Information-Theoretic Image Formation

Joseph A. O'Sullivan, *Senior Member, IEEE,* Richard E. Blahut, *Fellow, IEEE,* and Donald L. Snyder, *Fellow, IEEE*

(Invited Paper)

***Abstract*— The emergent role of information theory in image formation is surveyed. Unlike the subject of information-theoretic communication theory, information-theoretic imaging is far from a mature subject. The possible role of information theory in problems of image formation is to provide a rigorous framework for defining the imaging problem, for defining measures of optimality used to form estimates of images, for addressing issues associated with the development of algorithms based on these optimality criteria, and for quantifying the quality of the approximations. The definition of the imaging problem consists of an appropriate model for the data and an appropriate model for the reproduction space, which is the space within which image estimates take values. Each problem statement has an associated optimality criterion that measures the overall quality of an estimate. The optimality criteria include maximizing the likelihood function and minimizing mean squared error for stochastic problems, and minimizing squared error and discrimination for deterministic problems. The development of algorithms is closely tied to the definition of the imaging problem and the associated optimality criterion. Algorithms with a strong information-theoretic motivation are obtained by the method of expectation maximization. Related alternating minimization algorithms are discussed. In quantifying the quality of approximations, global and local measures are discussed. Global measures include the (mean) squared error and discrimination between an estimate and the truth, and probability of error for recognition or hypothesis testing problems. Local measures include Fisher information.**

***Index Terms*—Image analysis, image formation, image processing, image reconstruction, image restoration, imaging, inverse problems, maximum-likelihood estimation, pattern recognition.**

I. Introduction

IMAGE formation is the process of computing (or refining) an image both from raw sensor data that is related to that image and from prior information about that image. Information about the image is contained in the raw sensor data, and the task of image formation is to extract this information so as to compute the image. Thus it appears that information-theoretic notions can play an important role in this process. We will survey the emergent role that information theory now plays in the subject of image formation or may play in the future. This role could be to provide a rigorous framework for defining the imaging problem, for defining measures of optimality that can be used to judge estimates of images, for addressing issues associated with the development of algorithms based on these optimality criteria, and for quantifying the statistical quality of the approximations.

To this end, the domain of information theory may be divided into two parts: communication and observation. The problems of communication have been very successfully treated by information theory, in part because Shannon had the foresight to overlay the subject of communication with a clear partitioning into *sources*, *channels*, *encoders*, and *decoders*. Although Shannon's formalization seems quite obvious in our time, it was not so obvious half a century ago. In contrast, the problems of observation, including imaging, have been slower to yield to the methods of information theory, partly because the image formation problems are harder, and perhaps partly because a formal framework for the subject is still emerging. Even the terms *source*, *sensor*, and *image* can be slippery; our understanding of these terms is closely tied to and colored by our view of a particular physical problem. It is not yet common practice to study problems of image formation in terms of an abstract formalization that is not connected to a specific physical problem.

One may take the natural position that an image formation problem consists of a source to be imaged, a sensor that collects data about the source, and an algorithm that estimates the image from the data. Thus it seems that image formation closely corresponds to our commonplace notion of photography. However, upon closer examination one can find difficulties with this simple view. A physical scene has a richness and complexity well beyond what we may wish to model or can model. In some problems, the sensor data may contain very little information but the prior knowledge may be considerable. Then one uses the sensor data to supplement the prior model to produce the image. This is called *image enhancement* in some contexts and *model-based imaging* or *physics-based imaging* in others. In an extreme case of model-based imaging, the imaging task may well degenerate into the estimation of several parameters, or even a simple *yes* or *no* decision, meaning only that a previously designated object or target appears somewhere in the scene.

Similarly, the meaning of the term "sensor" can be hard to define. How much of the processing is part of the sensor and how much not? The placement or motion of a physical device does affect the data collected by the device. Is this placement or motion to be viewed as part of the sensor or as part of an encoder that prepares data for a sensor? Should one introduce the notion of an encoder into an imaging problem?

Manuscript received April 15, 1998; revised May 22, 1998. This work was supported in part by the Army Research Office under Grant ARO DAAH049510494.

J. A. O'Sullivan and D. L. Snyder are with the Department of Electrical Engineering, Washington University, St. Louis, MO 63130 USA.

R. E. Blahut is with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA.

Publisher Item Identifier S 0018-9448(98)06085-4.

Because we lack the greatness of Shannon, we have difficulty moving from these abstract questions to an abstract model. Instead, we tend to answer such questions only in a narrower context by relating back to specific physical situations. Nevertheless, we shall press forward in this paper to describe the emergent role of information theory in the imaging problem. Because imaging sensors of the future will provide massive amounts of data, and computers of the future will be able to process massive amounts of data, the theory that we seek is needed to guide the development of these image formation systems of the future. This implies that there is a need for a formal information-theoretic framework that can offer advice about how to process massive data sets so as to extract all of the information relevant to the task of image formation.

Many kinds of sensors passively collect data from an environment already rich with many kinds of signals, and these data sets may contain information about an object of interest. In many cases, this information is very deeply buried in the data. Powerful methods are necessary to examine the data by applying the various techniques of filtering, correlation, inference, and so forth. Seismic and acoustic systems may consist of large arrays of many small devices. Optical sensors and infrared sensors now contain very large detector arrays, such as charge-coupled device (CCD) arrays, in which individual pixels can be addressed and archived. Indeed, in a low-light environment, the time of occurrence of individual photon conversions can be reported one-by-one by each sensor pixel. Optical sensors used in imaging spectropolarimetry produce enormous quantities of data. Electromagnetic sensors in the microwave, ultra-high frequency (UHF), or very-high frequency (VHF) band can report massive amounts of data at every antenna. Lidars can now remotely probe the absorption spectrum of trace gases in the atmosphere. Even passive electromagnetic sensors at lower frequencies can report a considerable amount of useful data.

To enrich the collection of data, many sensors actively probe the environment with transmitted signals, such as radar, seismic, or lidar signals. This illumination may be necessary in order to create the necessary data-bearing reflections. It should also be noted, however, that in many cases, active probes are designed not just to increase the amount of illumination falling upon the scene, but rather to put that illumination into a form so that the received sensor data are in a convenient form. Although the environment may already contain many sources of energy that provide illumination and scattered reflections, this energy is not usually organized into waveforms that are easy to process by image formation algorithms.

To extract information from the collected data, models must be developed for the objects of interest, for the environment, and for the sensor. A system that observes a remote area must process signals that propagate long distances, and possibly through complex environments. A system that extracts information from weakly radiating objects will usually need large amounts of data and long integration times. A system that uses imaging radar for the detection of objects masked by foliage and other clutter or a system that uses acoustic sensors for the detection of underwater objects must treat the environment as a significant component of the image formation problem. Such systems may need to use prior information about the scene, or the equivalent, to augment the limitations of the sensor data.

Image formation using a prior model often can be treated as an inverse scattering problem. The measurements of the scattered signal are inverted to estimate the parameters of a model. Inverse algorithms iterate a forward algorithm, which calculates the far-field scattering of known illumination by a known object, and compares that to the measured field. The model parameters are then adjusted to reduce the discrepancy between calculations and measurements. This process is repeated until there is a satisfactory agreement.

While the task of image formation can be viewed abstractly simply as a problem of estimation, it can also be viewed as having a character and content of its own. The problems addressed, the cost functions used, and the specific models that are used for images and image sensors lead to new questions and mathematical techniques, such as the estimation of random processes on manifolds or other complex surfaces. Creating a formal information-theoretic framework forces one to think through general principles and to either justify or reject existing *ad hoc* procedures.

Thus we come to our thesis. The time is right for a far-reaching study into our notions of extracting images or other object information from very large data sets, including data sets from multiple sensors, and possibly enriched by archived models and archived data. Various communities will react to this statement differently. The information theory and statistics community will think of maximum-likelihood models, information-theoretic measures of performance, and data fusion. The statistics community will also invoke methods of correlation and statistical inference. The computer science community, under the term "data mining," will think of large archived data structures and various search engines to supplement sensor data. All, however, will agree that such methods can be very powerful, and can extract information that is very subtly and deeply buried in a massive data set. It is now timely and appropriate to attempt to survey a framework at this level for image formation. Insights will emerge from an information-theoretic framework that may not be seen when studying an individual application. This paper has been written as an early step in this direction.

This paper deals with the information-theoretic aspects of image formation. Another important area in imaging that benefits from information-theoretic methods is image compression. This is an area of active research with a large literature and is beyond the scope of this paper, which focuses on image formation. For an introduction to this literature, see for example [36], [65], and [14].

The paper is organized as follows. The problem is first structured in Sections II, III, and IV entitled "Image Space," "Sensor Data," and "Reproduction Spaces." Then performance measures are discussed in Sections V and VI, entitled "Information-Theoretic Measures," and "Performance Bounds." Sections VII and VIII, entitled "Image Formation" and "Computational Algorithms," are the core of the paper. Examples are given in Section IX, entitled "Modalites and Applications."

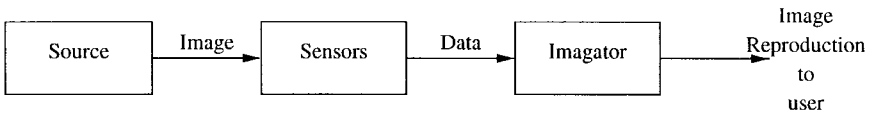

Fig. 1. Image formation model.

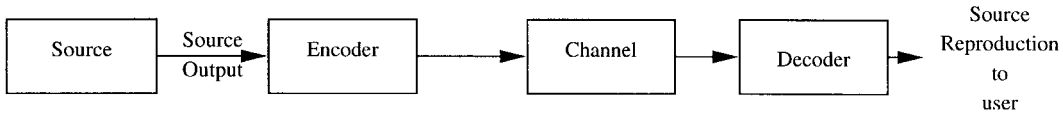

Fig. 2. Communications model.

II. Image Space

We shall discuss the image formation problem shown in Fig. 1. This model, which can describe many estimation problems, will be interpreted herein in the context of image formation. The image formation model shown in Fig. 1 is analogous to the standard communications model shown in Fig. 2. The communications model consists of a source, a channel, and a user, and these are connected by an encoder and a decoder. Information theory studies these abstract models of communications and image formation.

The image formation problem is concerned with an underlying image or scene that is analogous to the source output in a communications problem. The source selects one image from a set called an *image space*, and the selected image—or an adequate reproduction of that image—is to be provided to the user. The images in the image space are abstractions, perhaps similar to photographs of an underlying physical scene. Just as a photograph is a compressed representation of some underlying physical reality, so too, images in the image space are abstract compressed representations of a physical reality.

The sensors in Fig. 1 play the role of the channel. The "imagator" or image-formation algorithm plays the role of the demodulator and decoder. Unlike the communication model, the ways in which the images can be encoded or otherwise modulated into sensor waveforms is quite limited by the physics of the sensor interaction with the environment. The image data are encoded by nature into radiated signals such as electromagnetic waves, diffracted X-rays, acoustic waves, or seismic waves. These signals interact with the sensors to produce the data available in imaging problems.

In some cases, only parameters of the image are of interest, and the mapping from the parameters to the image may be viewed as a modulation of the data. The parameter space may be a low-dimensional space, with dimension corresponding to the position and orientation of an object of interest, or it may be of moderate dimension, such as when it consists of the parameters in a mixture model (as in a segmentation of the image). It may be of high dimension, such as a color spectrum that varies with position.

A. Nonparametric and Parametric Models

Image space is the set of model images that represent the true, underlying physical distributions that are measured by the sensors. The image space is denoted by $\mathcal{C} = \{c : \mathcal{D}_C \rightarrow \mathcal{R}_C\}$, where c is an image, $\mathcal{D}_C$ is the domain of the image, and $\mathcal{R}_C$ is the range of the image. Models of image space that describe an image by an infinite number of parameters, typically consisting of a real function on $\boldsymbol{R}^2$, are known as *nonparametric models*. A traditional nonparametric imaging problem may have domain $\mathcal{D}_C$ equal to $\boldsymbol{R}^2$ or a compact subset of $\boldsymbol{R}^2$. The range of the image c is commonly $\boldsymbol{R}$, $\boldsymbol{C}$, or $\boldsymbol{R}_+$, but sometimes it is a vector space $\boldsymbol{R}^m$, $\boldsymbol{C}^m$, or $\boldsymbol{R}_+^m$ consisting of the set of elements of $\boldsymbol{R}^m$ with nonnegative components. For example, densities of particles, attenuation functions, intensity functions, and power spectra have nonnegative values. Radar and coherent laser signals are complex-valued and may lead to complex-valued images of target reflectivity. To give a more elaborate example of an image, we note that a real-valued three-dimensional scene may be time-varying and so may be denoted $c(x, y, z, t)$. In this case, the domain is $\boldsymbol{R}^4$ and the range is $\boldsymbol{R}$. If viewed through a spectrally sensitive device, it may be advantageous additionally to model each point in space and time as having a spectrum associated with it. The domain is then $\mathcal{D}_C = \boldsymbol{R}_3 \times \boldsymbol{R} \times \boldsymbol{R}_+$, corresponding to a point in space, time, and frequency, so the domain is five-dimensional. Let $\boldsymbol{x} = [x, y, z]^T \in \boldsymbol{R}^3$ denote position, $t \in \boldsymbol{R}$ denote time, and $f \in \boldsymbol{R}_+$ denote frequency. Then $c(\boldsymbol{x}, t, f)$ is a point in the image, and $c(\boldsymbol{x}, t, \cdot) : \boldsymbol{R} \rightarrow \mathcal{D}_C$ is the frequency-dependent function associated with the point at position $\boldsymbol{x}$ at time t.

While typical scenes may be five-dimensional, or even larger if polarization effects are included, particular sensors may be insensitive to one or more of these dimensions. In that case, it is sufficient to project the five-dimensional function c onto the appropriate lower dimensional function. For example, if only a single measurement is made at a given fixed time, then the time variation may be ignored. If the measurement depends on the spectrum only through an inner product with a specified spectrum (the transfer function of the sensor), then the spectral dependence may be ignored, keeping only this projection. If the sensor is invariant to one of the three spatial dimensions, then that dimension may be ignored.

Models that describe an image by a finite number (or, rarely, by a countable number) of parameters are known as *parametric models*. Parametric models are important in model-based imaging or imaging. A *nonparametric model* for an image may consist of a restriction of the image to a function space or of a representation of the image as a countable linear combination of basis functions. The basis functions may correspond to a representation of the image in terms

of pixels, in terms of a transform-domain expansion, or in terms of an orthonormal set of functions. Even when this expansion is limited to a finite number of basis functions, the terminology would still be determined by the underlying view of the image as a function on $\boldsymbol{R}^2$, so the model could still be called a nonparametric model. An intermediate class consists of models that have a very large, but finite, number of parameters. These are called *hyperparametric models*. Most image space models, including the use of random field priors, are nonparametric descriptions. Hierarchical image models are usually hyperparametric models.

The standard imaging approach, which is to reconstruct an image as a finite array of pixels or voxels, is viewed as nonparametric. Occasionally it is desirable to combine a parametric model with a nonparametric model. This means that an expansion of an image in terms of basis functions may be found, and the coefficients in the expansion may be functions of the parameters of interest. This approach, because it relies on two steps, may not be optimal (by the data processing inequality) but may be necessary due to constraints on system implementation.

B. Priors

The observation space is denoted $\mathcal{O} = \{o : \mathcal{D}_{\mathcal{O}} \rightarrow \mathcal{R}_{\mathcal{O}}\}$ with domain $\mathcal{D}_{\mathcal{O}}$ and range $\mathcal{R}_{\mathcal{O}}$, respectively. A sensor maps the image space $\mathcal{C} = \{c : \mathcal{D}_{\mathcal{C}} \rightarrow \mathcal{R}_{\mathcal{C}}\}$ into the observation space; often this mapping is stochastic. Imaging problems are classified as either deterministic or stochastic according to whether the image and the sensor are described by deterministic or stochastic models. Typically, a deterministic model is used if little or nothing is known about the image and if the sensor noise is negligible.

Deterministic constraints are often a part of the model. These constraints will include nonnegativity constraints for functions such as intensity, attenuation, density, and scattering. The image may be known to take values in some convex subset of $\mathcal{C}$. Alternatively, the image may be parameterized in some way.

