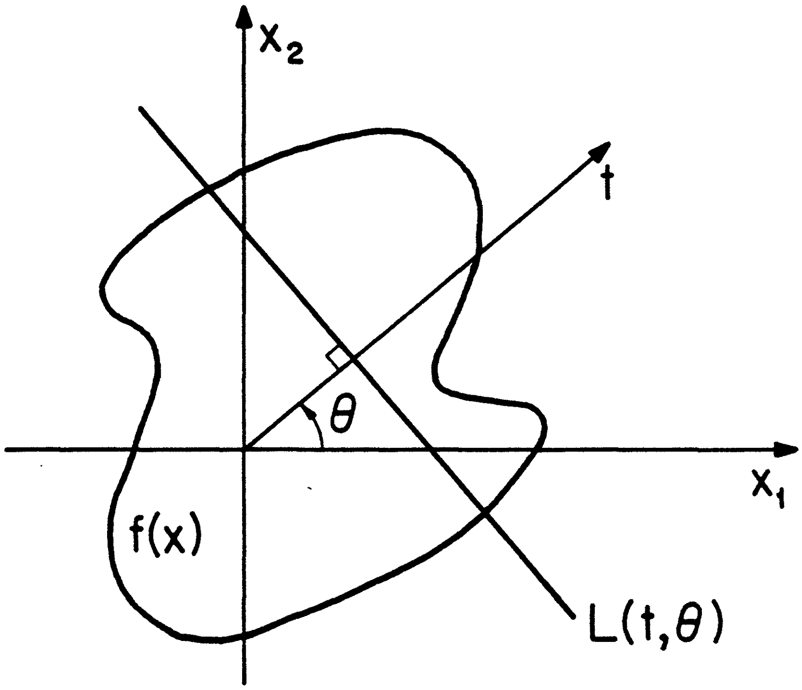


Figure 1: The geometry of the 2-D Radon transform.

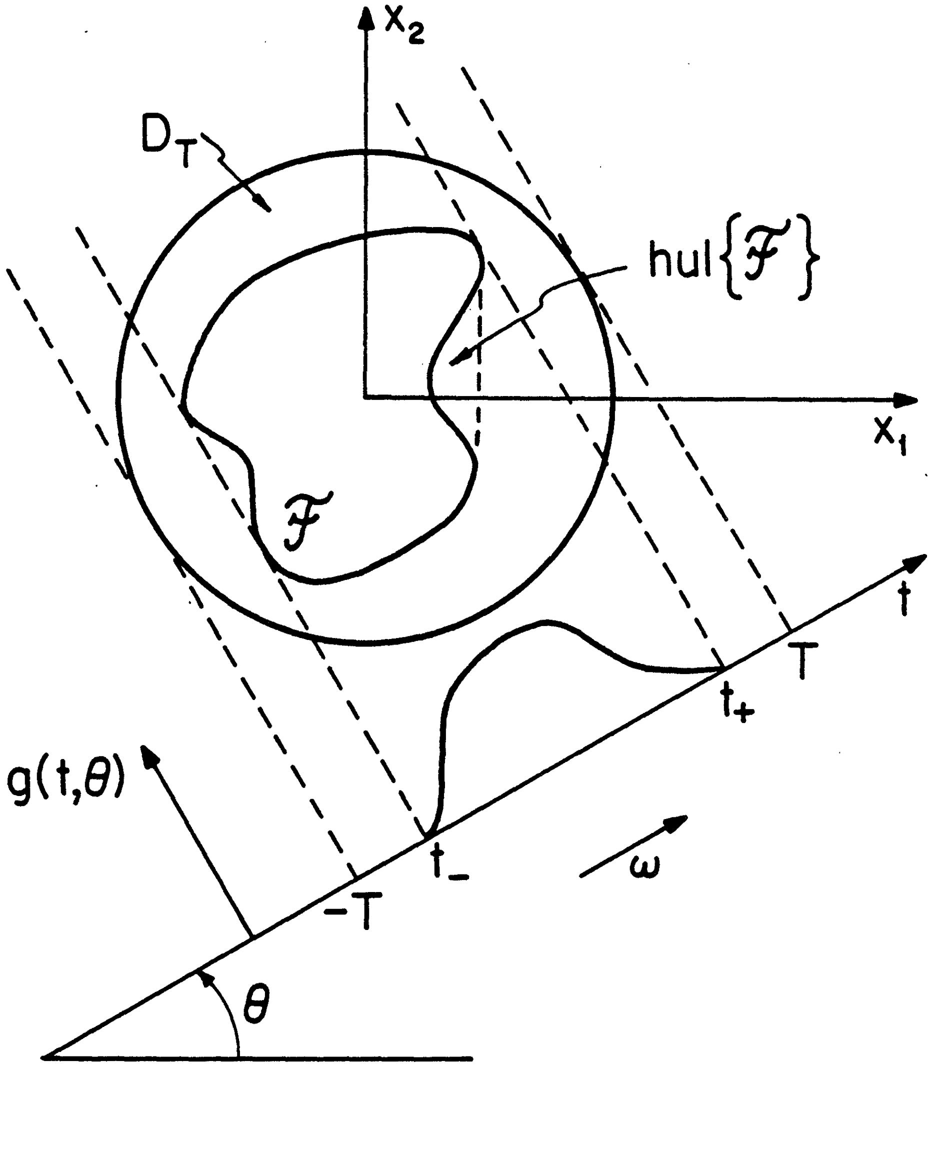


Figure 2: The convex support of an object and of a projection.

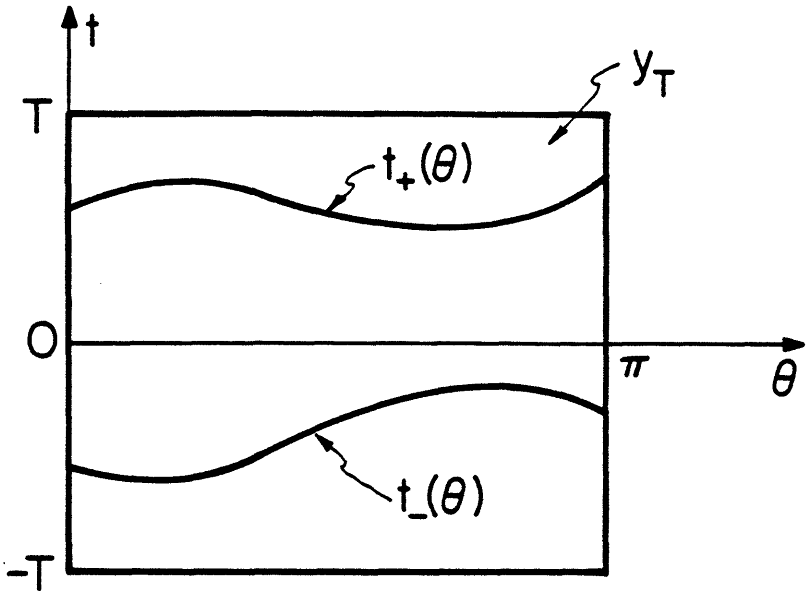
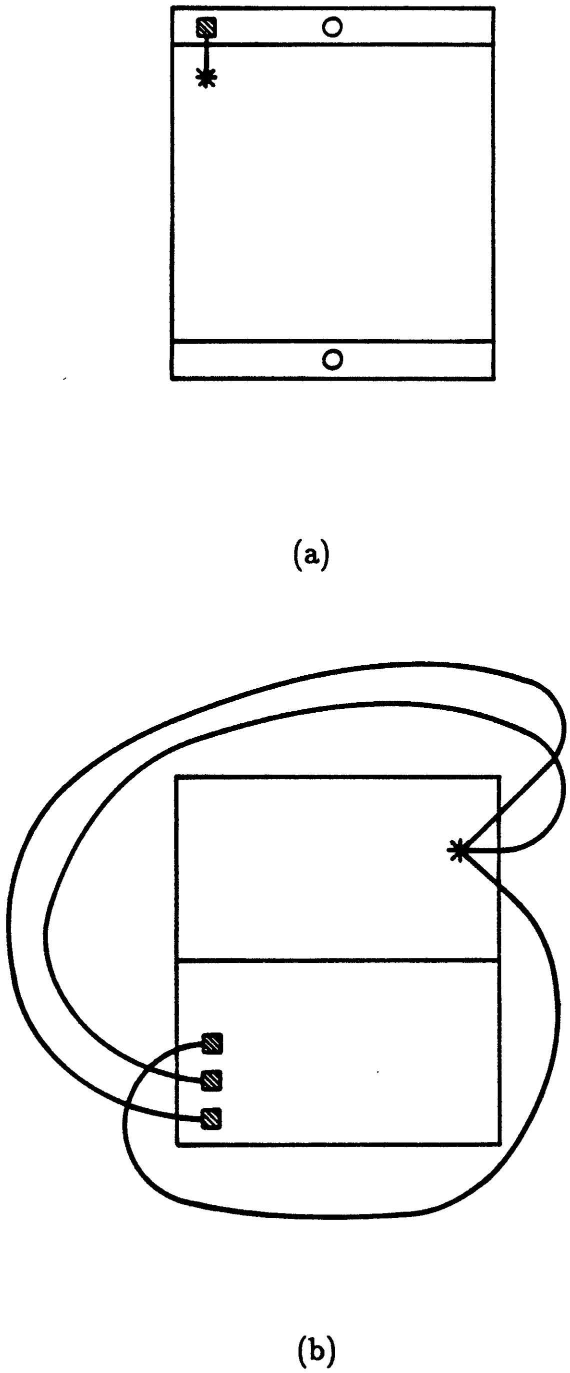


Figure 3: The support of a Radon transform.



(a)

(b)

Figure 4: (a) The vertical and (b) horizontal boundaries of the sinogram MRF.

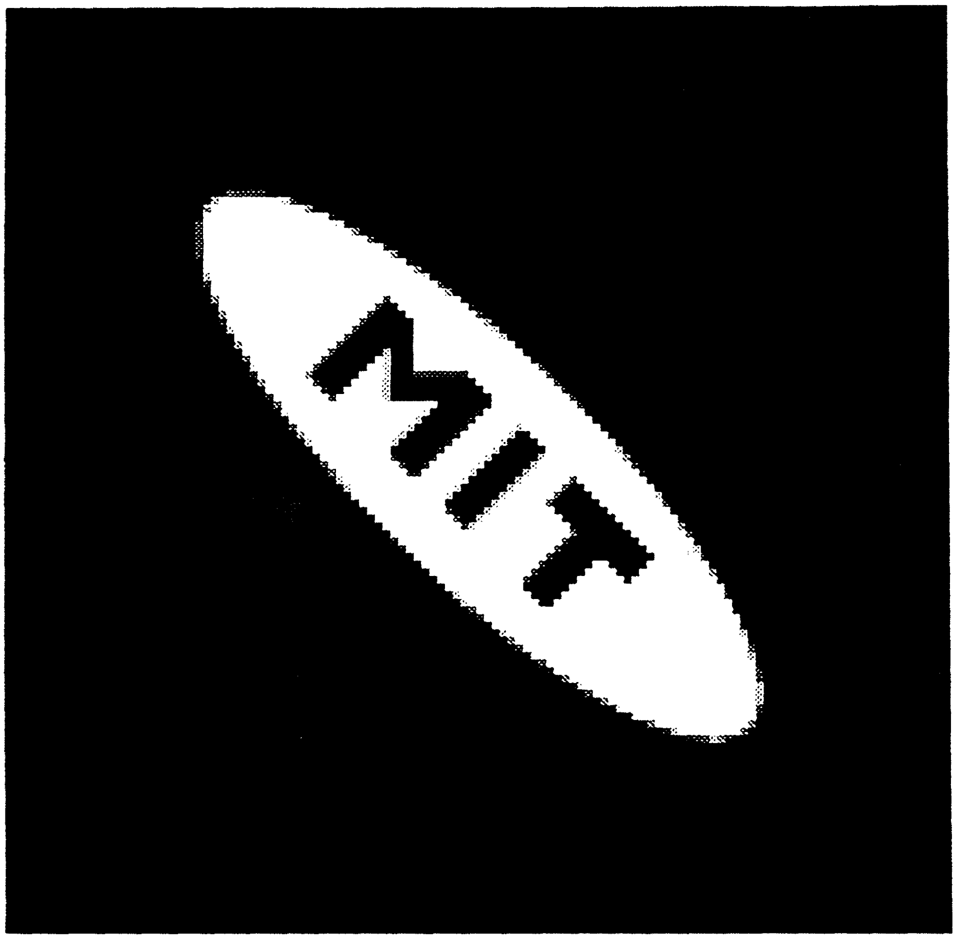


Figure 5: The MIT ellipse.