Other deterministic effects that must be captured include the projection effects of the sensor. For optical imaging, the projection onto the focal plane may be an orthographic or perspective projection; the projection onto the retina may be modeled as a spherical projection. For tomographic imaging, the data may be collected in parallel or fan beams.

In the remainder of this section, we consider the case in which something is known about the image. Prior knowledge about an image is most naturally incorporated through the use of a prior probability distribution $\mu(c)$ on the image space. Such priors may be specified directly on the image space $\mathcal{C}$, or may be specified on parameters in a parametric representation of the image space.

For some problems, the prior may be on a finite-dimensional parameter vector $\theta \in \Theta$ that characterizes the uncertainty in the image. For rigid objects, Θ may be the special Euclidean group of translations and rotations. In some passive scenarios (optical imaging, for example), there may be a scale parameter included in Θ. We assume that the set Θ is a finite-dimensional space with probability density function p_θ.

When there is such a parameter θ that characterizes the scene, it may provide a complete or a partial characterization. If it completely characterizes the scene, then c is a function of θ; that is, there is a deterministic mapping from Θ to $\mathcal{C}$ that assigns to each θ an image c. If it is a partial characterization, only part of the scene is characterized by θ. The remainder of the scene may be modeled either as an unknown deterministic function or as a stochastic process. In the latter case, a prior on the scene given the parameters is required and will be denoted by $\mu(c|\theta)$.

To recognize rigid objects in a scene automatically, the background is not directly of interest and so is regarded as a nuisance parameter. Conversely, to determine the background image in the presence of a rigid body, the rigid body may be regarded as the nuisance parameter. Alternatively, it may be of interest to estimate parameters associated with elements of the scene (such as positions and orientations of rigid bodies) and to form an image of the entire scene. This is the case in spiral tomographic imaging in the presence of high-density attenuators. Then the position and orientation of the high-density object are of interest, while simultaneously it is of interest to remove the streaking artifacts commonly seen in the images in the neighborhood of the object [103].

In object recognition problems, it may be that the image contains an object of interest, and the object is one of a finite number of object types. Further, it may be that only a determination of the object class is of interest. If the number of possible classes is fixed, then the problem of interest is one of hypothesis testing. Then the set of hypotheses will be denoted $\mathcal{H} = \{H_1, H_2, \cdots, H_m\}$. The corresponding prior probability distribution on this set of hypotheses, if there is a prior probability distribution, will be denoted $\boldsymbol{P} = \{P_1, P_2, \cdots, P_m\}$.

Finally, generalizations of these problems are often of interest. For example, in automatic object recognition problems, it may be required to estimate the number of objects in a scene, perform recognition on each detected object, estimate the position and orientation of each object, and form an image of the background. In this case, inference is performed on a very complicated, high-dimensional space.

C. Mathematical Representations

Scenes may be described either nonparametrically as functions taking values in a specified function space, parametrically in terms of known functions of parameters, or as some combination of these two. For example, a three-dimensional scene may have a known rigid object embedded in an unknown background. In this case, the image has both parametric and nonparametric components. The position and orientation of the rigid object take values in the six-dimensional space corresponding to both translations and rotations; the background is an unknown function.

This example illustrates a complicating aspect of many inference problems in imaging: the parameter space need not be isomorphic to $\boldsymbol{R}^n$. For example, the six-dimensional space of translations and rotations, denoted $SE(3)$, forms a non-Abelian group. Let $\boldsymbol{x} \in \boldsymbol{R}^3$ be a point on the rigid object. Let $\boldsymbol{R}(\theta)$ be a 3×3 rotation matrix, and $\boldsymbol{t} \in \boldsymbol{R}^3$ be a translation.

Then the point $\boldsymbol{x}$ is mapped according to

$$\boldsymbol{x} \mapsto \boldsymbol{R}(\theta)\boldsymbol{x} + \boldsymbol{t}. \tag{1}$$

An $n \times n$ rotation matrix takes values in the special orthogonal group of dimension n, denoted $SO(n)$. A rotation–translation pair takes values in the special Euclidean group $SE(n)$. In many problems, $SE(2)$ and $SE(3)$ are the groups that are relevant, these corresponding to translation and rotation in $\boldsymbol{R}^2$ and in $\boldsymbol{R}^3$, respectively.

More generally, let Θ be a finite-dimensional parameter space. The space Θ may be a subset of $\boldsymbol{R}^n$ or it may be a group. An image c in the image space may be written as the sum of two components

$$c = c_1(\theta) + c_2 \tag{2}$$

where $c_1 : \Theta \to \mathcal{C}$ is the part of the image determined by the finite parameterization and $c_2 \in \mathcal{C}$ is the nonparametric part of the image. This may be generalized further as $c = f(\theta, c_2)$, where f is an arbitrary function of θ and c_2.

A commonly used example of a parameterization is a hierarchical parameterization. A hierarchy of parameters is an ordered set of parameters, $(\theta_1, \theta_2, \cdots, \theta_m)$ with a Markov structure. Let $\theta_k \in \Theta_k$. If θ_1 corresponds to the finest scale of the hierarchy, then there is a mapping $c_1 : \Theta_1 \to \mathcal{C}$. For other scales in the hierarchy, there are mappings $h_k : \Theta_k \to \Theta_{k-1}$. In general, the mappings $\{c_1, h_2, \cdots, h_m\}$ may be stochastic. Inference is performed on the hierarchy, with different scales providing different pieces of information about the scene and objects within the scene.

A simple example of a hierarchical model would have two scales. Let Θ_2 be a set of possible objects. Assume there is one object of interest in the scene. Selection of an element of Θ_2 corresponds to the task of object detection. Suppose further that the mapping h_2 corresponds to translation and rotation of the object in the scene. The function c_1 is the image that results, given the object position, orientation, and type.

More complicated examples of hierarchical models may involve building complex objects up from simple objects. The hierarchy may progress from pixel values to edges, from edges to boundaries, from boundaries to regions, and from regions to object types.

D. Markov Random Fields

To employ probabilistic methods, one may regard an image as a random element drawn from a prespecified set of possible images. Then one must assign a prior probability distribution to the set of images, and this assignment leads to the introduction of the notion of a random field. A random field is a generalization of a random process to two or more dimensions. Random-field models are important in image formation because of analytic tractability, because they are a very good fit for many images in applications, because these methods are robust and still give satisfactory results even when not a good fit to a particular application, and because they can convert in an orderly way an ill-posed problem into a well-posed problem.

A *random field* is a multidimensional random process. For example, a Gaussian random field is determined by a mean function $\eta(\boldsymbol{x}, t, f)$ and an autocovariance function $K(\boldsymbol{x}_1, t_1, f_1; \boldsymbol{x}_2, t_2, f_2)$. A random field, possibly Gaussian, is an appropriate model for both real-valued images and complex-valued images. A prior that has been used successfully for imaging problems is the Markov random field.

Random variables characterized by conditional priors that account for local interactions are often used as natural and convenient priors in imaging problems. These conditional priors, placed directly on the image space $\mathcal{C}$ or on a subset or subspace of $\mathcal{C}$, are usually the most natural way to quantify our understanding of a problem. However, the fundamental probability distribution on the field is the joint probability distribution, and this is difficult or impossible to specify directly. One needs to verify that the chosen specification of conditional distributions is sufficient and consistent in the sense that a unique joint probability distribution corresponds to this set of conditional probability distributions. The simplest example is the Ising random field, which consists of a binary random variable defined at each site of the integer lattice $\boldsymbol{Z}^2$ with each random variable conditional on the value realized at each of the four nearest neighbors. An important aid in describing such collections of conditional priors is the Hammersley–Clifford theorem which states that under certain conditions, the most natural conditional probability functions do uniquely define global probability functions.

The notion of a Markov random field extends the notion of a Markov process to multidimensional spaces by generalizing the concept of order dependence that is fundamental in the definition of a Markov process. The well-known works of Ising contain the earliest application of Markov random fields. Later, Onsager used the classic Ising random-field model to characterize magnetic domains. Important early applications to the imaging problem include the work of Besag [4], who discusses a broad variety of Markov random fields and their applications, and of Geman and Geman [34] as well as Chellappa [13].

The generalization of a one-dimensional Markov random process to a multidimensional Markov random field is not straightforward because the concepts of past and future, which are quite natural in one dimension, do not have counterparts in higher dimensional spaces. Instead, the concept of a *neighbor* is used. A random process with index set T is given by $\{x(t) \mid t \in T, T = \{t_1, t_2, \cdots, t_k\}\}$. The random process is called a random field if the elements of T are vectors from a multidimensional space, such as the two-dimensional plane $\boldsymbol{R}^2$. Assume that the random variables $x(t_1), x(t_2), \cdots, x(t_k)$ are continuous random variables and that their joint probability density function $p_{x(t_1),x(t_2),\cdots,x(t_k)}(X_1, X_2, \cdots, X_k)$ exists. We shall also require for each i, $1 \le i \le k$, that the joint density of the $k-1$ random variables $\{x(t_j); j \ne i, 1 \le j \le k\}$ is strictly greater than zero

$$p_{x(t_1),\cdots,x(t_{i-1}),x(t_{i+1}),\cdots,x(t_k)}(X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_k) > 0. \tag{3}$$

The reason for imposing this positivity condition is to ensure the existence of the conditional densities

$$p_{x(t_i)|x(t_1),\cdots,x(t_{i-1}),x(t_{i+1}),\cdots,x(t_k)} \times(X_i \mid X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_k) \quad (4)$$

for $1 \le i \le k$. Site t_j is said to be a *neighbor* of site t_i if the above conditional density for $x(t_i)$ is dependent on X_j. Thus

$$\begin{aligned} &p_{x(t_i)|x(t_1),\cdots,x(t_{i-1}),(t_{i+1}),\cdots,x(t_k)} \\ &\quad\times(X_i \mid X_1,\cdots,X_{i-1},X_{i+1},\cdots,X_k) \\ &= p_{x(t_i)|x(t_j):t_j\in N_i}(X_i \mid \{X_j \text{ for all } j \text{ such that } t_j \in N_i\}) \end{aligned} \quad (5)$$

where N_i is the set of neighbors of t_i. The neighbor relationship is symmetric: if $t_j \in N_i$, then $t_i \in N_j$.

A random process $\{x(t)|t \in T\}$ with a discrete index set $T = \{t_1,t_2,\cdots,t_k\}$ and a set of neighborhoods $\{N_1,N_2,\cdots,N_k\}$ and conditional densities is called a *Markov random field*. A common example of a Markov random field prior is a Gauss–Markov random field. For a Gauss–Markov random field, each of the probability density functions in (5) is Gaussian.

In any numerical implementation, the region of computation must be truncated to a region of finite size. When this happens, some assumptions about boundary conditions must be made, and these assumptions may play a crucial role in the resulting image reconstructions because the Markov random field model may exhibit phase transitions due to the boundary conditions, resulting in poor reconstructions.

The boundary conditions generally fall into three categories: periodic boundaries, random boundaries, and fixed boundaries. For periodic boundaries, the neighborhood structure is made periodic, so that every lattice site has the same neigborhood structure. For other choices, the lattice is truncated, so specific assumptions must be made about how that truncation takes place. Lattice sites whose neighborhood structure differs from others due to truncation are called boundary sites. For random boundaries, the boundary site values are chosen randomly from the marginal distribution. For fixed boundaries, the boundary values are set to specific values such as the mean of the marginal distribution.

We shall now give several examples of Markov random fields.

The simplest example of a Markov random field is one in which the neighbors of t_i for each i, are *all* the other indices $\{t_j, j \ne i, 1 \le i \le k\}$. For this choice,

$$N_1 = \{t_2,t_3,\cdots,t_k\}$$
$$N_k = \{t_1,t_2,\cdots,t_{k-1}\}$$

and

$$N_i = \{t_1,\cdots,t_{i-1},t_{i+1},\cdots,t_k\}$$

for $2 \le i \le k-1$. This example, with no restrictions on the statistical dependencies between the variables, indicates that Markov random fields can be quite general.

A temporal Markov random process is another simple example of a Markov random field. Suppose that $\{x(t)|t\in T\}$ is a random process with a discrete index set $T = \{t_1,t_2,\cdots,t_k\}$ where each t_i is a real number representing time. The neighborhood structure is $N_k = \{t_{k-1}\}, N_1 = \{t_2\}$, and $N_i = \{t_{i-1},t_{i+1}\}$. This is the standard definition of a first-order Markov random process.

Another simple example of a Markov random field can be constructed as follows. Define the index set to be of the form

$$T = \{t_{ij} \mid 0 \le i \le M_\ell - 1, 0 \le j \le M_j - 1\}$$

which may arise as a Markov field on the lattice points of a discrete pixelization of a finite rectangular region of the plane. Suppose that each interior point t_{ij} of a region has four neighbors defined as the four lattice points lying to the North, East, South, and West, so the neighborhood of t_{ij} is

$$N(t_{ij}) = \{t_{i,j+1},t_{i+1,j},t_{i,j-1},t_{i-1,j}\}.$$

Each boundary point of a region has its neighbors defined similarly except that only lattice points within the region can be neighbors. This set of neighborhoods along with a specified conditional density for the process at each lattice point defines a two-dimensional Markov random field with nearest neighbor dependency.

It can be cumbersome to perform inference on the random field directly, so typically some discretization of the problem is required. Such a discretization may be a pixelization or an expansion in some other basis. For Gaussian random field models, the natural expansion would be in terms of the eigenfunctions of the autocovariance function. For other random field models, different expansions may be appropriate, as discussed below.

III. Sensor Data

The observation data are available at the output of the set of sensors. These data may be modeled either deterministically or stochastically, as appropriate for the given application.

Stochastic models arise naturally in many situations. In a radar receiver there is internal noise, and there may be external sources of noise as well. Both of these impairments of the desired signal are unrelated to the scene in question, and both are often modeled as additive Gaussian noise in the microwave frequency band or the intermediate frequency band, or as complex Gaussian noise at complex baseband. In optical bands, the data arise from photoconversions in the sensor. This situation is naturally modeled using a Poisson counting process model or, if readout noise is significant, as a Poisson–Gaussian mixture model. In positron-emission tomography, the production of annihilation photons within the patient volume is accurately modeled as a spatial point process; the resulting positron detections are well-modeled as Poisson distributed events.

In many problems, however, noise is insignificant, and stochastic models are not as well motivated, so deterministic models are used. In photographic applications, the intensity incident on the film is blurred by the lens and other known quantities in the optical path. For such optical imaging systems, the data are collected as images in a focal plane; there is also a known perspective or orthographic projection of the

three-dimensional scene to the focal plane. In X-ray tomographic applications, the intensity may be high enough that the Poisson statistics of the detection process can be ignored and the data reasonably modeled as being deterministic; standard analysis of computed tomography systems models the data this way, and the problem of recovery of the unknown attenuation function is treated as a deterministic inverse problem.

In general, both deterministic and stochastic effects must be taken into account. The finite size of detectors has a deterministic effect on the collected data. The bandwidth of radar signals, the motion between the antenna and the scene, and the geometry and electrical characteristics of radar antennas yield deterministic effects on radar data and any images generated from the radar data such as synthetic-aperture radar images. These may then need to be combined with stochastic models for the detection process and receiver noise. In optical systems, a deterministic model for the effects of the lens and the known geometry may need to be combined with a stochastic model for the detector and a stochastic model for turbulence in the optical path.

If the model for the available data is a deterministic one, then the observation o is a deterministic function of the image c, $o = f(c)$. Examples of such models are discussed in the applications section. A stochastic model is defined by a conditional distribution on the observation given the image. When coupled with a deterministic projection, the likelihood often can be written $L(o \mid f(c))$, where f is a known function.

There may be parameters that enter into the problem in various ways. For high-precision problems of machine vision, a known object may be viewed using a camera that has unknown parameters. In the calibration step, the focal length, the image center on the focal plane, and parameters for lens distortion and focal plane nonuniformity often need to be estimated from the data. In actual use, the position and orientation of the object must be estimated. For typical machine vision problems, a deterministic model is used and general optimization procedures are applied to find the parameters [45], [97]. In other problems, analogous parameters describing blurring functions, optical centers, and projections may need to be estimated, either in calibration steps or in every image.

Stochastic models may be combined with priors on scenes to find joint likelihoods for the data and for the underlying scenes. The joint distribution on the observations and the scene is then the product of the sensor's conditional distribution and the prior, $\pi(o \mid c)\mu(c)$.