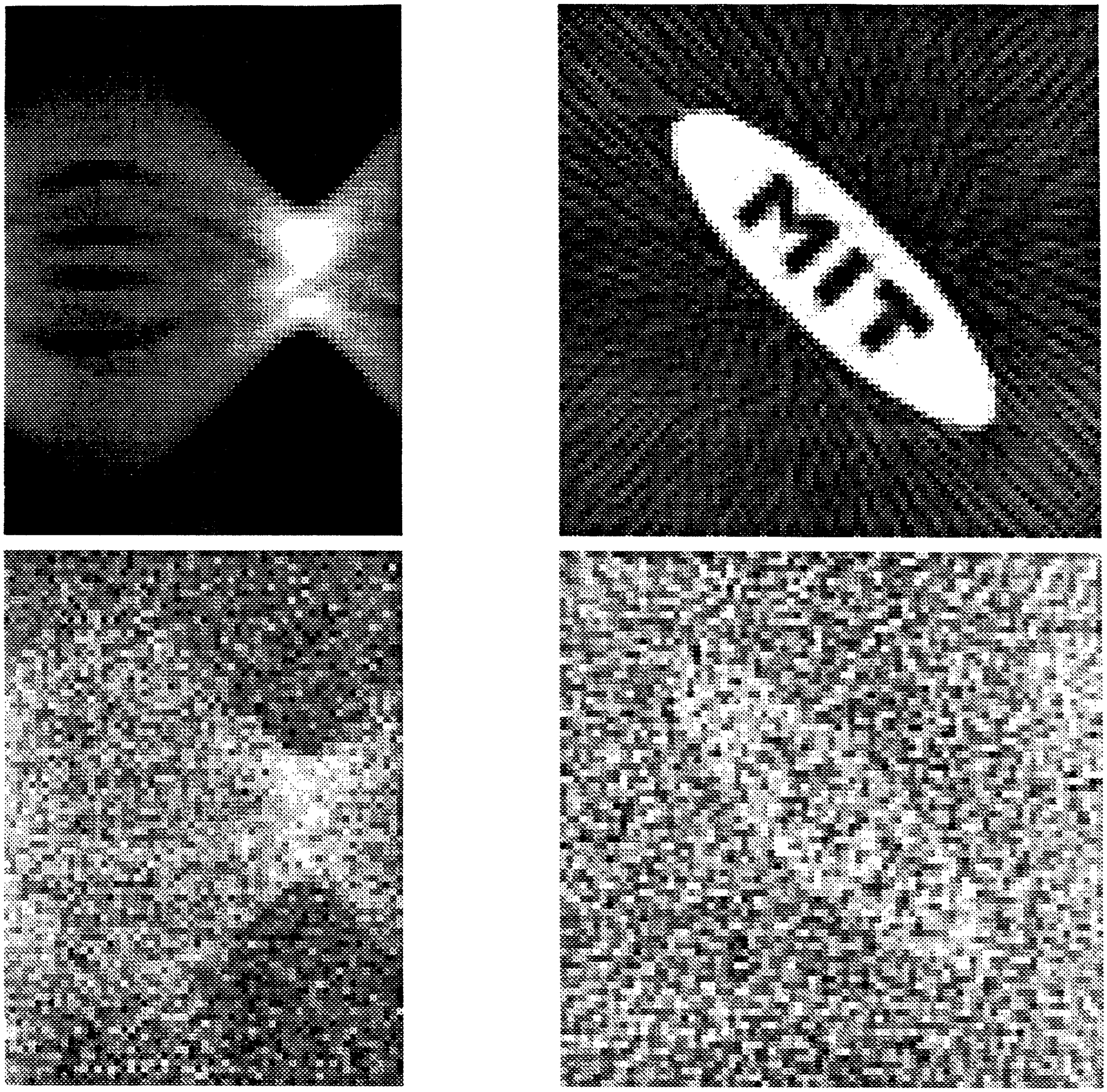


Figure 6: (a) A noise-free sinogram, (b) and its reconstruction. (c) A noisy sinogram (SNR=3.0dB), (c) and its reconstruction.

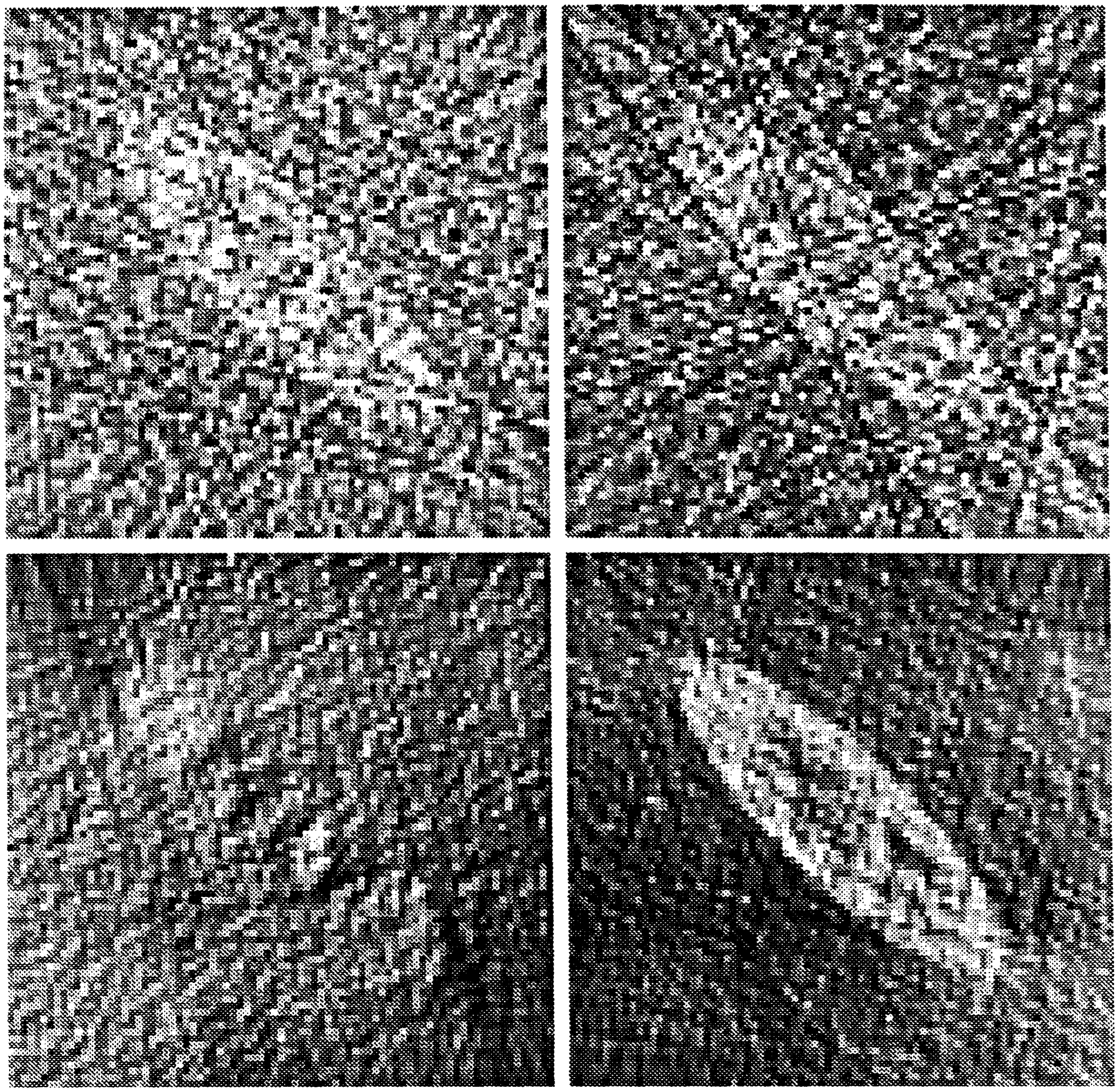


Figure 7: Reconstructions from a noisy sinogram (SNR=10.0dB) from (a) 15 sparse views, (b) 10 sparse views, (c) left-most 40 views, and (d) right-most 40 views.

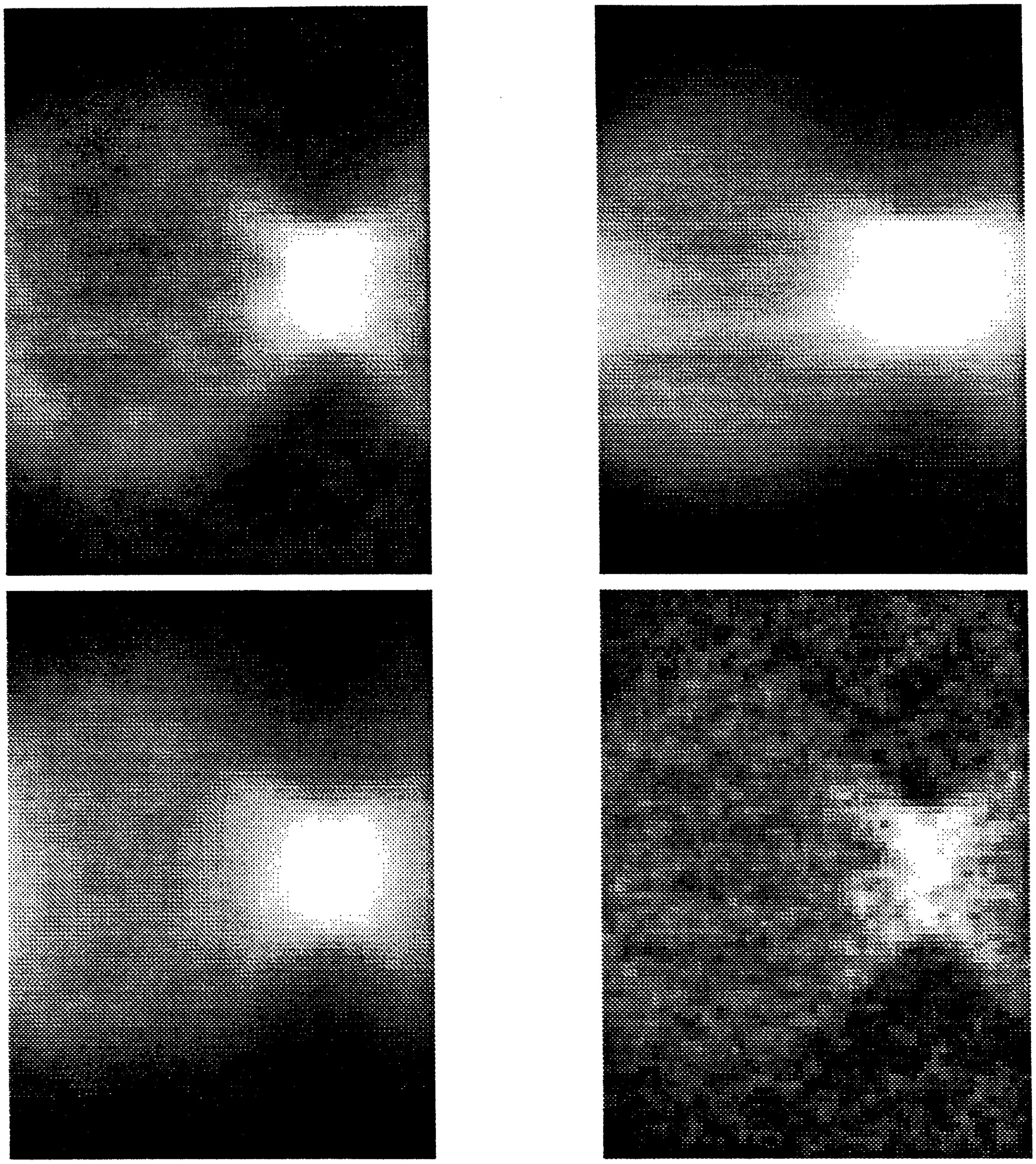


Figure 8: Estimates produced by the primal-dual algorithm with (a) $\gamma = 0.05$ and $\beta = 0.01$, (b) $\gamma = 0.5$ and $\beta = 0.01$, (c) $\gamma = 0.05$ and $\beta = 0.1$, and (d) $\gamma = 0.005$ and $\beta = 0.001$.

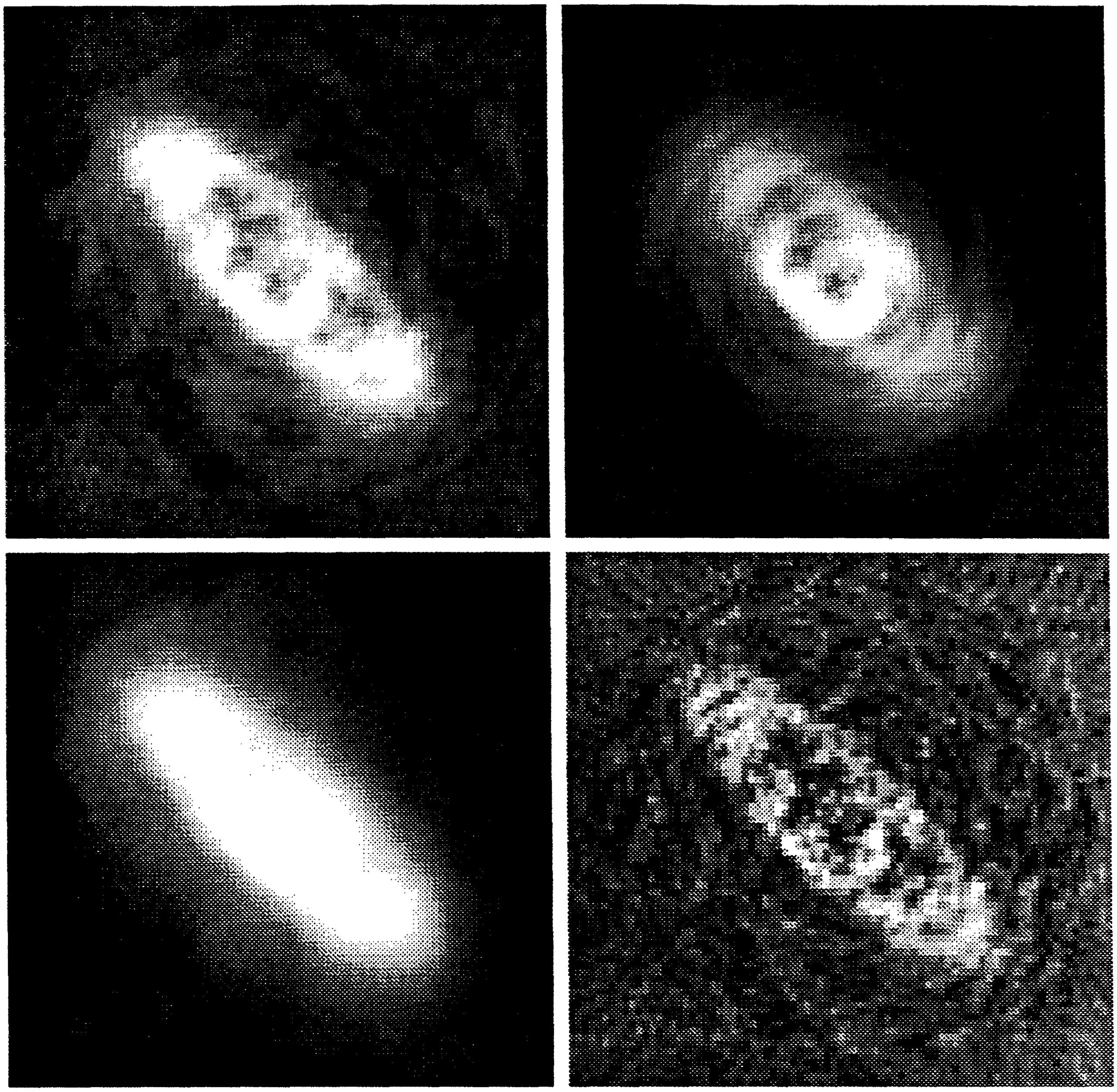


Figure 9: CBP reconstructions from Fig. 8.