IV. Reproduction Spaces

The image space is designed to model as closely as possible an idealized representation of the physical processes that generate the data. There will always be aspects of the underlying physical situation that are not captured in the image space model. Indeed, one of the most challenging problems in information-theoretic imaging is the development of models for the underlying physical processes that are adequate for the problem at hand, but not so complicated as to present intractable mathematics. The reproduction space is the set of functions in which a computational algorithm for image formation produces its output values. The selection of a reproduction space must anticipate the needs and limitations of the computational algorithm. In the selection of the reproduction space, there is a tradeoff between its ability to represent images in $\mathcal{C}$ closely and the computational complexity of the resulting algorithm.

The estimated image itself is often a discrete approximation of the underlying conceptual image which usually is a continuous distribution. The discreteness could be due to quantization of values associated with the image, but usually also includes a representation of the continuous image that is sampled or pixelated in some way. For example, in astronomical imaging, there is an underlying intensity distribution corresponding to the distribution of the astronomical object being viewed. This intensity function is defined on a continuous domain. Computed images are presented as discrete values on an array of pixels, which may be considered to be an approximation of this true underlying distribution, this approximation or estimate consisting of a summation of pixel values that scale appropriate basis functions.

There is also a tradeoff between bias and variance, or a tradeoff between approximation error and estimation error. The higher the dimension of the reproduction space, the more closely the underlying image in $\mathcal{C}$ may be represented. That is, the discrepancy between the closest element of the reproduction space and the true image decreases as the size of the reproduction space grows (or its bias decreases). On the other hand, as the dimension of the reproduction space grows, the statistical variation in the estimate grows. That is, the discrepancy between the estimate in the reproduction space and the element of the reproduction space closest to the truth increases (or its variance increases). For any given imaging problem and some measure of the sum of these two terms, there is usually an optimal size of the reproduction space that minimizes the measure.

The reproduction space is typically a subset or a subspace of the image space. Often, the reproduction space is parameterized and the parameters may be varied as data are collected to refine existing image estimates as more data becomes available. The refinements should take values in successively higher dimensional subsets of image space. The use of sieves provides a framework for indexing the parameter in the refinements in order to achieve consistency of the image estimates as the amount of data collected increases (see Grenander [42]).

Much of the theory underlying this formulation falls within the subject of approximation theory. We will not attempt to survey the results within this broad research area, but will summarize some of the aspects that are relevant to image formation problems. In applications, computational issues may influence the choice of the reproduction space used.

Information-theoretic discrepancy measures are useful throughout this paper. For each space, we assume that there is a discrepancy measure between two elements of that space.

Definition: A discrepancy measure on a space $\mathcal{X}$ is any mapping $d : \mathcal{X} \times \mathcal{X} \to \boldsymbol{R}_+$ such that $d(x_1, x_2) \geq 0$ for all $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$ and $d(x_1, x_2) = 0$ if and only if $x_1 = x_2$ almost everywhere.

For the spaces $\mathcal{C}, \mathcal{O}, \Theta$, and $\mathcal{H}$, we will require that discrepancy measures $d_{\mathcal{C}}, d_{\mathcal{C}}, d_{\Theta}$, and $d_{\mathcal{C}}$ have been defined.

A. Regularization

In most imaging problems, the underlying image takes values in an infinite-dimensional space. The reconstruction or estimation of the image based on some given data then implies that an infinite number of variables that represent the image must be determined. Such problems are usually ill-posed, in the sense described below, and some form of regularization is required to reduce the sensitivity of the reconstruction to variations in the observed data, mismatches between the data and the model adopted for the data, and to choices made in the implementation of the reconstruction algorithm (such as the number of bits of precision and the order in which arithmetic operations are performed). One approach to regularization is the representation of an image using pixels or voxels, but this can exhibit the undesirable consequences of "dimensional instability," as described by Tapia and Thompson [55], and a tradeoff between reconstruction accuracy (when a continuous image is approximated by a piecewise-continuous one) and resolution. This necessity for making tradeoffs accompanies other forms of regularization as well; that tradeoff may be between bias and variance, or between approximation error and estimation error. Tikhonov [96], Hadamard, and Joyce and Root [53] have addressed the instability problems caused by ill-posedness and have suggested approaches for regularization to avoid such problems.

Anticipating the formal statement of an image formation problem below, suppose that the optimal image $\hat{c}(o)$ is defined as the c that achieves

$$\hat{c}(o) = \arg\min_{c \in \mathcal{C}} \psi(o, c) \tag{6}$$

where $\psi(o, c)$ is a given objective function. The notation $\hat{c}(o)$ in (6) implies that there is a mapping from $\mathcal{O}$ to $\mathcal{C}$. If there are multiple $c \in \mathcal{C}$ that achieve the minimum, then we require that one of them is chosen arbitrarily. While typical examples of possible choices for ψ are discussed in Section VII, we introduce a few possibilities here.

For stochastic models, ψ may equal the negative of the log-likelihood function of o. If c has a deterministic model, then this is the conditional log-likelihood of o given c. If c has a stochastic model, then this is the joint log-likelihood of the pair (o, c), and the minimization is equivalent to finding the maximum *a posteriori* (MAP) estimate of c given o.

For deterministic models, ψ may be the discrepancy between a predicted observation $f(c)$ and the true observation o. As discussed below, additional terms may be included in ψ.

Definition: A statement posing the arg minimum is *well-posed* if for each $o \in \mathcal{O}$, there is a unique $c \in \mathcal{C}$ that achieves the minimum, and

$$\sup_{o \in \mathcal{O}} \lim_{o_1 \to o} \frac{d_{\mathcal{C}}(\hat{c}(o_1), \hat{c}(o))}{d_{\mathcal{O}}(o_1, o)} \tag{7}$$

exists and is finite. A statement posing the arg minimum that is not well-posed is *ill-posed*.

The intention behind this definition is that, when the problem is well-posed, small changes in the data should produce small changes in the estimate. The size of a change is measured using a discrepancy measure. If a problem is well-posed, then the value of the limit in (7) is one measure of the degree to which the problem is well-posed. For deterministic inverse problems, the condition number of the mapping may be preferred as an alternative measure.

For a deterministic inverse problem, with the data modeled as $o = f(c)$, the norm of f is defined by

$$||f|| = \sup_{c \in \mathcal{C}} \lim_{c_1 \to c} \frac{d_{\mathcal{O}}(f(c_1), f(c))}{d_{\mathcal{C}}(c_1, c)}. \tag{8}$$

If f is invertible (that is, for each $o \in \mathcal{O}$, there is a unique $c \in \mathcal{C}$ such that $f(c) = o$), then the norm of f^{-1} is defined as in (8). The condition number of f is then

$$\chi(f) = ||f||\ ||f^{-1}||. \tag{9}$$

Note that this definition is the square of the standard definition of condition number if d is squared error. This definition of condition number is suitable for our consideration because a small change in c, from c to c_1, yields a change in o that is less than $||f|| d_{\mathcal{C}}(c_1, c)$. In turn, the discrepancy between the resulting estimate of c and the original value is less than

$$||f||\ ||f^{-1}|| d_{\mathcal{C}}(c_1, c) = \chi(f) d_{\mathcal{C}}(c_1, c). \tag{10}$$

As discussed extensively in the literature of numerical analysis (see Golub and Van Loan [37]), the condition number quantifies the sensitivity of the problem to numerical and approximation errors.

Ill-posed problems such as the image formation problem defined in (6) using the objective function ψ may be associated with a family of well-posed problems, by defining a family of subsets of $\mathcal{C}$ that converge to $\mathcal{C}$. Suppose that a one-parameter family of sets $\{\mathcal{C}^{\nu}, \nu > 0\}$ is given and that the closure of this family equals $\mathcal{C}$

$$\text{Closure}[\cup_{\nu > 0} \mathcal{C}^{\nu}] = \mathcal{C}. \tag{11}$$

This assumes the definition of a topology on $\mathcal{C}$. In many cases, the interest is in constructing, for each $c \in \mathcal{C}$, a sequence of functions $c^{\nu} \in \mathcal{C}^{\nu}$ such that

$$\lim_{\nu \to 0} d_{\mathcal{C}}(c^{\nu}, c) = 0. \tag{12}$$

Definition: Suppose that the family of subsets $\mathcal{C}^{\nu}$ satisfies (11) and that for each $o \in \mathcal{O}$ there is a unique $\hat{c}(o)$ that achieves the minimum in (6). The family of problems

$$\hat{c}^{\nu}(o) = \arg\min_{c \in \mathcal{C}^{\nu}} \psi(o, c) \tag{13}$$

is a *regularization* of the image formation problem defined in (6) if for each ν the problem defined by (13) is well-posed and if for almost all $o \in \mathcal{O}$

$$\lim_{\nu \to 0} \hat{c}^{\nu}(o) = \hat{c}(o). \tag{14}$$

If the discrepancy measure is a distance or a squared distance, then for each ν, the mapping $\hat{c}^{\nu} : \mathcal{O} \to \mathcal{C}$ is continuous.

For an ill-posed problem, the function $\hat{c}$ is discontinuous, so that even small amounts of noise in the observations can lead to large changes in the estimates. For each nonzero value of ν, however, the mapping $\hat{c}^\nu$ is continuous so the regularized solution is less sensitive to noise. The functions $\hat{c}^\nu$ converge to $\hat{c}$ as ν goes to zero, so the sensitivity to noise increases as ν goes to zero. For pixelization of images, this problem is called "dimensional instability" by Tapia and Thompson [95], who observe that the estimates become increasingly ill-behaved and unstable as the discretization is refined even as more data are collected.

This definition of a regularization is a modification of the statements in Youla [106], Grenander [42, p. 358], and Kirsch [54, pp. 24–26], modified to account for the statement of the imaging problem as an inverse problem, the solution of which is defined in terms of the objective criterion ψ. Various models, including linear, nonlinear, stochastic, and deterministic models are included within this statement. A feature of this definition of regularization is that it is in terms of the reproduction space rather than the mappings $\hat{c}^\nu$. This is a departure from the statements by Youla and Kirsch who require directly that the mappings $\hat{c}^\nu$ are continuous. Grenander's definition [42, p. 358] is in terms of a one-parameter family of operators acting on $\mathcal{C}$. That is, his mapping is meant to regularize $\hat{c}(o)$, by using operations such as lowpass filtering (projections onto subspaces or subsets of $\mathcal{C}$). Later in [42], and in other settings using his method of sieves, Grenander uses a concept of regularization consistent with the definition given here.

A regularization method in general provides a framework within which the ill-posedness can be addressed quantitatively. We discuss the use of penalties, prior probability distributions, kernel sieves, and choice of reproduction space as regularization methods. The simplest and most common way to regularize a problem is by the use of pixelization. An image displayed using pixels is really a projection onto a finite-dimensional subspace. A measure of the size of a pixel is the regularization parameter ν. There are many other standard restrictions of images to subspaces, with $1/\nu$ corresponding roughly to the dimension of the subspace.

A penalty regularization alters the objective function by adding a penalty to it. Tikhonov [96] introduced a quadratic penalty. More generally, a penalty can be added to the objective function as a discrepancy between the estimate and a nominal value $d(c, c_0)$.

Given a model for the data in terms of a conditional likelihood function $L(o \mid f(c))$, and a discrepancy measure $d_{\mathcal{C}}$ on $\mathcal{C}$, there may be an optimal ν that minimizes a tradeoff between approximation error and estimation error, as described next.

Assume that $c^{\nu *} \in \mathcal{C}^\nu$ is the unique element of $\mathcal{C}$ that minimizes

$$c^{\nu *} = \arg\min_{c_1 \in \mathcal{C}^\nu} d_{\mathcal{C}}(c_1, c) \tag{15}$$

where $c \in \mathcal{C}$ is the true image. Then the sum of the estimation error and the approximation error is

$$\delta^\nu(c, o) = d_{\mathcal{C}}(\hat{c}^\nu(o), c^{\nu *}) + d_{\mathcal{C}}(c^{\nu *}, c). \tag{16}$$

For a stochastic problem, this error is a random variable. Typically, the first term (the estimation error) is monotonically increasing and the second term (the approximation error) is monotonically decreasing as ν decreases. For a typical problem, there is typically an optimal ν that minimizes a measure of $\delta^\nu(c, o)$ such as its expected value. The motivation for defining the sum in (16) is that for some discrepancy measures of interest, if $\mathcal{C}^\nu$ is a linear subspace

$$d_{\mathcal{C}}(\hat{c}^\nu(o), c) = d_{\mathcal{C}}(\hat{c}^\nu(o), c^{\nu *}) + d_{\mathcal{C}}(c^{\nu *}, c). \tag{17}$$

This holds for discrimination and squared error as discrepancy measures. Choosing ν to minimize the expected value of $d_{\mathcal{C}}(\hat{c}^\nu(o), c)$ is a precise way to define an optimal regularization. If $\mathcal{C}^\nu$ is a convex set, then for these same discrepancy measures

$$d_{\mathcal{C}}(\hat{c}^\nu(o), c) \geq d_{\mathcal{C}}(\hat{c}^\nu(o), c^{\nu *}) + d_{\mathcal{C}}(c^{\nu *}, c). \tag{18}$$

In this case, choosing ν to minimize the expected value of (16) corresponds to minimizing a lower bound on the mean of $d_{\mathcal{C}}(\hat{c}^\nu(o), c)$.

B. Pixelization

The reproduction space often can be viewed as consisting of a linear combination of basis functions. The most common example of such a representation is when the basis functions are indicator functions on some domain, and the resulting representation is referred to as a pixelization. When an image is represented as an array, the elements in the array are the coefficients in the linear combination. The basis functions need not always be viewed as indicator functions, however. If the image is assumed to have a fixed bandwidth and the representation is on a fine enough scale, the coefficients may be viewed as samples of the image. In that case, the basis functions are optimal interpolation functions. If the data are actually the result of integrating the image against a known kernel, then the coefficients may be viewed as the values used in a discrete approximation of the integral. If the integration is modeled as a Riemann sum, then the basis functions are indicator functions. If the integration is modeled as using a trapezoid rule for numerical integration, then the basis functions are first-order splines.

We shall describe some of the issues first in a one-dimensional setting and then a multidimensional setting. Let $a(t)$ be a function of time, and let a_k, for $k = 1, 2, \cdots, n$, be samples of $a(t)$ at times kT. If $a(t)$ is represented by the samples, there is an assumed nominal representation. One representation is as a linear combination of indicator functions

$$\hat{a}(t) = \sum_{k=1}^{n} a_k \Phi_T(t - kT) \tag{19}$$

where

$$\Phi_T(t) = \begin{cases} 1, & -T/2 \leq t < T/2 \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Another choice, for lowpass functions, is as a linear combination of interpolation functions

$$\hat{a}(t) = \sum_{k=1}^{n} a_k p_T(t - kT) \tag{21}$$

where

$$p_T(t) = \operatorname{sinc}(t/T) \tag{22}$$

and $\operatorname{sinc}(t) = \sin(\pi t)/\pi t$. Other choices can be made for the interpolation function, and the values a_k do not necessarily correspond to samples of $a(t)$. In orthogonal representations, the values typically represent inner products of the function with basis functions.

For the multidimensional setting, let $\Phi^\nu = \{\phi_{\boldsymbol{k}}^\nu,\ \boldsymbol{k} \in \mathcal{L}\}$ be an orthonormal set of functions, where $\boldsymbol{k}$ is a discrete index taking values on the lattice $\mathcal{L}$, and where ν is a parameter roughly corresponding to the resolution of the functions. Assume that $\mathcal{C}$ consists of square integrable functions, and denote the inner product on $\mathcal{C}$ by $\langle \cdot, \cdot \rangle$. Then an ideal expansion of $c \in \mathcal{C}$ using the basis Φ^ν is obtained as

$$\hat{c}^\nu = \sum_{\boldsymbol{k} \in \mathcal{L}} c[\boldsymbol{k}] \phi_{\boldsymbol{k}}^\nu \tag{23}$$

where

$$c[\boldsymbol{k}] = \langle c, \phi_{\boldsymbol{k}}^\nu \rangle. \tag{24}$$

The expansion in (23) is a representation of $\hat{c}$ in the subspace $\mathcal{C}^\nu \subset \mathcal{C}$ consisting of all linear combinations such that

$$\sum_{\boldsymbol{k} \in \mathcal{L}} |c[\boldsymbol{k}]|^2 < \infty. \tag{25}$$

The parameter ν indexes the subspaces so that

$$\lim_{\nu \to 0} \mathcal{C}^\nu = \mathcal{C} \tag{26}$$

in the sense that for all $c \in \mathcal{C}$

$$\lim_{\nu \to 0} \langle c - \hat{c}^\nu,\ c - \hat{c}^\nu \rangle = 0. \tag{27}$$

The statement in (27) is valid for deterministic convergence. For the stochastic setting, the corresponding statement involves stochastic convergence. For convergence in a mean-square sense, given a prior on the image space $\mathcal{C}$ such that $E\{\langle c, c \rangle\} < \infty$, the sequence of functions $\hat{c}^\nu$ converges to c in a mean-square sense if

$$\lim_{\nu \to 0} E\{\langle c - \hat{c}^\nu,\ c - \hat{c}^\nu \rangle\} = 0. \tag{28}$$

A more general setting involves convergence of the mean discrepancy $E\{d_{\mathcal{C}}(\hat{c}^\nu, c)\}$ to zero. Other modes of convergence of $\hat{c}^\nu$ to c may also be studied.