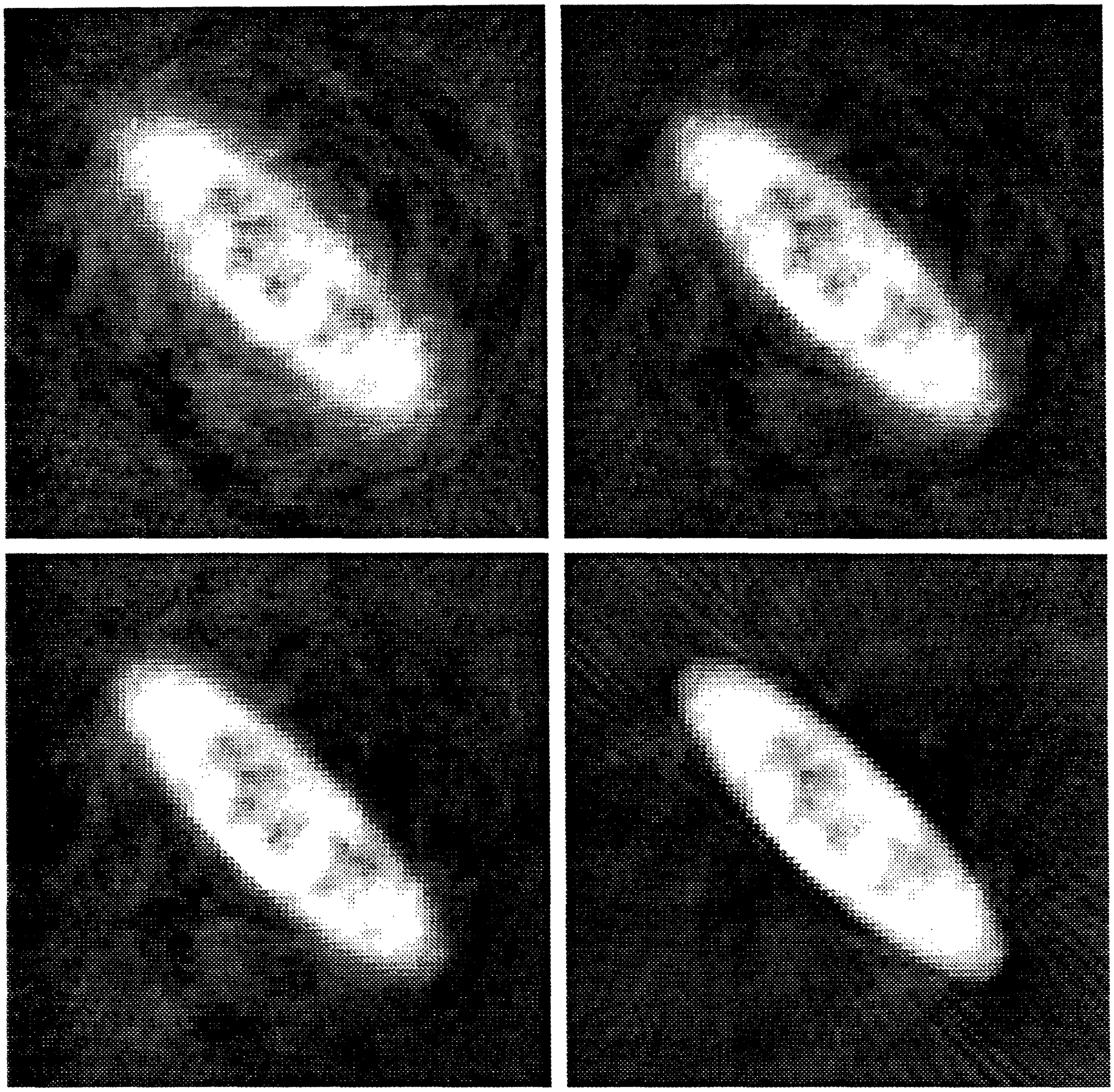


Figure 10: Effect of known support for (a) $\kappa = 0.0$, (b) $\kappa = 5.0$, (c) $\kappa = 10.0$, and (d) $\kappa = 10,000.0$.

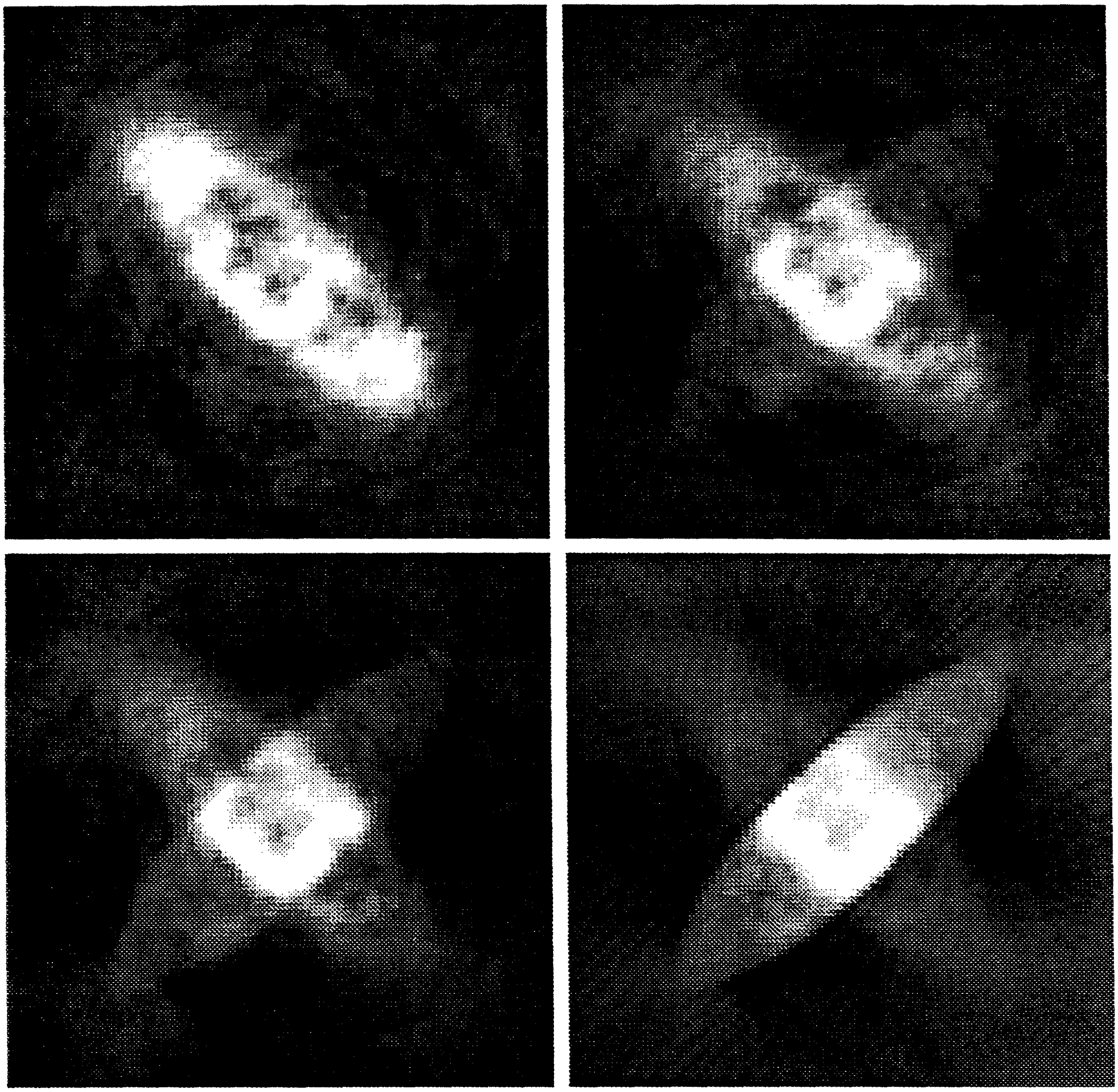


Figure 11: Effect of incorrect support for (a) $\kappa = 0.0$, (b) $\kappa = 5.0$, (c) $\kappa = 10.0$, and (d) $\kappa = 10,000.0$.

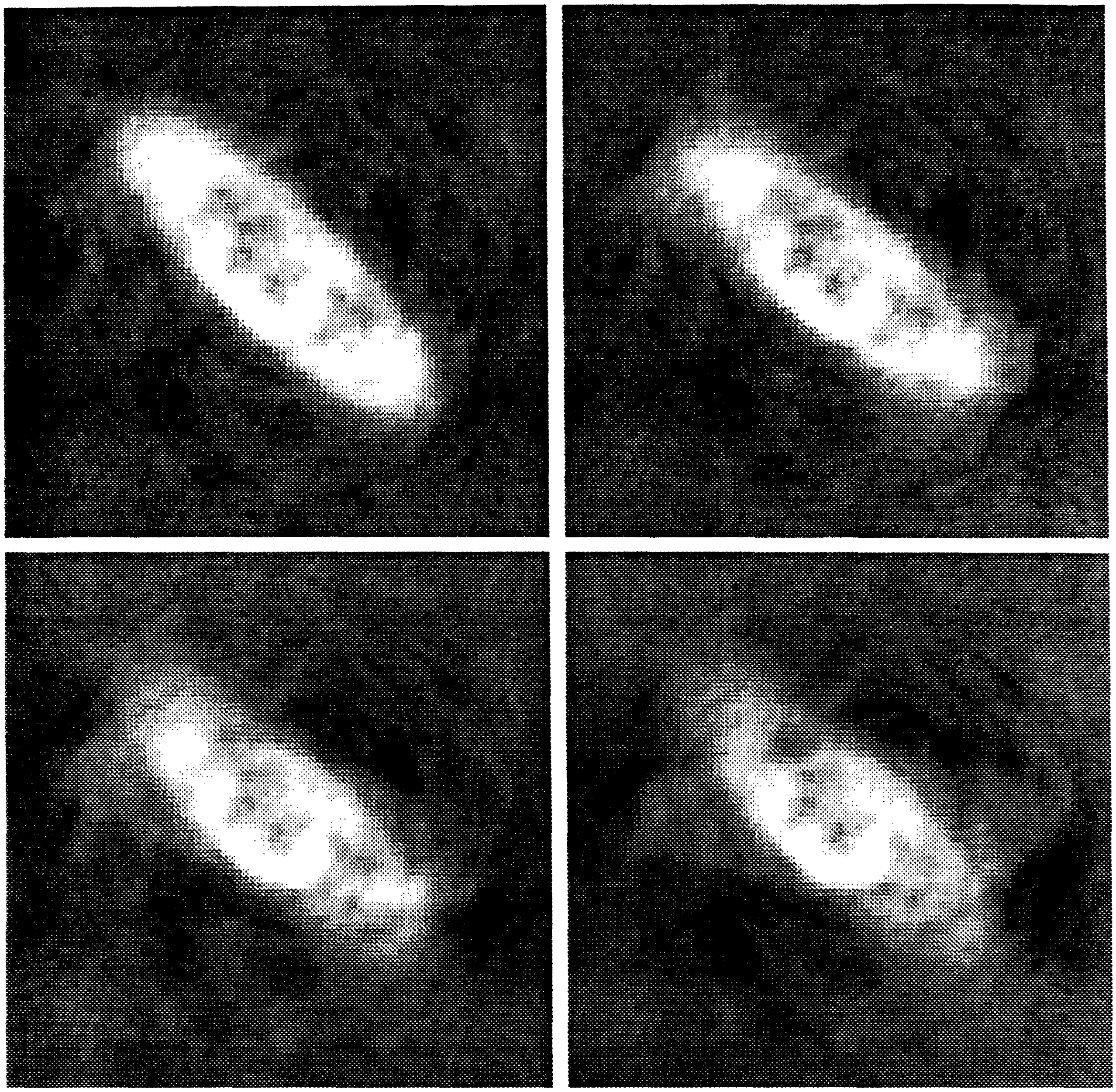


Figure 12: Effect of using an incorrect support which is rotated counter-clockwise by (a) 0.0 degrees, (b) 15.0 degrees, (c) 30.0 degrees, and (d) 45.0 degrees.

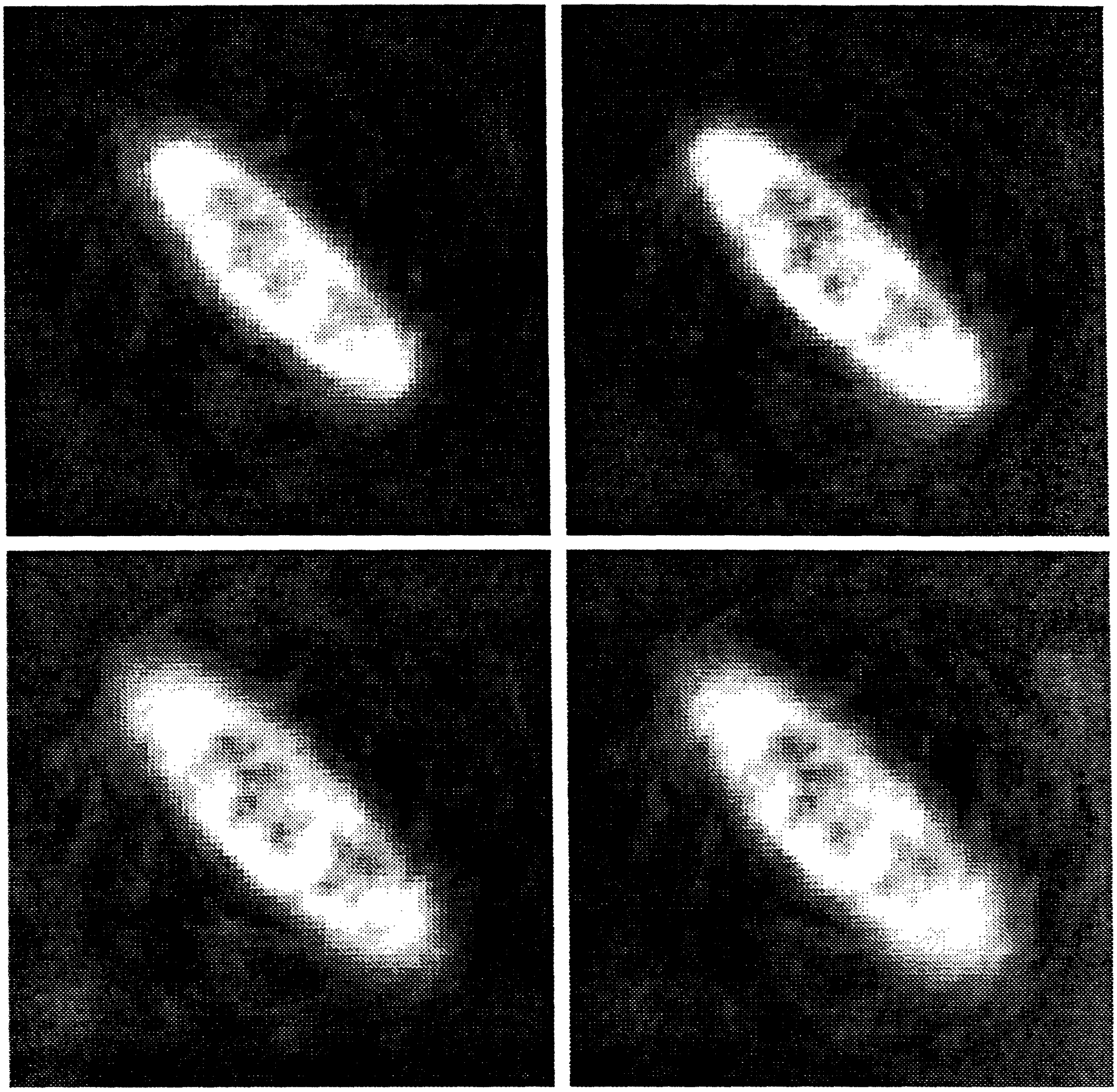


Figure 13: Effect of using an incorrect support which is too small in (a) and (b) and too large in (c) and (d).