This description can be modified to allow for $\mathcal{C}^\nu$ to consist of functions that do not form an orthonormal set. This is the typical case for polynomial splines and for some multiresolution expansions. The extension involves using a different function to extract the coefficients than is used in the expansion itself. For polynomial splines there is an additional complication that, for expansions that are not ideally chosen, the functions often are not linearly independent. There have been some descriptions of the use of "frames" to cover this situation (see [26] and [46]).

Often, the expansion functions $\phi_{\boldsymbol{k}}^\nu$ are translated versions of a single-basis function ψ^ν. Specifically, assume that $\mathcal{C} = L_2(\boldsymbol{R}^n)$, and that the sample points occur on a regular lattice in $\boldsymbol{R}^n$ with lattice basis elements $\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_n$. Any point on the lattice is then specified by a unique integer vector $\boldsymbol{k} \in \boldsymbol{Z}$ and equals $\sum_{i=1}^n k_i \boldsymbol{\xi}_i$. We then have

$$\phi_{\boldsymbol{k}}^\nu(\boldsymbol{x}) = \psi^\nu\left(\boldsymbol{x} - \sum_{i=1}^{n} k_i \boldsymbol{\xi}_i\right). \tag{29}$$

The parameter ν is a measure of the size of the Voronoi cells in the lattice.

In polynomial spline expansions, the basis function ψ^ν is a polynomial in its n arguments. Clearly, there are infinitely many choices for the degree of the polynomial in its arguments. For a general discussion of splines, see the books by Chui and by Wahba [16], [102]. The simplest polynomial spline is a constant over an interval and zero outside of that interval. For the lattice described above, let

$$\psi_0^\nu(\boldsymbol{x}) = \begin{cases} \frac{1}{\sqrt{A_0}}, & \|\boldsymbol{x}\|^2 < \|\boldsymbol{x} - \sum_{i=1}^{n} k_i \boldsymbol{\xi}_i\|^2, \quad \forall \boldsymbol{k} \neq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{30}$$

Note that ψ_0^ν is proportional to an indicator function on the Voronoi cell of the lattice point at the origin and that A_0 is the volume of the Voronoi cell. The reproduction space in this case consists of images that are piecewise-constant.

C. Penalty and Constraint Methods

A common method of regularization of image estimates is to use penalty or constraint methods. To motivate these methods, a specific class of examples is used.

Suppose that $\mathcal{C}^\nu = \{c \in \mathcal{C} : d_{\mathcal{C}}(c, c_0) \leq 1/\nu\}$, where c_0 is a nominal value. Then the regularization problem (13) is a constrained optimization problem. The constraint may be incorporated using a Lagrange multiplier, α, changing the criterion to

$$\psi(o, c) + \alpha d_{\mathcal{C}}(c, c_0). \tag{31}$$

For a wide class of discrepancy measures and criteria ψ, if the constraint is satisfied with equality, then there is a one-to-one correspondence between the value of α and the value of the constraint $1/\nu$. This shows an equivalence between constraint methods and penalty methods, where the additional term $\alpha d_{\mathcal{C}}(c, c_0)$ is viewed as a penalty.

If a squared-error discrepancy measure is used, then this yields a quadratic penalty; typically c_0 would be chosen to be zero. If discrimination is used as the discrepancy measure, then this yields entropy-type penalties. For example, if c_0 is a constant then $d_{\mathcal{C}}(c, c_0)$ is the Shannon entropy of the function c.

This approach also yields a method to combine positive-valued images c with real- or complex-valued data o. Then the

function $\psi(o,c)$ may be a squared error, and the discrepancy measure $d_{\mathcal{C}}$ may be discrimination. Similarly, if the images are real- or complex-valued and the data are positive, ψ may be discrimination and $d_{\mathcal{C}}$ may be squared error.

Roughness penalties are often based on Good's roughness measure [38], [61], [69], [95]. When restricted to a pixelization, Good's roughness measure may be written as a sum of discriminations between the image and shifted copies of the image [69].

D. Transform-Domain Representations

A traditional engineering approach to the study of function representation is to manipulate a function in its transform-domain representation. This includes Fourier representations and wavelet representations as special cases. In Fourier-domain representations, there are several options for considering convergence of the representations. One is to assume that the image space is space-limited. The set of Fourier-series coefficients obtained by considering a periodic extension of the image space completely characterizes the image space. The parameter $1/\nu$ may correspond to the number of coefficients used in an expansion.

A second Fourier-domain option is to assume that the image is effectively bandlimited with bandwidth proportional to $1/\nu$. For each value of ν, the image is represented by its samples using the Nyquist–Shannon interpolation formula. This is equivalent to the ideal interpolation discussed in the pixelization subsection above.

There are other options for Fourier-domain representations. Typically, they correspond to expansions of the image space in the Fourier domain.

In wavelet representations, the parameter ν may correspond to the number of scaling levels considered or to sampling in the time-scale domain.

E. Sieves

A powerful method for regularization, which incorporates additional structure, is the method of sieves due to Grenander [42]. In the method of sieves, a sequence of subsets of the image space is defined and used to address issues of convergence of image estimates as the amount of data increases. For this discussion, we follow [42], assuming the underlying image c is a deterministic parameter and that there is a stochastic model for the data o given c. The conditional likelihood for the observation o given the underlying image c is denoted $\pi(o \mid c)$. In this setting, there is a true underlying image which is denoted c_0.

Definition: A one-parameter family of subsets of $\mathcal{C}$, $\{\mathcal{C}^\nu, \nu > 0\}$ is a *sieve* if the following conditions hold:

1) for almost every $o \in \mathcal{O}$, the maximum-likelihood estimate of c,
$$\hat{c}_{\mathrm{ML}}(o) = \arg\max_{c \in \mathcal{C}} \pi(o \mid c) \tag{32}$$
exists and is unique;
2) the closure of the union of subsets equals $\mathcal{C}$
$$\text{Closure}\left[\cup_{\nu>0} \mathcal{C}^\nu\right] = \mathcal{C}; \tag{33}$$
3) for each ν, the maximum-likelihood estimate restricted to $\mathcal{C}^\nu$
$$\hat{c}^\nu(o) = \arg\max_{c \in \mathcal{C}^\nu} \pi(o \mid c) \tag{34}$$
exists and is unique.

Note that this definition is essentially the same as the definition of regularization. The most visible use of sieves has been in studying the consistency of estimates. Let $o_1^n = \{o_1, o_2, \cdots, o_n\}$ be independent and identically distributed (i.i.d.) observations with distribution function $\pi(\cdot, c_0)$. Denote the maximum-likelihood estimate of c restricted to $\mathcal{C}^\nu$ by $\hat{c}_n^\nu(o_1^n)$. Grenander [42, Ch. 9] proves that under conditions on the continuity and boundedness of the restricted log-likelihood function, and uniqueness and continuity of the discrimination function, there is a sequence of $\nu(n)$ so that $\hat{c}_n^{\nu(n)}(o_1^n)$ converges to c_0, with probability one.

The parameter ν in the definition of a sieve is referred to as the mesh size. In some instances, the sequence of spaces is nested (monotonic) in the sense that if $\nu_1 > \nu_2$, then $\mathcal{C}^{\nu_1} \subset \mathcal{C}^{\nu_2}$. This might appear to be a natural condition, but it is not necessary. Further details are given by Grenander [42], Chow and Grenander [15], Moulin [64], and Moulin, O'Sullivan, and Snyder [67].

Let $\mathcal{X} = \boldsymbol{R}^n$, so that $c : \boldsymbol{R}^n \to \boldsymbol{R}$. Let $k_\nu(\boldsymbol{x})$ be a function indexed by ν, referred to as the kernel. A *kernel sieve* is a set $\mathcal{C}^\nu$ all of whose elements can be written as the result of a convolution with k_ν
$$c(\boldsymbol{x}) = \int k_\nu(\boldsymbol{x} - \boldsymbol{\xi})\gamma(\boldsymbol{\xi})\, d\boldsymbol{\xi}. \tag{35}$$
The functions k_ν converge in distribution to a Dirac delta function as $\nu \to 0$. One choice, discussed in [89], is a circularly symmetric Gaussian kernel with space parameter ν
$$k_\nu(\boldsymbol{x}) = \frac{1}{(2\pi\nu)^{n/2}} e^{-\frac{1}{2\nu}\boldsymbol{x}^T\boldsymbol{x}}. \tag{36}$$

F. Convergence of Sequences

The asymptotic properties of estimators are often important in estimation problems. As the amount of information increases, the estimates should converge to the truth in some sense. For regularization as described above, a setting for studying convergence is introduced.

Let $d_{\mathcal{C}}(\hat{c}, c)$ be a discrepancy measure. Suppose now that the estimate $\hat{c}$ is formed from the observations. Let T be a measure of how informative the observations are. For example, T could be the time-integration interval over which data are collected or the number of independent observations in a dataset. Suppose for each T, that the estimate lies within $\mathcal{C}^{\nu(T)}$. Then, $d_T = d_{\mathcal{C}}(\hat{c}^\nu(o_T), c)$ is a random variable indexed by T. Its randomness arises due to the observations and possibly due to c being a random process. The family of estimators $\hat{c}$ is said to be consistent in the "R" sense if the random variables d_T converge to zero in the "R" sense. Here "R" may be almost everywhere, mean-square, in probability, or in distribution. That is, convergence of a discrepancy measure between the truth and the estimate in some sense.

G. Convex Constraints

The reproduction space may be constrained either by prior knowledge or by the limitations of the available sensor data. We shall be especially interested in those constraints that satisfy a convexity property because they are analytically tractable and arise frequently. In particular, the set of probability distributions on a finite set is a convex set, and information-theoretic constraints usually satisfy a convexity property. For deterministic models, the problem becomes one of finding an element of the convex set that is most consistent with the data in terms of minimizing the discrepancy between the predicted data and the available data (as discussed in Section VII). A general introduction to convex constraints is given by Combettes [17].

A set $\mathcal{A}$ is *convex* if each convex combination of two elements from $\mathcal{A}$ is also in $\mathcal{A}$. That is, for any $c_1, c_2 \in \mathcal{A}$, then for all $0 \leq \lambda \leq 1$, $\lambda c_1 + (1-\lambda)c_2 \in \mathcal{A}$. The intersection of any number of convex sets is a convex set. A *convex constraint* is a statement that the image lies in a given convex set.

Many constraints that arise very naturally are convex constraints. One example is a nonnegativity constraint. If c is real-valued, then the nonnegativity constraint that $c(\boldsymbol{x}) \geq 0$ for all $\boldsymbol{x}$ is a convex constraint. This nonnegativity constraint can also be used within the class of complex-valued functions because the set of functions $c(\boldsymbol{x})$ that are both real and nonnegative is a convex set within the set of complex functions. Thus the convex constraint can be used to enforce a very natural property of images while still allowing the larger space of complex functions to play a role in the theory.

Our second example of a convex constraint is an energy constraint for square-integrable functions. Suppose that the image space $\mathcal{C}$ is equipped with an inner product $\langle \cdot, \cdot \rangle$, and a corresponding norm

$$\|c\|^2 = \langle c, c \rangle. \tag{37}$$

Let $\mathcal{A}_B \subset \mathcal{C}$ be defined as the set of images whose norm is less than some constant B

$$\mathcal{A}_B = \{c \in \mathcal{C} : \|c\| < B\}. \tag{38}$$

By the triangle inequality on the inner product norm

$$\begin{aligned} \|\lambda c_1 + (1-\lambda)c_2\| &\leq \|\lambda c_1\| + \|(1-\lambda)c_2\| \\ &= \lambda^2 \|c_1\| + (1-\lambda)^2 \|c_2\| \end{aligned} \tag{39}$$

so we can conclude that $\mathcal{A}_B$ is a convex set. More generally, convex constraints often may be defined in terms of discrepancy measures. For this to hold, we need that for each $c_0 \in \mathcal{C}$

$$\mathcal{A}_B = \{c \in \mathcal{C} : d_{\mathcal{C}}(c, c_0) < B\} \tag{40}$$

is a convex set. If, in addition, the sets $\mathcal{A}_B$ in (40) are compact, then they may be used to define a regularization. Simply set $\nu = 1/B$ and $\mathcal{C}^\nu = \mathcal{A}_{1/\nu}$.

Our third example of a convex constraint is a support constraint. Let the set $\mathcal{X}_S \subset \mathcal{X}$ be the support of the function. A typical support constraint is the requirement that $c(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \notin \mathcal{X}_S$. Clearly, $\lambda c_1(\boldsymbol{x}) + (1-\lambda)c_2(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in \mathcal{X}_s$ if the same holds for $c_1(\boldsymbol{x})$ and $c_2(\boldsymbol{x})$ individually. Likewise, a support constraint on the Fourier transform of the function $c(\boldsymbol{x})$ is also a convex constraint.

H. Stochastic Complexity and Shrinkage Techniques

If a prior on the image is not known, the regularization approaches described above may not be desired. Alternatives to these approaches include stochastic complexity, wavelet shrinkage techniques, and complexity regularization. The goal is either to define a universal prior, the use of which will achieve asymptotically near-optimal performance for a variety of true priors, or without reference to a prior to derive a simple algorithm that achieves near-optimal performance for a variety of underlying image spaces.

To give specific examples, assume that

$$c = \sum_{\boldsymbol{k} \in \mathcal{L}} c[\boldsymbol{k}] \phi_{\boldsymbol{k}}^\nu \tag{41}$$

where $\{\phi_{\boldsymbol{k}}^\nu, \boldsymbol{k} \in \mathcal{L}\}$ is a set of basis functions such as a Fourier basis or a wavelet basis, and the coefficients $c[\boldsymbol{k}]$ are real-valued. In the image processing community, wavelet bases play an important role because of several empirically observed properties such as sparseness of significant coefficients. Suppose that we have the simple case where the observations are

$$o = c + w \tag{42}$$

where o, c, and w are real-valued, so that the coefficients of o are given by

$$o[\boldsymbol{k}] = c[\boldsymbol{k}] + w[\boldsymbol{k}]. \tag{43}$$

The specific problem addressed by these methods is to estimate the coefficients $c[\boldsymbol{k}]$. The Bayesian shrinkage, minimum description length, and complexity regularization techniques may be extended to the general case by simply adding the equivalent of a log-prior to the objective function.

Definition: A function $\hat{c} : \boldsymbol{R} \to \boldsymbol{R}$ is a called shrinkage function if

$$|\hat{c}(x)| \leq |x| \quad \text{and} \quad x\hat{c}(x) \geq 0. \tag{44}$$

Examples of shrinkage functions are the soft threshold

$$\hat{c}(x) = \min(0, x - \lambda \operatorname{sgn}(x)) \tag{45}$$

and the hard threshold

$$\hat{c}(x) = \begin{cases} 0, & |x| < \lambda \\ x, & |x| \geq \lambda. \end{cases} \tag{46}$$

These simple shrinkage functions have been shown, with proper choice of the thresholds, to provide good asymptotic performance [28]. The threshold $\sigma\sqrt{2 \ln N}$, where N is the number of terms in the basis expansion and σ is the standard deviation of the additive noise, is called the universal threshold. Minimax methods may be used to determine the threshold [28].

Bayesian methods within this category define a prior on c that models the coefficients $c[\boldsymbol{k}]$ as independent and identically

distributed, with common density function $p(\cdot)$. Model the noise samples $w[\boldsymbol{k}]$ as independent and identically distributed Gaussian random variables with zero mean and variance σ^2. The optimal estimate under this model has the form

$$\begin{aligned}\hat{c}(x) &= \arg\max_c p(x \mid c)p(c) \\ &= \arg\min_c \frac{1}{2\sigma^2}(x-c)^2 - \ln p(c) \end{aligned} \tag{47}$$

which reflects the Gaussian assumption on $w[\boldsymbol{k}]$. The resulting estimated function has coefficients

$$c[\boldsymbol{k}] = \hat{c}(o[\boldsymbol{k}]). \tag{48}$$

If $p(c)$ is a Gaussian density with zero mean and variance σ_c^2, then the result is

$$\hat{c}(x) = \frac{\sigma_c^2}{\sigma^2 + \sigma_c^2} x \tag{49}$$

which is the standard Wiener filter. If $p(c)$ is a Laplacian density

$$p(c) = \frac{1}{\sigma_c\sqrt{2}} e^{\frac{-|c|\sqrt{2}}{\sigma_c}} \tag{50}$$

then the soft threshold is the optimal estimator

$$\hat{c}(x) = \begin{cases} x - \lambda \operatorname{sgn}(x), & |x| > \lambda \\ 0, & |x| \le \lambda \end{cases} \tag{51}$$

where $\lambda = \sqrt{2}\frac{\sigma^2}{\sigma_c}$. The universal threshold $\sigma\sqrt{2\ln N}$ corresponds to $\sigma_c = \sigma(\ln N)^{-1/2}$ [66].

Moulin and Liu [66] study the more general set of general Gaussian distributions whose log-priors are proportional to $-|c|^\alpha$, for $0 < \alpha \le 2$. They note that the resulting estimator has a threshold whenever the derivative of the log-prior is not continuous at $c = 0$; this is the case for $0 < \alpha \le 1$. The estimator converges to the hard threshold shrinkage function as α tends to zero.

Physics-based models provide prior information that can affect the process of image formation, and also other signal processing tasks such as detection, estimation, classification, and compression. The notion of *minimum description length*, or *Rissanen length*, plays a fundamental role in this study. The Rissanen-length estimation criterion minimizes the quantity $\log \pi(o|c) + L(c)$ over c, where $L(c)$ is a measure of complexity.

There are several closely related techniques for regularizing estimates using complexity methods and shrinkage estimators. Several of these techniques have nearly optimal performance for a variety of measures, if the true image can be assumed to belong to broad classes of signals such as Besov classes [29], [30].

V. Information-Theoretic Measures

There are various measures that can be used to quantify the performance of an imaging system. The information that is available prior to acquiring data that is to be used for making inference about a scene is quantified by a prior probability distribution on the space of images. Measurements provided by the sensors add information to this prior information. The value of the new information in terms of implications on performance depends on the goal of the imaging system. The goal may be some combination of detection, recognition, parameter estimation, and scene estimation. For these goals, there are associated performance measures including probability of detection, probability of false alarm, probability of correct classification, mean-squared error, and discrepancy between the estimated image and the true image.

Each of the information measures discussed in this section is important, and each has a role in a specific class of problem. One measure is not fundamentally more important than any other. They all share an ability to quantify the information provided by a measurement, and they all depend on the likelihood of the data. In this sense, the likelihood function itself is more fundamental than any single measure of performance. Each measure reduces the likelihood function to a form that is more appropriate to a particular problem. We note, however, that for some imaging situations there is as yet no information or discrepancy measure that is entirely satisfactory. This is especially true when seeking to emulate the performance of human observers of images.

A well-known statement of information theory is the data processing theorem, which says that processing cannot increase information; processing can only refine information by presenting it in a more accessible form. The data processing theorem is an important statement whose validity is based on a formal definition of the term information. In common use, the term information is often used in a casual and imprecise way. There is always a danger of allowing the imprecision in our everyday notion of information to confuse the precision necessary in formal work.

A trivial, though perhaps not obvious, corollary of the data processing theorem is the statement that appending more data to a problem cannot decrease the amount of information and so cannot decrease the performance of an optimal algorithm.

Closely related to the data processing theorem is the concept of a sufficient statistic. The data processing theorem, and notions of uncertainty and entropy lead to the concepts of maximum entropy and minimum discrimination and thereby to the Cramer–Rao bounds, the Fisher information, least squares processing, and the maximum entropy principle.

A. Discrepancy Measures

As defined above, a discrepancy measure $d_{\mathcal{X}}$ on a space $\mathcal{X}$ is a mapping $d_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \boldsymbol{R}_+$ such that $d_{\mathcal{X}}(x_1, x_2) \ge 0$, with equality if and only if $x_1 = x_2$.

Discrepancy measures are assumed to be defined for each of the spaces in the imaging problem. Often it is natural to define discrepancy measures as sqared distances on the spaces. For example, if the space $\mathcal{X}$ has an inner product, $\langle \cdot, \cdot \rangle$, then a natural discrepancy measure is squared error

$$d_{\mathcal{X}}(x_1, x_2) = \|x_1 - x_2\|^2 = \langle x_1 - x_2, x_1 - x_2 \rangle. \tag{52}$$

If the space $\mathcal{X}$ consists of positive-valued functions, then, as discussed below, discrimination is a natural discrepancy measure. In many problems the image spaces are assumed to

be linear spaces equipped with distances and norms in addition to discrepancy measures.

There have been several information-theoretic derivations of discrepancy measures presented in the literature, including [23], [51], [52], and [85]. These lead to characterizations of various discrepancy measures including squared error, discrimination, Ali–Silvey distances [1] or f-divergences [20], Bregman distances [7], and the Itakura–Saito distance [50]. From the axiomatic derivation of Csiszár [23], the discrimination for positive-valued functions and squared error for real-valued (and complex-valued) functions play unique roles in the analysis.

The discrimination function was introduced by Kullback [55], [67], under the name *information for discrimination*. Kullback took the view that the discrimination is an information measure that is more fundamental in some sense than the entropy.

1) Axiomatic Formulation Shannon [83] gave a reasonable set of axioms that a measure of information should satisfy. Shannon's approach leads to the logarithm as a measure of information. Csiszár [23], in the tradition of this approach to the entropy function and the mutual information, gives an axiomatic development for selecting discrepancy functions. Suppose a solution $\boldsymbol{c}$ to the matrix vector equation $\boldsymbol{Hc} = \boldsymbol{a}$ is sought. Starting with a set of reasonable axioms that a measure of discrepancy should satisfy, Csiszár concludes that if the elements of $\boldsymbol{H}$, $\boldsymbol{c}$, and $\boldsymbol{a}$ are required to be real-valued and are otherwise arbitrary, then the only function consistent with his axioms is the squared error $\|\boldsymbol{Hc} - \boldsymbol{a}\|^2$. It is well known that the choice of $\boldsymbol{c}$ that minimizes the squared error is then $\hat{\boldsymbol{c}} = (\boldsymbol{H}^T\boldsymbol{H})^{\#}\boldsymbol{H}^T\boldsymbol{a}$, where the notation $\boldsymbol{M}^{\#}$ denotes the pseudoinverse of $\boldsymbol{M}$. On the other hand, if all entries in $\boldsymbol{H}$, $\boldsymbol{c}$, and $\boldsymbol{a}$ are required to be both real and nonnegative, as is often the case for inverse problems in imaging, then the only discrepancy function consistent with Csiszár's axioms is the discrimination

$$I\left(\sum_x h(\cdot,x)c(x)\|a(\cdot)\right) = \sum_y a(y)\ln\left[\frac{a(y)}{\sum_x h(y,x)c(x)}\right] - \sum_y\left[a(y) - \sum_x h(y,x)c(x)\right]. \tag{53}$$

Explicit analytical expressions for the $c = \hat{c}$ minimizing discrimination are difficult to obtain, and so numerical methods are appropriate.

2) A Discrepancy Inequality For any convex set $\mathcal{A} \subset \boldsymbol{R}^n$, both the discrimination

$$d(y,x) = I(y\|x) = \sum_i\left[y_i\log\frac{y_i}{x_i} - y_i + x_i\right] \tag{54}$$

and the squared euclidean distance $d(y,x) = \|y-x\|^2$ satisfy the inequality

$$d(y,x) \geq d(y,y^*) + d(y^*,x), \qquad \forall x,y \in \mathcal{A} \tag{55}$$

where

$$y^* = \arg\min_{y\in\mathcal{A}} d(y,x). \tag{56}$$

If $\mathcal{A}$ is further restricted to be an affine subspace, then

$$d(y,x) = d(y,y^*) + d(y^*,x), \qquad \forall x,y \in \mathcal{A}. \tag{57}$$

The proof of this statement for discrimination is based on the following argument [18]. Because y^* achieves the minimum, $d(y,x) \geq d(y^*,x)$ for all $y \in \mathcal{A}$; so for all $\epsilon > 0$, and $y^* + \epsilon\Delta y \in \mathcal{A}$

$$\frac{1}{\epsilon}[d(y^* + \epsilon\Delta y, x) - d(y^*,x)] \geq 0. \tag{58}$$

Take the limit as ϵ goes to zero to obtain

$$\nabla_y d(y^*,x)\Delta y \geq 0. \tag{59}$$

Now let $\Delta y = y - y^*$, and rearrange this to get

$$\begin{aligned}\sum (y_i - y_i^*)\left(\log\frac{y_i^*}{x_i}\right) &= \sum_i\left[y_i\log\frac{y_i^*}{y_i} + y_i - y_i^* + y_i\log\frac{y_i}{x_i}\right.\\ &\quad \left. - y_i + x_i - y_i^*\log\frac{y_i^*}{x_i} + y_i^* - x_i\right]\\ &= -d(y,y^*) + d(y,x) - d(y^*,x) \geq 0.\end{aligned} \tag{60}$$

In the case that $\mathcal{A}$ is an affine subspace, both Δy and $-\Delta y$ are allowable directions, yielding equality in (58).

The inequality (55) (or the equality (57) when $\mathcal{A}$ is an affine subspace) can be viewed as a statement of the tradeoff between approximation error and estimation error. The term $d(y^*,x)$ is a measure of the approximation error because it is a discrepancy between the closest element in the convex set $\mathcal{A}$ and x. The term $d(y,y^*)$ is a measure of the estimation error since it is a discrepancy between the estimated value y and the closest element in the convex set $\mathcal{A}$. The inequality says that the discrepancy between the estimate and x is bounded below by the sum of these two terms.

For the squared-error discrepancy $d(y,x) = \|y-x\|^2$, the same argument as above holds, until (59) becomes $2\langle y - y^*, y^* - x\rangle \geq 0$, which immediately implies

$$\begin{aligned}\|y-x\|^2 &= \|y-y^*\|^2 + \|y^*-x\|^2 + 2\langle y-y^*, y^*-x\rangle\\ &\geq \|y-y^*\|^2 + \|y^*-x\|^2.\end{aligned} \tag{61}$$

The interpretation of this inequality is the same as interpretation of the discrimination criterion.

B. Mutual Information

The uncertainty associated with a random vector θ is quantified by its differential entropy (or, if θ takes on only discrete values by its entropy). Denote this differential entropy by $h(\theta)$

$$h(\theta) = -E\{\log p(\theta)\}. \tag{62}$$

The observations may be viewed as decreasing the uncertainty in the underlying parameters. In this view, the mutual information between the parameters and the observation quantifies the

decrease in uncertainty obtained by making an observation, because

$$I(o;\theta) = h(\theta) - h(\theta|o) \tag{63}$$

where

$$h(\theta|o) = -E\{\log p(\theta|o)\}. \tag{64}$$

This view is helpful in coding applications. To be more precise, the rate-distortion curves for θ with the observation o lies below the rate-distortion curve for θ in the absence of observations. For any fixed distortion, the difference in the two curves is never greater than $I(o;\theta)$.

C. Fisher Information

For parameter estimation, a standard performance measure, which will be discussed in Section VI-A, is the Cramer–Rao bound and its extensions. The information provided by sensor measurements can be quantified in terms of the reduction in the variance of a parameter estimate due to the measurement. This is equivalent to looking at the increase in the Fisher information, as outlined below.

Suppose that θ, the parameters to be estimated, take values in $\boldsymbol{R}^n$. Let the prior probability density function on θ be $p(\theta)$ and let the conditional distribution for the sensor data be $\pi(o \mid c(\theta))$. The posterior density on θ is proportional to $\pi(o \mid c(\theta))p(\theta)$; denote it by $p(\theta \mid o)$. Prior to making any measurements, the Fisher information matrix is

$$\boldsymbol{J}_\theta = E\left\{ \frac{\partial \ln p(\theta)}{\partial \theta} \frac{\partial \ln p(\theta)}{\partial \theta}^T \right\} \tag{65}$$

where the partial derivatives are assumed to exist and yield column vectors, and T denotes transpose. After an observation, the Fisher information matrix is

$$\boldsymbol{J}_{\theta|o} = \boldsymbol{J}_{o|\theta} + \boldsymbol{J}_\theta \tag{66}$$

where $\boldsymbol{J}_\theta$ is given above and

$$\boldsymbol{J}_{o|\theta} = E\left\{ \frac{\partial \ln \pi(o \mid c(\theta))}{\partial \theta} \frac{\partial \ln \pi(o \mid c(\theta))}{\partial \theta}^T \right\}. \tag{67}$$

Let $\boldsymbol{R}$ be the mean-squared-error matrix for any specified estimator. Then the matrix $\boldsymbol{R} - \boldsymbol{J}_{\theta|o}^{-1}$ is nonnegative definite.

The matrix $\boldsymbol{J}_{o|\theta}$ quantifies the increase in the Fisher information obtained from the observations. For estimation problems, the relative increase in the Fisher information is one way to quantify the value of observations and the value in making additional observations.

The difference between matrices can be measured in several ways. One way is by examining the increase in the Fisher information, as in (66). Another is to examine the corresponding decrease in the inverse of the Fisher information matrix, which determines the Cramer–Rao bound on estimation (see Section VI-A)

$$\boldsymbol{J}_\theta^{-1} - \boldsymbol{J}_{\theta|o}^{-1} = \left(\boldsymbol{J}_\theta + \boldsymbol{J}_\theta \boldsymbol{J}_{o|\theta}^{-1} \boldsymbol{J}_\theta\right)^{-1}. \tag{68}$$

A third is to measure the decrease in the inverse of the Fisher information relative to the prior

$$\boldsymbol{J}_\theta^{T/2}\left(\boldsymbol{J}_\theta^{-1} - \boldsymbol{J}_{\theta|o}^{-1}\right)\boldsymbol{J}_\theta^{1/2} = \left(\boldsymbol{I} + \boldsymbol{J}_\theta^{T/2} \boldsymbol{J}_{o|\theta}^{-1} \boldsymbol{J}_\theta^{1/2}\right)^{-1}. \tag{69}$$

Here, $\boldsymbol{J}_\theta^{1/2}\boldsymbol{J}_\theta^{T/2} = \boldsymbol{J}_\theta$.

Other bounds for parameter estimations can be examined in a similar way.

There is a link between differential entropy and Fisher information given by de Bruijn's identity [19, pp. 494–495]. Let Z be a Gaussian random variable with mean zero and variance one. Let ϕ equal θ plus a scalar times Z

$$\phi = \theta + \sqrt{t}Z. \tag{70}$$

Then, de Bruijn's identity says that (assuming natural logarithms)

$$2\frac{\partial h(\phi(t))}{\partial t} = J_\theta \tag{71}$$

where $h(\phi(t))$ is the differential entropy of ϕ parameterized by t. In terms of the observation

$$2\frac{\partial h(o \mid \phi(t))}{\partial t} = J_{o|\theta} \tag{72}$$

so

$$2\frac{\partial h(o \mid \phi(t))}{\partial t} + 2\frac{\partial h(\phi(t))}{\partial t} = J_{o|\theta} + J_\theta = J_{\theta|o}. \tag{73}$$

Another interpretation of the Fisher-information matrix is given by Amari [2] in his discussion of the differential geometry of statistical models. Here, a parametric model $\pi(o|c(\theta))$ is interpreted as defining a manifold S in the space of all models $\pi(o)$. Amari defines an inner product between vectors in tangent planes of S as covariances, then arguing that $J_{o|\theta}$ is the metric tensor in the resulting Riemannian space. Amari uses this framework to establish general asymptotic properties of maximum-likelihood estimators of θ, such as asymptotic efficiency, consistency, and normality.