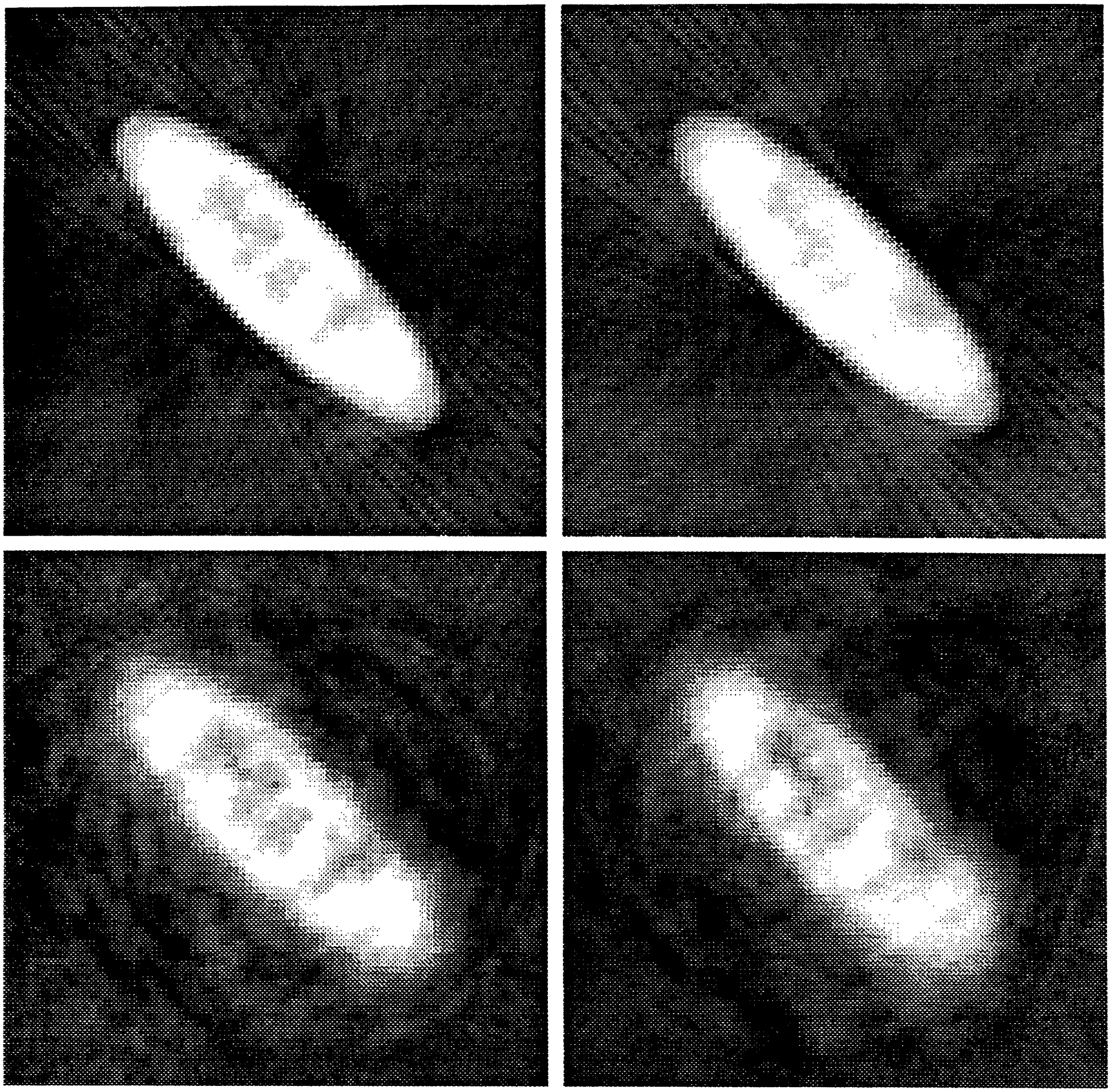


Figure 14: Sparse-angle studies with $\gamma = 0.05$ and $\beta = 0.01$. (a) 15 observed projections and known support. (b) 10 observed projections and known support. (c) 15 observed projections and no support. (d) 10 observed projections and no support.

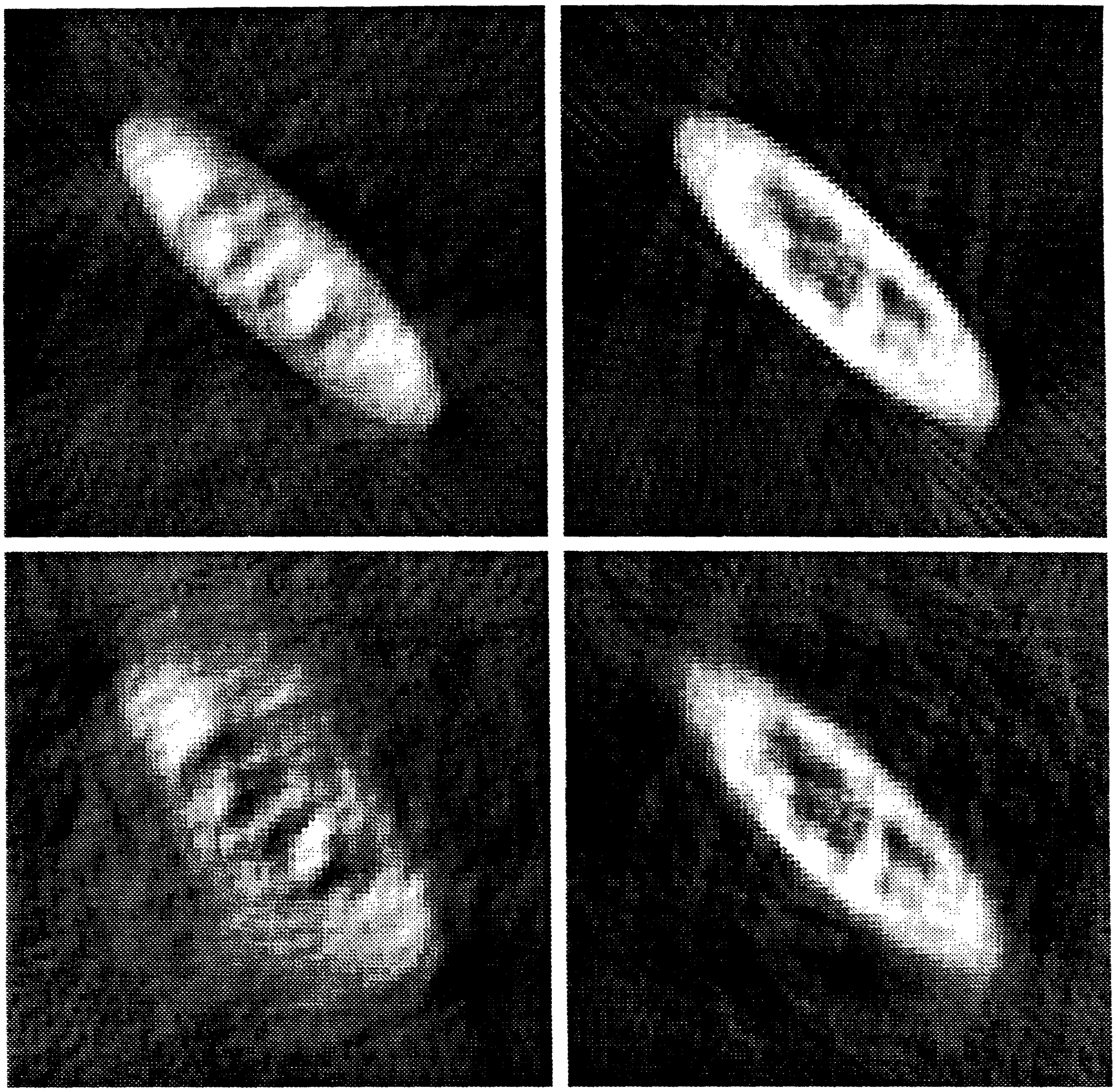


Figure 15: Limited-angle studies with $\gamma = 0.05$ and $\beta = 0.01$. (a) Left 40 projections and known support. (b) Right 40 projections and known support. (c) Left 40 projections and no support. (d) Right 40 projections and no support.