VI. Performance Bounds

A goal of information-theoretic image formation is to bound achievable performance in terms of the information measures. Whenever these measures cannot be evaluated analytically, an important technique is to append information that is actually unknown so that the bounds can be evaluated analytically. The actual performance cannot be better than such a bound. An early use of this technique is in the classical book on communication theory by Wozencraft and Jacobs [104]. They introduce a "genie" in their analysis of the performance of coding systems. In a general form of the argument, the genie is assumed to provide gratuitous side information that embellishes the actual data. The performance without the genie's side information cannot be better than the performance with this extra information. The technique of introducing a genie to embellish the actual data set so that a bound on performance can be computed is quite similar to the technique of Dempster, Laird, and Rubin [27] who introduce a "complete data set" so that calculation of the maximum-likelihood solution becomes analytically tractable.

A. Cramer–Rao Bounds

The simplest problem of estimation theory involves an unknown parameter $\theta \in \boldsymbol{R}^n$ and a random measurement $o \in \mathcal{O}$ from which θ is to be determined. The measurement $o \in \mathcal{O}$ has probability distribution function $\pi(o \mid c(\theta))$ depending on the parameter θ. The unknown parameter θ must be estimated based upon an observation of o. The estimate of θ, given the measurement o, is a function $\hat{\theta}(o)$. This estimate $\hat{\theta}$ is itself a random variable because it is a function of the random measurement o.

The quality of an estimator is often judged by its mean value

$$E[\hat{\theta}] = E[\hat{\theta}(o)] \tag{74}$$

and by its mean-squared-error

$$\boldsymbol{R} = E[(\hat{\theta}(o) - \theta)(\hat{\theta}(o) - \theta)^T]. \tag{75}$$

When θ is not random, an unbiased estimator of θ is any function $\hat{\theta}(o)$ satisfying

$$E[\hat{\theta}] = \theta. \tag{76}$$

For any unbiased estimator, the matrix

$$\boldsymbol{R} - \boldsymbol{J}_{o|\theta}^{-1} \tag{77}$$

is nonnegative definite; this is the Cramer–Rao bound. When θ is random, then

$$\boldsymbol{R} - [\boldsymbol{J}_{o|\theta} + \boldsymbol{J}_{\theta}]^{-1} \tag{78}$$

is nonnegative definite.

Other bounds on the variance of an estimate can sometimes be tighter, including the Ziv–Zakai bound, the Barankin bound, and the Bhattacharyya bound. The Cramer–Rao bound when the parameters to be estimated are constrained to lie in a nonopen subset of $\boldsymbol{R}^n$ is developed by Gorman and Hero [41]; this bound is useful in imaging problems, for example, when the intensity has a known support or is smooth. For imaging problems, the Fisher information matrix can be, and usually is, too large to invert practically so that computing the Cramer–Rao bound on the error covariance in estimating all of the parameters that define the image is infeasible. To address this problem, Hero and Fessler [47] and Hero, Usman, Sauve, and Fessler [48] have developed a recursive procedure for computing submatrices of the inverse of the Fisher matrix; this can be especially useful for establishing Cramer–Rao bounds on subsets of the parameters (corresponding to a region in an image) that are of particular interest.

B. Bounds on Groups

If the parameters in the problem are not real-valued, then bounds other than the Cramer–Rao bound may be appropriate. A measure analogous to squared-error must be defined on the space, and bounds on errors in terms of the mean of this measure found. One approach detailed in this section is valid for group-valued parameters.

If a scene has objects of interest that are rigid bodies, the group actions consist of translation and rotation. When restricted to the plane, the group is $SO(2)$ which is isomorphic to the circle. Rotations in three dimensions take values in $SO(3)$. In either case, the group of rotations is compact, so there is a maximum distance between any two elements of the group. For small errors, as are typically encountered in high signal-to-noise-ratio problems, expansion in a local coordinate system followed by standard Cramer–Rao analysis in those coordinates is appropriate. When the estimation errors are not local, however, the curvature of the parameter space becomes important, and this local analysis does not apply. If this curvature is ignored, then it is possible to get so-called lower bounds that get arbitrarily large as a parameter (typically, signal-to-noise ratio) gets small. But this is impossible because the largest error possible on a compact set is bounded.

One approach that avoids this difficulty has been proposed by Grenander, Miller, and Srivastava [44]. It is explained here within the context of rotation groups, but can be extended to other groups.

Elements of $SO(n)$ are mapped to their $n \times n$ matrix group representative so that matrix multiplication is equivalent to the group action. For $SO(2)$, the matrices are of the form

$$\boldsymbol{O}(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{79}$$

where θ is the one-dimensional parameter of the group. Any norm on $n \times n$ matrices induces a norm on the group. Define the Hilbert–Schmidt norm by

$$\|A\|_{\mathrm{HS}}^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}^2. \tag{80}$$

The squared distance between two elements of the group equals the distance between their matrix representatives

$$d_{\mathrm{HS}}(\theta_1, \theta_2) = \|\boldsymbol{O}(\theta_1) - \boldsymbol{O}(\theta_2)\|_{\mathrm{HS}}^2. \tag{81}$$

This is referred to as the Hilbert–Schmidt distance squared, and it is the natural extension of squared error in $\boldsymbol{R}^n$ to $SO(n)$. Note that

$$\begin{aligned} \|\boldsymbol{O}(\theta_1) - \boldsymbol{O}(\theta_2)\|_{\mathrm{HS}}^2 &= \mathrm{Tr}\,[(\boldsymbol{O}(\theta_1) - \boldsymbol{O}(\theta_2))(\boldsymbol{O}(\theta_1) - \boldsymbol{O}(\theta_2))^T] \\ &= 2n - 2\,\mathrm{Tr}\,[\boldsymbol{O}(\theta_1)\boldsymbol{O}(\theta_2)^T]. \end{aligned} \tag{82}$$

Group-valued estimators may be evaluated in terms of this Hilbert–Schmidt distance squared. The Hilbert–Schmidt estimator is the minimum expected Hilbert–Schmidt distance-squared estimator (the extension of the minimum mean-squared-error estimator to the special orthogonal group $SO(n)$)

$$\hat{\theta}_{\mathrm{HS}} = \arg\min_{\tilde{\theta}} E[d_{\mathrm{HS}}(\theta, \tilde{\theta}) \mid o]. \tag{83}$$

That is, $\hat{\theta}$ is the orientation that minimizes the expected Hilbert–Schmidt squared error given the observations. Note that the estimator must be defined in this way because mean values are not defined on the group. There is only one operation available to combine two elements of the group; this operation is not necessarily addition.

There is a straightforward algorithm to compute the Hilbert–Schmidt estimator if the posterior is known. From (82), the optimal estimate is the one that maximizes

$E\{\mathrm{Tr}\,[\boldsymbol{O}(\theta)\boldsymbol{O}(\tilde{\theta})^T]|o\}$. First, compute $\boldsymbol{A} = E\{\boldsymbol{O}(\theta)|o\}$, where the expectation is well-defined because this is a linear combination of $n \times n$ matrices. Next, find the singular value decomposition of $\boldsymbol{A}$ as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T. \tag{84}$$

Assume that $\boldsymbol{\Sigma}$ is ordered so that the smallest eigenvalue is in the lower right corner. Then let $\boldsymbol{D}$ be a diagonal matrix whose diagonal entries are all one, except possibly for the lower right corner. The estimate is

$$\hat{\boldsymbol{O}}_{\mathrm{HS}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \tag{85}$$

and the lower right entry of $\boldsymbol{D}$ is chosen to ensure that the determinant of $\hat{\boldsymbol{O}}_{\mathrm{HS}}$ is equal to one. Finally, $\hat{\theta}_{\mathrm{HS}}$ is the group element corresponding to $\hat{\boldsymbol{O}}_{\mathrm{HS}}$. The performance of any group-valued estimator is bounded as follows [44].

Theorem: Let $\hat{\theta}(o) \in SO(n)$ be any estimator. Then

$$E[d_{\mathrm{HS}}(\theta, \hat{\theta}(o))] \geq E[d_{\mathrm{HS}}(\theta, \hat{\theta}_{\mathrm{HS}}(o))]. \tag{86}$$

It is interesting to note that for small variations, the Hilbert–Schmidt squared distance is essentially the same as would be obtained using a linearization of the space. To see this for $n = 2$, note that (82) becomes $4 - 4\cos(\theta_1 - \theta_2)$. For small differences $\theta = \theta_1 - \theta_2$, $\cos\theta \approx 1 - \theta^2/2$, and

$$4 - 4\cos(\theta_1 - \theta_2) \approx 2(\theta_1 - \theta_2)^2. \tag{87}$$

This is twice the squared error between the angles, so for small errors this is equivalent to linearizing the space and using squared error in $\boldsymbol{R}$.

C. Resolution

The resolution achieved in image formation is an important attribute that is often cited. However, a universally acceptable definition of resolution as a performance measure is elusive. The Rayleigh criterion is often used to quantify the resolution of optical images. The width of the main lobe of the point-spread function of an imaging system is another frequently used measure of resolution. However, these measures are typically applied to image data rather than to post-processed data. Model-based processing can result in sharper image detail and, hence, improved resolution; good examples of this are the images that result from processing image data acquired in the presence of spherical aberration in the Hubble space telescope. The resolution achieved with image restoration and image estimation is more difficult to quantify. One approach is through computer simulation in which a known object is synthetically imaged and the resulting image processed for restoration. The restored imaged can be correlated against a test image formed by convolving the known object with a given point-spread function and then adjusting a "width" parameter of this function to achieve maximum correlation. An example is a circular Gaussian point spread in which the width (or spread) parameter is adjusted for maximum correlation, then used as a measure of resolution [11], [73]. The benefit of such an approach in practice depends on how well the synthetic image data matches data actually produced by the imaging system of interest.

D. Information Rate Functions

If the goal is object detection (binary hypothesis testing), then the performance may be quantified by error rates. For example, the Chernoff information determines the rate of the minimum probability of error detector, and hence can be used as a measure of information contained in a measurement. Let π_1 and π_0 be the probability distributions on the sensor data under hypotheses H_1 and H_0, respectively. Let the log-moment-generating function for the loglikelihood ratio be denoted by $\phi(s)$

$$\phi(s) = \log E_1\left\{\exp\left(s\log\frac{d\pi_1}{d\pi_0}\right)\right\} \tag{88}$$

where E_1 denotes expectation with respect to π_1. The information rate function for the problem is given by [9], [31]

$$I(x) = \sup_{0 \geq s \geq -1}[sx - \phi(s)]. \tag{89}$$

Then the Chernoff information equals $I(0)$ [99, p. 123].

Similarly, Stein's lemma says that fixing the probability of one type of error and minimizing the probability of the other type yields the relative entropy between the distributions under H_0 and H_1 as the measure of the information provided by a measurement. Put another way, let

$$x_1 = E_1\left\{\log\frac{d\pi_1}{d\pi_0}\right\} = D(\pi_1 \mid \pi_0) \tag{90}$$

and

$$x_0 = E_0\left\{\log\frac{d\pi_0}{d\pi_1}\right\} = D(\pi_0 \mid \pi_1) \tag{91}$$

then [19, pp. 309–311]

$$I(-x_0) = x_0. \tag{92}$$

Because the Chernoff information and the rate in Stein's lemma are just samples of the information rate function $I(x)$, for $x_0 \leq x \leq x_1$, the rate function may be the proper measure of information provided by the sensor for detection problems.

The purpose of an imaging system may be to recognize, detect, or locate an object within the image; the image itself may be only of passing interest. In such a case, the performance of the imaging system is measured by the performance of the recognition or detection function. The overall performance of a system that is designed to recognize objects within an image is ultimately determined by the performance of the final decision that declares an object present. The detection algorithm may be constructed in stages where the output of one stage is fed to the next, or the detection algorithm may be designed in terms of a single, unified optimization problem. To predict the performance of an object recognition algorithm, we ask how the system would operate if the receiver knew everything except the decision.

Assume that two scenes are completely specified and that the available data are Poisson-distributed with means $T\boldsymbol{\Lambda}_1$ and $T\boldsymbol{\Lambda}_0$, where T is a measure of the signal-to-noise ratio such as the integration time. Then the log-likelihood function for the data $\boldsymbol{y}$ may be written

$$l(\boldsymbol{y}) = T[I(\boldsymbol{y}/T, \boldsymbol{\Lambda}_0) - I(\boldsymbol{y}/T, \boldsymbol{\Lambda}_1)] \tag{93}$$

where

$$I(x,\boldsymbol{\xi}) = \sum x_i \log \frac{x_i}{\xi_i} - x_i + \xi_i$$

is discrimination. From Stein's lemma, for a fixed probability of false alarm, the probability of detection converges to one exponentially fast in T with exponent $-TI(\boldsymbol{\Lambda}_0, \boldsymbol{\Lambda}_1)$. Thus discrimination predicts asymptotic rates.

The rate depends on the clutter. If $\boldsymbol{\Lambda}_0$ has only clutter and $\boldsymbol{\Lambda}_1$ has a target and clutter, then this rate is a measure of the clutter complexity in the sense that in quantifies the clutter's ability to reduce the detection rate. For systems with small point-spread functions, the discrimination between images without and with a target is a *local* measure. This measure can be used to gain a confidence measure on the output of an object recognition system. Suboptimal algorithms may be compared on the basis of their rates. That is, a fixed algorithm will exhibit a performance that varies with signal-to-noise ratio. For large signal-to-noise ratio, the rate is all that matters.

VII. Image Formation

We call the process of forming images from data acquired with a sensor *imagation* or *image formation*. In our use of this term, imagation includes: image reconstruction (the common term used for building a tomographic image from projection data), image restoration (a term used for correcting image data that are marred by camera defects or motion), image estimation (used for forming images from data that are stochastic), and image formation from data that indirectly depend on an image, such as in synthetic-aperture radar.

The type of models used for describing the image and data spaces, as well as the discrepancy measures that are adopted for assessing performance, influence the approaches used for imagation. Deterministic models and the use of least squares, discrimination, and maximum-entropy discrepancy-measures with and without constraints lead to one set of approaches, while stochastic models and the use of likelihood discrepancy-measures with and without priors and constraints leads to another.

A. Maximum Likelihood

The maximum-likelihood method is a long-standing method for estimating unknown, deterministic parameters that influence a set of stochastic data. The maximum-likelihood principle is a general principle of data reduction in which when reducing a set of data $\boldsymbol{x}$ described by a log-likelihood function $\Lambda(\gamma) = \log p(\boldsymbol{x} \mid \gamma)$, one chooses a γ that maximizes the log-likelihood function

$$\hat{\gamma} = \arg\max_{\gamma} \Lambda(\gamma). \tag{94}$$

A maximum-likelihood estimate has the desired properties that it is asymptotically unbiased and efficient.

Its use for imagation follows the usual prescription of formulating a model for the data acquired with an imaging system, with this model being in the form of a probability distribution $\pi(o:c)$ that is a functional of the image c; $\pi(o:c)$ is called the *likelihood* or *data likelihood* in this context. A maximum-likelihood estimate of the image $\hat{c}$ is an image c that maximizes the log-likelihood functional

$$\hat{c} = \arg\max_{c \in \mathcal{C}} \log \pi(o:c). \tag{95}$$

For an image restricted to be a function of a parameter vector θ, the image estimate is $\hat{c} = c(\hat{\theta})$, where

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \log \pi(o : c(\theta)). \tag{96}$$

For example, the image could be of a known object whose position and orientation are unknown. In this instance, the likelihood can be regarded as a function of the parameters, which would be estimated by maximizing the likelihood in the usual manner.

If the image is regarded simply as an unknown function, then the problem is often ill-posed, and some regularization is required. One approach is to discretize the image, treating it as piecewise constant over pixels or by representing it as a linear combination of orthonormal functions as discussed in Section IV-B. This in effect converts the imagation problem into a parameter-estimation problem, and the maximum-likelihood method can in principle be used straightforwardly to estimate the parameterized image. For example, if the data source is modeled as a spatial Poisson process $\{N(A), A \in \sigma(\boldsymbol{R}^2)\}$ with an intensity function $\{c(x), x \in \boldsymbol{R}^2\}$, the log-likelihood functional is

$$\ln \pi(N:c) = -\int_{\boldsymbol{R}^2} c(x)\,dx + \int_{\boldsymbol{R}^2} \ln c(x) N(dx). \tag{97}$$

This likelihood is unbounded over the space of nonnegative functions, so a maximum-likelihood estimate does not exist. Overcoming this difficulty requires the use of some form of regularization as discussed in Section IV-A, which can be in the form of imposing a discretization, imposing a prior distribution on image values, imposing a penalty functional that restricts the roughness of the estimated image values, or using Grenander's sieves [42] to restrict maximizers to a subset of the nonnegative functions.

B. Maximum a Posteriori

Maximum *a posteriori* probability (MAP) estimation is also a long-standing method of estimating parameters from observed data; it is used when the parameters to be estimated are random and have a known prior probability distribution $p(\theta)$. If the data likelihood is $\pi(r|\theta)$ for some given data r, then a MAP estimate $\hat{\theta}$ of θ is a maximizer of the posterior distribution $p(\theta|r)$. Because this conditional distribution is proportional to the product $\pi(r|\theta)p(\theta)$ of the data likelihood and the prior, this procedure is analogous to maximum-likelihood (ML) estimation of the parameters but with the likelihood scaled by the prior. MAP imagation is similar to ML imagation with a prior distribution on image values included in the functional being maximized.

C. Maximum Entropy and Minimum Discrimination

The *Jaynes maximum-entropy principle* is a principle of data reduction that says that when reducing a set of data into the

form of an underlying model, one should be maximally noncommittal with respect to missing data. If one must estimate a probability distribution $\boldsymbol{q}$ on the data source satisfying certain known constraints on $\boldsymbol{q}$, such as

$$\sum_k q_k f_k = t \tag{98}$$

then, of those distributions that are consistent with the constraints, one should choose as the estimate of $\boldsymbol{q}$ the probability distribution $\hat{\boldsymbol{q}}$ that has maximum entropy. A nice example can be given for a probabilistic source with a real output. Suppose the source produces a real-valued random variable X whose mean and variance are known, and otherwise the probability distribution governing the source is unknown. Then the maximum-entropy principle says that one should estimate that the probability density $q(x)$ is a Gaussian probability density with the given mean and variance. This is a consequence of the well-known fact that a Gaussian random variable has the largest differential entropy of any random variable of a given mean and variance.

The maximum-entropy and maximum-likelihood principles are equivalent when the constraint to be enforced when estimating a probability distribution is not in the form of some given moments but, rather, of some given data. When given some statistical data from which a distribution or image is to be estimated, one approach is to use those data to estimate some moments and then to use these estimated moments as if they were the exact (deterministic) moments when maximizing entropy. However, the estimated moments are exact only in the limit of a large data set and otherwise are random, resulting in a conceptual inconsistency. As discussed by Miller and Snyder [62], when entropy is maximized subject only to the constraint of some given, statistical data rather than deterministic moments, the resulting maximum-entropy estimates are also maximum-likelihood estimates.

The *Kullback minimum-discrimination principle* is an alternative principle that applies when one is given both a probability distribution $\boldsymbol{p}$ as a prior estimate of $\boldsymbol{q}$ and also a set of constraints, such as moment constraints, that the probability distribution $\boldsymbol{q}$ must satisfy. Under this principle, the optimal $\boldsymbol{p}$ is

$$\hat{\boldsymbol{p}} = \arg\min_{\boldsymbol{p}\in\mathcal{P}} I(\boldsymbol{p}||\boldsymbol{q}) \tag{99}$$

where $\mathcal{P}$ is the set of probability distributions that satisfy the moment constraints [57], [56], and $I(\cdot\|\cdot)$ is discrimination. If the prior estimate $\boldsymbol{q}$ is a uniform distribution, then this principle yields the maximum-entropy distribution subject to the moment constraints.

D. Minimum Discrepancy: Least Squares and Discrimination

For observations in $\mathcal{O}$, assume the information-theoretic discrepancy measure $d_{\mathcal{O}}$. In deterministic problems, there is some model for the observed data o in terms of the underlying image c. Let this model be a function $f : \mathcal{C} \to \mathcal{O}$. Then the minimum-discrepancy problem is to find the $c \in \mathcal{C}$ that minimizes the predicted discrepancy from the observation

$$\hat{c}(o) = \arg\min_{c\in\mathcal{C}} d_{\mathcal{O}}(f(c), o). \tag{100}$$

This formalism includes least-squares, minimum-discrimination, and related methods.

If $\mathcal{O}$ consists of real- or complex-valued functions, then the squared error is the natural discrepancy measure. Let $\langle\cdot,\cdot\rangle$ denote an inner product on $\mathcal{O}$. The discrepancy measure is then

$$d_{\mathcal{O}}(o_1, o_2) = \langle o_1 - o_2, o_1 - o_2\rangle. \tag{101}$$

The least-squares problem is

$$\hat{c}(o) = \arg\min_{c\in\mathcal{C}} \langle f(c) - o, f(c) - o\rangle. \tag{102}$$

If $\mathcal{O}$ consists of positive-valued functions, then discrimination is the natural discrepancy measure. Assume $o : \boldsymbol{R}^n \to \boldsymbol{R}_+$, and let $o(\boldsymbol{y})$ be the value of the observation at the point $\boldsymbol{y} \in \boldsymbol{R}^n$. The discrimination is defined as

$$\begin{aligned} d_{\mathcal{O}}(o_1, o_2) &= \int \left[o_1(\boldsymbol{y}) \ln \frac{o_1(\boldsymbol{y})}{o_2(\boldsymbol{y})} - o_1(\boldsymbol{y}) + o_2(\boldsymbol{y}) \right] d\boldsymbol{y} \\ &= I(o_1||o_2). \end{aligned} \tag{103}$$

In applications, the observations are often vectors rather than functions, in which case (103) is written as a summation as in (54) rather than an integral. For either the integral or the summation form, the optimization problem as stated in (100) becomes

$$\hat{c}(o) = \arg\min_{c\in\mathcal{C}} I(f(c)||o). \tag{104}$$

For linear inverse problems, this formulation leads to the generalized iterative scaling or SMART (simultaneous multiplicative algebraic reconstruction technique) algorithm [12], [25].

The discrimination may be used with the arguments reversed. In this second formulation, the minimum discrepancy statement in (100) becomes

$$\hat{c}(o) = \arg\min_{c\in\mathcal{C}} I(o||f(c)). \tag{105}$$

Since discrimination is not symmetric in its arguments, the criteria (104) and (105) can have different solutions and they lead to very different algorithms.

VIII. Computational Algorithms

Information-theoretic image formation yields images by the optimization of performance metrics. Analytical intractability usually accompanies any attempt to form images in this way, so numerical algorithms must be used; many algorithms used have a strong infomration-theoretic motivation. An important exception can occur with some linear problems having Gaussian statistics and quadratic metrics, but even in these cases numerical methods for performing matrix inversions or solving integral equations are often needed. For example, the original method by Rockmore and Macovski [76] for forming maximum-likelihood images for emission tomography was realized practically only when Shepp and Vardi [84] later

introduced the expectation-maximization method of Dempster, Laird, and Rubin [27] as a means of constructing algorithms for computing the maximum-likelihood image. A wide variety of computational algorithms are in use for producing images numerically. Standard methods of numerical optimization, such as gradient descent, are widely used. In this section, we shall review methods based on information-theoretic concepts that have become popular and are presently finding their way into practical imaging systems. This includes the expectation-maximization method and its recent extensions introduced by Fessler and Hero [32]. Also mentioned is a similar method introduced by Snyder, Schulz, and O'Sullivan [92] for deterministic problems in which the discrepancy metric is the discrimination in the form of (105). A related algorithm originally proposed by Darroch and Ratcliff [25] for computing the distribution that maximizes entropy subject to constraints has been used to solve linear inverse problems by Byrne [12] and others. This is the generalized iterative scaling or SMART algorithm for deterministic problems in which the discrepancy metric is the discrimination in the form of (104). Stochastic search by means of the jump-diffusion algorithm of Grenander and Miller [43] is another powerful tool which we shall review briefly.

There are many other algorithms with an information-theoretic motivation that are not discussed here. Simulated annealing for imaging problems was discussed by Geman and Geman [34]; the jump-diffusion algorithm is a stochastic search without the annealing process. In addition to these stochastic search methods to combat the multimodality of optimization criteria, there are deterministic methods including the graduated nonconvexity algorithm [5] and its improvements [68]. The SMART algorithm is a descendent of ART and MART, algorithms that have been used in image reconstruction algorithms for many years (see the references in [12]). There have been many methods proposed for increasing the convergence rate of the expectation–maximization (EM) algorithm. Among the more promising techniques in the literature are those based on partitioning of the data space such as the ordered-subset EM algorithm and its variants [8], [49].

A. The Expectation–Maximization Method

The expectation–maximization method of Dempster, Laird, and Rubin [27] is a general approach for formulating recursive algorithms that can be used to determine the maximum-likelihood estimate of a parameter vector θ in terms of some measured data r. To apply the method requires judgment, and the structure of the problem must be appropriate. When successfully applied to a particular problem, the EM method yields a particular algorithm that is specific to that problem. Indeed, the EM method can even yield more than one algorithm for the same problem because there can be more than one way to apply the method. For imaging problems, θ is composed of the unknown parameters, such as pixel values, that comprise the image to be estimated, and r is composed of the data values produced by the imaging system. If some prior information regarding θ is available, this can readily be incorporated into the method by either adding the logarithm of the prior on θ to the log-likelihood function being maximized, thereby producing MAP estimates of the parameters, or by adding a penalty function during the maximization.

The EM method begins by selecting some hypothetical data, r_{cd}, called the "complete data." There is considerable flexibility is making this selection, and making a good choice has largely been based on experience drawn from a familiarity with the physical problem at hand and its mathematical model. The choice can influence the behavior of the recursive algorithm that results, such as its rate of convergence, so careful consideration is warranted. Roughly speaking, the choice should be such that there is a function $h(\cdot)$ such that the actual data r, here termed the "incomplete data," can be recovered from the complete data, $r = h(r_{cd})$, and such that the log-likelihood function $L_{cd}(\theta)$ of the complete data can be formulated and the required analytical steps can be accomplished. At the very least, the conditional likelihood of the incomplete data given the complete data must be independent of the parameters θ. The issue of selecting complete data is discussed further in the next section.

The recursion proceeds as follows. Suppose that $\hat{\theta}^{\text{old}}$ is an estimate of θ that has been formed at some stage of the recursion. To get to the next stage, an E-step and an M-step must be performed. The E-step consists of evaluating the conditional expectation of the complete-data log-likelihood given the incomplete data and the parameter estimate available at that stage resulting in a function $Q(\theta|\hat{\theta}^{\text{old}})$, defined by

$$Q(\theta|\hat{\theta}^{\text{old}}) = E[L_{cd}(\theta)|r, \hat{\theta}^{\text{old}}].$$

The M-step is then performed to obtain an updated and possibly improved parameter estimate, $\hat{\theta}^{\text{new}}$ according to

$$\hat{\theta}^{\text{new}} = \arg\max_{\theta} Q(\theta|\hat{\theta}^{\text{old}}).$$

In many interesting situations, this maximization can be performed analytically, but in others numerical optimization is required, resulting in an iteration nested within the EM recursions.

Following Dempster, Laird, and Rubin [27], it is straightforward to demonstrate that the recursion produces a nondecreasing sequence

$$L_{id}(\hat{\theta}^{(0)}) \leq L_{id}(\hat{\theta}^{(1)}) \leq \cdots \leq L_{id}(\hat{\theta}^{(k)}) \leq \cdots$$

of log-likelihoods $L_{id}(\theta)$, of the incomplete data r. The limit, $\hat{\theta}^{(\infty)}$, if it exists, satisfies the necessary conditions for a maximizer of $L_{id}(\theta)$. To see this, let $p_{id}(r : \theta)$ and $p_{cd}(r_{cd} : \theta)$ denote the likelihood functions for the incomplete and complete data, respectively. These are related according to

$$p_{id}(r : \theta) = \int p(r|r_{cd}) p_{cd}(r_{cd} : \theta)\, dr_{cd}. \tag{106}$$

Also, define the conditional likelihood $p(r_{cd}|r : \theta)$ according to Bayes rule

$$p(r_{cd} \mid r : \theta) = \frac{p(r \mid r_{cd}) p_{cd}(r_{cd} : \theta)}{p_{id}(r : \theta)}. \tag{107}$$

Then

$$L_{id}(\theta) = L_{cd}(\theta) - \log p\,(r_{cd} \mid r : \theta) + \log p\,(r \mid r_{cd}) \tag{108}$$

where

$$L_{id}(\theta) = \log p_{id}(r : \theta)$$

and

$$L_{cd}(\theta) = \log p_{cd}(r_{rd} : \theta).$$

Multiplying both sides of this equation by $p(r_{cd}|r : \theta')$ and integrating over r_{cd} then yields

$$L_{id}(\theta) = Q(\theta \mid \theta') - \int p(r_{cd} \mid r : \theta') \log p(r_{cd} \mid r : \theta)\, dr_{cd} + \int p(r_{cd} \mid r : \theta') \log p(r \mid r_{cd})\, dr_{cd}. \tag{109}$$

It follows from this expression that

$$L_{id}(\hat{\theta}^{\text{new}}) - L_{id}(\hat{\theta}^{\text{old}}) = Q(\hat{\theta}^{\text{new}} \mid \hat{\theta}^{\text{old}}) - Q(\hat{\theta}^{\text{old}} \mid \hat{\theta}^{\text{old}}) - \int p(r_{cd} \mid r : \theta^{\text{old}}) \times \log \left[\frac{p(r_{cd} \mid r : \theta^{\text{new}})}{p(r_{cd} \mid r : \theta^{\text{old}})} \right] dr_{cd}. \tag{110}$$

The entropy (discrimination) inequality, $-\int p \log q/p \geq 0$, then yields

$$L_{id}(\hat{\theta}^{\text{new}}) - L_{id}(\hat{\theta}^{\text{old}}) \geq Q(\hat{\theta}^{\text{new}} \mid \hat{\theta}^{\text{old}}) - Q(\hat{\theta}^{\text{old}} \mid \hat{\theta}^{\text{old}}). \tag{111}$$

Noting that the maximization step in the expectation–maximization method implies that

$$Q(\hat{\theta}^{\text{new}} \mid \hat{\theta}^{\text{old}}) \geq Q(\hat{\theta}^{\text{old}} \mid \hat{\theta}^{\text{old}}) \tag{112}$$

then establishes that $L_{id}(\hat{\theta}^{\text{new}}) \geq L_{id}(\hat{\theta}^{\text{old}})$ and, hence, that the sequence produced recursively via the expectation–maximization method does not reduce the incomplete-data log-likelihood at any stage.

If an estimate of θ is sought that maximizes $L_{id}(\theta)+\alpha\Phi(\theta)$, corresponding to estimating θ subject to a penalty constraint with penalty function $\Phi(\theta)$ or to estimating θ with a log-prior $\log p(\theta) = \alpha\Phi(\theta)$, then the expectation–maximization method can be used with the maximization step becoming

$$\hat{\theta}^{\text{new}} = \arg\max_{\theta} [Q(\theta \mid \hat{\theta}^{\text{old}}) + \alpha\Phi(\theta)].$$

Wu [105], Shepp and Vardi [84], and Csiszár and Tusnady [24] address the convergence properties of the sequence of parameter estimates and corresponding sequence of incomplete-data log-likelihoods towards local and global maximizers of the incomplete-data log-likelihood.

B. Space-Alternating Generalized Expectation–Maximization

The expectation–maximization method as originally formulated maximizes a conditional expectation $Q(\theta|\hat{\theta}^{\text{old}})$ of a single complete-data log-likelihood function $L_{cd}(r_{cd})$ and simultaneously updates estimates of all the parameters comprising the parameter vector θ. While this method does permit maximum-likelihood estimates of images to be obtained numerically, it is slow in convergence, and penalty functions to enforce regularization and priors can make the maximization step difficult. Fessler and Hero [32] address these deficiencies in a method they term "space-alternating generalized expectation–maximization," or SAGE. In their SAGE method, parameters in θ are grouped into subsets that are sequentially updated by alternating between multiple, small, hidden-data spaces rather than a single, large complete-data space. The result is a numerical approach that, in comparison to the usual expectation–maximization method, produces maximum-likelihood estimates of an image with a convergence rate that is potentially greater and with a complexity that may be less in the presence of constraints.