References

[1] R. H. T. Bates, K. L. Garden, and T. M. Peters, "Overview of computerized tomography with emphasis on future developments," *Proc. IEEE*, vol. 71, no. 3, pp. 356–297, 1983.

[2] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging.* New York: IEEE Press, 1988.

[3] L. Axel, P. H. Arger, and R. A. Zimmerman, "Applications of computerized tomography to diagnostic radiology," *Proc. IEEE*, vol. 71, no. 3, pp. 293–297, 1983.

[4] S. R. Deans, *The Radon Transform and Some of Its Applications.* New York: John Wiley and Sons, 1983.

[5] A. Macovski, "Physical problems of computerized tomography," *Proc. IEEE*, vol. 71, no. 3, pp. 373–378, 1983.

[6] M. E. Davison and F. A. Grunbaum, "Tomographic reconstructions with arbitrary directions," *Comm. Pure Appl. Math.*, vol. 34, pp. 77–119, 1979.

[7] T. Inoye, "Image reconstruction with limited angle projection data," *IEEE Trans. Nucl. Sci.*, vol. NS-26, no. 2, pp. 2666–2669, 1979.

[8] J. A. Reeds and L. A. Shepp, "Limited angle reconstruction in tomography via squashing," *IEEE Trans. on Medical Imaging*, vol. MI-6, pp. 89–97, June 1987.

[9] A. K. Louis, "Picture reconstruction from projections in restricted range," *Math. Meth. in the Appl. Sci.*, vol. 2, pp. 209–220, 1980.

[10] S. L. Wood, *A system theoretic approach to image reconstruction.* PhD thesis, Stanford University, May 1978.

[11] M. H. Buonocore, W. R. Brody, and A. Macovski, "Fast minimum variance estimator for limited angle CT image reconstruction," *Medical Physics*, Sept./Oct. 1981.

[12] B. R. Frieden and C. K. Zoltani, "Maximum bounded entropy: application to tomographic reconstruction," Tech. Rep., US Army Ballistic Research Laboratory, 1985. Technical Report BRL-TR-2650.

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

[14] M. H. Buonocore, *Fast Minimum Variance Estimators for Limited Angle Computed Tomography Image Reconstruction.* PhD thesis, Stanford University, 1981.

[15] D. C. Youla and H. Webb, "Image restoration by the method of convex projections: part 1 — theory," *IEEE Trans. Med. Imaging*, vol. MI-1, pp. 81–94, 1982.

[16] M. I. Sezan and H. Stark, "Tomographic image reconstruction from incomplete view data by convex projections and direct Fourier inversion," *IEEE Trans. Med. Imag.*, vol. MI-3, no. 2, pp. 91–98, 1984.

[17] M. I. Sezan and H. Stark, "Image restoration by convex projections in the presence of noise," *Applied Optics*, vol. 22, no. 18, pp. 2781–2789, 1983.

[18] J. H. Park, K. Y. Kwak, and S. B. Park, "Iteractive reconstruction-reprojection in projection space," *Proc. IEEE*, vol. 73, pp. 1140–1141, June 1985.

[19] J. H. Kim, K. Y. Kwak, S. B. Park, and Z. H. Cho, "Projection space iteration reconstruction-reprojection," *IEEE Trans. Med. Imag.*, vol. MI-4, September 1985.

[20] D. J. Rossi, *Reconstruction from projections based on detection and estimation of objects.* PhD thesis, Massachusetts Institute of Technology, 1982.

[21] D. J. Rossi and A. S. Willsky, "Reconstruction from projections based on detection and estimation of objects–parts I and II: performance analysis and robustness analysis," *IEEE Trans. ASSP*, vol. ASSP-32, no. 4, pp. 886–906, 1984.

[22] Y. Bresler and A. Macovski, "3-d reconstruction from projections based on dynamic object models," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, March 1984.

[23] Y. Bresler and A. Macovski, "Estimation of 3-d shape of blood vessels from x-ray images," in *Proc. IEEE Comp. Soc. Int. Symp. on Medical Images and Icons*, July 1984.

[24] Y. Bresler and A. Macovski, "A hierarchical Bayesian approach to reconstruction from projections of a multiple object 3-d scene," in *Proc. 7th International Conference on Pattern Recognition*, August 1984.

[25] K. M. Hanson, "Tomographic reconstruction of axially symmetric objects from a single radiograph," 1984. SPIE, vol. 491, High Speed Photography (Strasbourg).

[26] M. Soumekh, "Binary image reconstruction from four projections," in *Proceedings of the 1988 Int'l Conf. Acoust. Speech. Sig. Proc.*, pp. 1280–1283, April 1988.

[27] B. K. P. Horn, *Robot Vision.* Cambridge, MA: MIT Press, 1986.

[28] P. C. Fishburn, J. A. Reeds, and L. A. Shepp, "Sets uniquely determined by projections," 1986. Preprint.

[29] P. D. Lax and R. S. Phillips, "The Paley-Wiener theorem for the Radon transform," *Comm. Pure and Appl. Math.*, vol. 23, pp. 409–424, 1970.

[30] S. Helgason, *The Radon Transform.* Boston, MA: Birkhauser, 1980.

[31] D. Ludwig, "The Radon transform on Euclidean space," *Comm. Pure Appl. Math.*, vol. 19, pp. 49–81, 1966.

[32] J. L. Prince, *Geometric Model-Based Estimation From Projections.* PhD thesis, Massachusetts Institute of Technology, 1988.

[33] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and Bayesian restoration of images," *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. PAMI-6, pp. 721–741, Nov. 1984.

[34] G. Cross and A. Jain, "Markov random field texture models," *IEEE Trans. Pat. Analysis and Mach. Int.*, vol. PAMI-5, January 1983.

[35] R. Kinderman and J. L. Snell, *Markov Random Fields and Their Applications.* Providence: American Mathematical Society, 1980.

[36] S. Geman and D. E. McClure, "Bayesian image analysis: and application to single photon emission tomography," Tech. Rep., Brown University, 1985. Preprint to appear in 1985 Proc. Amer. Stat. Assoc. Statistical Computing.

[37] D. G. Luenberger, *Linear and Nonlinear Programming.* Reading, MA: Addison-Wesley, second ed., 1984.

[38] F. B. Hildebrand, *Methods of Applied Mathematics.* Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1965.

[39] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods.* New York: Academic Press, 1982.

[40] C. J. Kuo, B. C. Levy, and B. R. Musicus, "A local relaxation method for solving elliptic PDE's on mesh-connected arrays," *SIAM J. Sci. Stat. Comput.*, vol. 8, no. 4, pp. 550–573, 1987.

[41] R. S. Varga, *Matrix Iterative Analysis.* Englewood Cliffs, N.J.: Prentice-Hall, 1962.

[42] G. T. Herman, *Image Reconstruction from Projections.* New York: Academic Press, 1980.

[43] J. L. Prince and A. S. Willsky, "Estimation algorithms for reconstructing a convex set given noisy measurements of its support lines," Tech. Rep. LIDS-P-1638, M.I.T. Laboratory for Information and Decision Systems, January 1987.

[44] J. L. Prince and A. S. Willsky, "Reconstructing convex sets from support line measurements," Tech. Rep. LIDS-P-1704, M.I.T. Laboratory for Information and Decision Systems, September 1987. Submitted to IEEE PAMI.