The SAGE method can be summarized as follows. Let θ be a p-dimensional vector of parameters to be estimated, and index these parameters using the set of integers $\{1, 2, \cdots, p\}$. Let S and $\tilde{S}$ be subsets of these indices such that $S \cup \tilde{S} = \{1, 2, \cdots, p\}$ and $S \cap \tilde{S} = \emptyset$. Denote by θ_S the m-dimensional vector of elements of θ having indices in S, where m is the number of indices in S. Similarly, define $\theta_{\tilde{S}}$ to be the vector of dimension $p - m$ formed from the remaining elements of θ. In general, θ may be partitioned into more than just two subvectors in this way using multiple disjoint index sets S^i, $i = 1, 2, \cdots$ whose union covers $\{1, 2, \cdots, p\}$. Functions $f(\theta_S, \theta_{\tilde{S}})$ of the m- and $p - m$-dimensional vectors θ_S and $\theta_{\tilde{S}}$ are interpreted as equal to the function $f(\theta)$ of the p-dimensional vector θ. In the SAGE method, updates are performed by sequencing through the different index sets $S = S^i$ and updating only those parameters in θ_S while holding the other parameters $\theta_{\tilde{S}}$ fixed.

Hidden-data spaces must also be defined and selected; doing so requires that complete data r_{cd}^S be selected in the usual way for estimating θ_S but now assuming that $\theta_{\tilde{S}}$ is known. Let $\hat{\theta}^{(0)}$ be an initial estimate of θ. A sequence of estimates that results in a nondecreasing sequence of incomplete-data log-likelihoods is produced by the Fessler–Hero SAGE algorithm [32], which repeats the following iteration:

Step 1. Choose an index set $S = S^i$;
Step 2. Choose complete data $r_{cd}^{S^i}$ for θ_{S^i};
Step 3. (E-step) Compute $Q_i(\theta_{S^i}|\hat{\theta}^{(i)})$;
Step 4. (M-step)

$$\hat{\theta}_{S^i}^{(i+1)} = \arg\max_{\theta_{S^i}} Q_i(\theta_{S^i} \mid \hat{\theta}^{(i)}) \tag{113}$$

$$\hat{\theta}_{\tilde{S}^i}^{(i+1)} = \hat{\theta}_{\tilde{S}^i}^{(i)}; \tag{114}$$

Step 5. (optional) Repeat Step 3 and Step 4.

The ith iteration consists of this sequence of steps. The iterations are repeated for $i = 0, 1, 2, \cdots$, halting when the iterates reach an equilibrium. Thus it is necessary to prove that the iterates of the algorithm do converge to an equilibrium. The convergence properties of the SAGE algorithm and considerations to be made in selecting complete data are discussed by Fessler and Hero [32], and extensions are given in [33].

C. The Random Sampling Method

The EM and SAGE methods for numerically producing maximizers of likelihood functionals, both with and without priors, proceed deterministically: a sequence of functions or images is produced that is predetermined by the data given

and the function chosen to initiate the iteration. Although also iterative, the jump-diffusion method does not proceed deterministically but, rather, via a random search for maximizers or for estimates, such as the conditional mean (i.e., minimum mean-squared-error estimate). The method as introduced by Grenander and Miller [43] and discussed by Miller, Srivastava, and Grenander [63], [94], provides a numerical method for sampling from complicated distributions when the parameter space has both discrete and continuous components. It has been used effectively in a variety of applications, including identifying the number and shape of mitochondria in electron microscope images [43], deforming a labeled anatomy in a textbook to match a patient's anatomy [98], and detecting the number and orientation of targets in infrared images [59]. These various applications share the characteristic of having quantities in an image that are both discrete, such as the number of objects or the labeling of objects by their type, and continuous, such as the position and orientation of objects or spatially varying intensities in a scene containing the objects. The "jumps" in the method provide estimates of the discrete quantities by means of a stochastic search of the Metropolis–Hastings type, and the "diffusions" yield estimates of the continuous quantities through a stochastic optimization [35].

For example, let x_N represent parameters (such as the poses) and a_N the types of N objects in a scene. Denote the logarithm of the posterior likelihood of the data by $L(x_N : N, a_N)$ for given N and a_N. The approach is to formulate a diffusion process $\{x_N(t), t \geq 0\}$ that has the property that the log-distribution of $x_N(t)$ converges with increasing t towards $L(x_N : N, a_N)$. This diffusion is produced by the stochastic differential equation

$$dx_N(t) = \nabla_x L(x_N : N, a_N)\, dt + dw_N(t)$$

where $w_N(\cdot)$ is a standard N-dimensional Wiener process. Jumps between different choices of N and a_N are performed at the times of a Poisson process, and decisions of whether to select new values for N and a_N or to retain old ones are made in a manner similar to decisions made with the Metropolis–Hastings method of stochastic search. The cited references can be consulted for further details of the jump-diffusion approach.

D. Iterative Minimization of Discrimination

An iterative method that is similar to the expectation–maximization algorithm can be used to produce minimizers of discrimination for deterministic linear inverse problems. This approach has been suggested by Snyder, Schulz, and O'Sullivan [92]. Similar approaches are given by Vardi and Lee [100] and Byrne [12].

Linear inverse problems that can be approached with this method have the form

$$a(r) = \sum_x h(r, x) c(x) \tag{115}$$

where the three functions $a(\cdot)$, $h(\cdot, \cdot)$, and $c(\cdot)$ are nonnegative, with $a(\cdot)$ and $h(\cdot, \cdot)$ being given and $c(\cdot)$ to be determined. Joyce and Root [53] and many others have commented on the notoriously ill-posed character of many linear inverse problems. Various approaches have been suggested for solving them while introducing regularization to stabilize solutions. Most of these approaches are based on least squares optimization with constraints to enforce regularization, such as described by Tikhonov and Arsenin [96]. Youla [107] has proposed a method for accommodating nonnegativity constraints with least squares optimization.

As already noted, Csiszár [23] identified the important role of discrimination as a discrepancy measure for optimization when comparing nonnegative functions or images. Recognizing that a solution to the linear inverse problem described by (115) will necessarily be an approximation, a function $\hat{c}(\cdot)$ is sought such that the function $b(r : \hat{c})$, defined by

$$b(r : \hat{c}) = \sum_x h(r, x) \hat{c}(x) \tag{116}$$

is a good approximation to the given function $a(r)$ in the sense that the discrimination $I(a||b)$ between $b(r : \hat{c})$ and $a(r)$ is minimized. Let $c^{(0)}(x) > 0$ be a nonnegative function selected as an initial guess. Then, the sequence of functions $\{c^{(k)}(x), k = 0, 1, \cdots\}$ produced by the following recursion produces a corresponding sequence of discriminations $I(a||\hat{b}^{(k)})$ that is nonincreasing, where $\hat{b}^{(k)}(r) \equiv b(r : \hat{c}^{(k)})$

$$\hat{c}^{(k+1)}(x) = \hat{c}^{(k)}(x) \frac{1}{H_0(k)} \sum_r \left[\frac{h(r, x)}{\sum_{x'} h(r, x') \hat{c}^{(k)}(x')} \right] a(r). \tag{117}$$

Properties of the sequence $\{\hat{c}^{(k)}(x), k = 0, 1, \cdots\}$ and conditions for convergence are discussed by Snyder, Schulz, and O'Sullivan [92]; these are established using results from Cover [18] and Vardi, Shepp, and Kaufmann [101]; see also Vardi and Lee [100]. Applications to tomographic imaging are given by Wang, Snyder, O'Sullivan, and Vannier [103], and by Robertson, Yuan, Wang, and Vannier [75].

E. Generalized Iterative Scaling or SMART

The generalized iterative scaling algorithm was originally introduced to find the distribution that maximizes entropy subject to a set of linear (mean-value) constraints by Darroch and Ratcliff [25]. It was shown by Byrne to minimize the discrimination in the form of (104) for linear inverse problems with nonnegative data, using an alternating minimization approach [12]. Byrne referred to this algorithm as SMART for the simultaneous multiplicative algebraic reconstruction technique. Csiszár [22] showed that generalized iterative scaling can be interpreted as alternating I-projections and the convergence is thus covered by his more general results [21]. Byrne explicitly showed that this algorithm is in fact an alternating minimization algorithm whose convergence is covered by Csiszár and Tusnady [24]. O'Sullivan [70] discussed several alternating minimization algorithms including this one.

For linear inverse problems as in (115), the problem is to minimize $I(b||a)$, where $b(r : \hat{c})$ is the estimate for a as in (116). Let $c^{(0)}(x) > 0$ be a nonnegative function selected as

an initial guess. Then, the sequence of functions $\{c^{(k)}(x), k = 0, 1, \cdots\}$ produced by the following recursion produces a corresponding sequence of discriminations $I(\hat{b}^{(k)}||a)$ that is nonincreasing, where $\hat{b}^{(k)}(r) \equiv b(r : \hat{c}^{(k)})$

$$\hat{c}^{(k+1)}(x) = \hat{c}^{(k)}(x) \prod_r \left[\frac{a(r)}{\sum_{x'} \hat{c}^{(k)}(x') h(r, x')} \right]^{h(r,x)/\sum_{r'} h(r',x)}. \tag{118}$$

If there is a nonnegative solution c to (115), then the iterates $\hat{c}^{(k)}(x)$ converge to the solution of (115) that minimizes $I(c||\hat{c}^{(0)}(x))$ [25], [12].

F. Projection onto Convex Sets

The operation of projection onto a closed convex set in a Hilbert space is an example of a nonlinear procedure that can be explained in simple abstract terms. It is not normally viewed as a statistical method. Projection onto convex sets plays a role in image formation because the constraints on the image space are often convex. Moreover, the topic of projection onto convex sets can be expanded into the study of the powerful methods of alternating maximization [24], [70], [106]. These methods of alternating maximization applied to problems of information theory appeared earlier in the literature [3], [6] in the context of computing channel capacity and rate-distortion functions.

A general discussion of the topic of projection onto convex sets can be found in the paper of Combettes [17] and the work of Youla [106], Youla and Webb [108], and Segan and Stark [82]. The projection is unique and often can be found by analytically tractable methods, including iterative methods. Because the intersection of a finite number of convex sets $\mathcal{A}_\ell$ is convex, one may wish to project onto $\mathcal{A} = \cap_\ell \mathcal{A}_\ell$ by interatively projecting onto the individual $\mathcal{A}_\ell$. This procedure need not converge in general, but will always converge if the individual $\mathcal{A}_\ell$ are affine subspaces.

IX. Modalities and Applications

Some representative applications of information-theoretic imaging are described in this section. For each, the application is reviewed briefly and a likelihood model given for the data acquired for image formation. The applications are drawn from optical imaging, tomographic imaging, and radar imaging. The models include deterministic and random data, with the random data modeled by Poisson processes, Poisson–Gaussian mixtures, and Gaussian processes.

A. Deterministic Models

Imagation in which deterministic models are used for images and sensor data is often derived as a solution to a linear inverse problem in the form of a Fredholm integral equation

$$\int_X h(y, x) c(x)\, dx = a(y), \qquad y \in Y \tag{119}$$

where $\{a(y), y \in Y\}$ are the sensor data, $\{h(y, x), y \in Y, x \in X\}$ is a (point-spread) function characterizing the sensor, and $\{c(x), x \in X\}$ is the image to be formed. Some examples that illustrate the nature of the image and data spaces, X and Y, respectively, and functions that are encountered are given next.

Optical Imaging In optical imaging problems, $a(\cdot)$ represents the data acquired by a camera, Y is typically a two-dimensional subset of the plane $\boldsymbol{R}^2$, $h(\cdot, \cdot)$ is the point-spread function of the optical elements of the camera, such as telescope and microscope lenses, field stops, and mirrors, $c(\cdot)$ is the scene being imaged, and X is typically a subset of $\boldsymbol{R}^2$ or $\boldsymbol{R}^3$. For coherent imaging, where phase information is maintained, the functions $a(\cdot)$, $h(\cdot, \cdot)$, and $c(\cdot)$ are complex-valued functions. For incoherent imaging, these functions are real-valued and nonnegative functions. For multispectral, hyperspectral, polarimetric, or spectropolarimetric imaging, these functions are vector-valued.

Tomographic Imaging In tomographic imaging problems, $a(\cdot)$ represents the logarithm of the data acquired by the tomograph, Y is typically a subset of $\boldsymbol{R}^n$, with $n = 2$ or $n = 3$ corresponding to planar or volumetric imaging, $h(\cdot, \cdot)$ is the point-spread function of the tomograph, $c(\cdot)$ is the X-ray absorption density being imaged, and X is typically a subset of $\boldsymbol{R}^2$ or $\boldsymbol{R}^3$. For example, in helical-scan X-ray tomographic imaging in the fan-beam geometry, $y \in Y$ is three-dimensional with $y = (\beta, \gamma, z)$, where β is the angular position of the X-ray source, γ is the angle of a particular source to detector element, and z is the axial position of the source, $x \in X$ is three-dimensional with $x = (x_1, x_2, x_3 = z)$ being the coordinates of a point location in the target volume, and, for perfectly collimated source–detector combinations

$$\begin{aligned} h(y, x) &\equiv h(\gamma, \beta, z; x_1 x_2, x_3) \\ &= \delta[D \sin\gamma - x_1 \cos(\beta + \gamma) - x_2 \sin(\beta + \gamma)] \\ &\quad \cdot \delta(z - x_3) \end{aligned} \tag{120}$$

where D is the distance from the source to the axis of rotation, $z = x_3 = p\beta$, and p is the pitch of the helical scan. All functions in the linear inverse problem of tomographic imaging are constrained to be nonnegative.

Radar Imaging Complex-valued reflectance functions and real, nonnegative scattering functions are images of radar targets formed from high-resolution radar range data. If the signal transmitted by the radar is $s_T(t)$, the ideal echo-signal received from a point reflector is $cs_T(t - \tau)e^{j2\pi f(t-\tau/2)}$, where c is the strength of the reflector, τ is the two-way propagation delay of the transmitted signal to and from the point reflector, and f is the Doppler frequency shift due to relative motion between the radar transmitter and the reflector along the line of sight. For a spatially extended reflector, the received signal to a first approximation is the superposition of the signal reflected from each point; this neglects, for example, secondary reflections of the signal from one location on the reflector to another before returning

to the radar receiver. The received signal is then

$$a(t) = \int_{f_{\min}}^{f_{\max}} \int_{\tau_{\min}}^{\tau_{\max}} s_T(t-\tau) e^{j2\pi f(t-\tau/2)} c(f,\tau)\, df d\tau \tag{121}$$

where $(f_{\min}, f_{\max})$ is the range of Doppler shifts, $(\tau_{\min}, \tau_{\max})$ is the range of propagation delays that cover the reflector. This is in the form of (119) with Y being the time interval of the measurement, X being the two-dimensional space of delay-Doppler shifts

$$h(y,x) \equiv h(t; f, \tau) = s_T(t-\tau) e^{j2\pi f(t-\tau/2)}$$

and $c(x) \equiv c(f,\tau)$.

The linear inverse problem described (119) is routinely discretized to facilitate numerical solutions. While this can be accomplished in various ways, the result can generally be placed in the form of an algebraic, linear inverse problem of the form

$$\sum_x h(y,x)c(x) = a(y) \tag{122}$$

where x and y are discrete-valued or, alternatively, in matrix–vector form

$$\boldsymbol{Hc} = \boldsymbol{a} \tag{123}$$

in which $\boldsymbol{a}$ is a vector-valued discretization of the given data, $\boldsymbol{H}$ is a discretization of the kernel of the Fredholm equation (119), and $\boldsymbol{c}$ is a discretization of the unknown function that is sought. If $\boldsymbol{H}$ is invertible, then the obvious solution is $\boldsymbol{c} = \boldsymbol{H}^{-1}\boldsymbol{a}$. However, this ideal solution is usually impractical because $\boldsymbol{H}$ often is not invertible or is poorly conditioned so that solutions are extremely sensitive to the detailed choices made in designing a numerical implementation and to the effects of finite-precision arithmetic. Joyce and Root [53] provide a good discussion of this issue.

B. Stochastic Models

Sensor noise can be significant in inverse problems encountered in imaging. A variety of noise models are useful with the most successful results in applications occurring when the noise model selected is a good representation of the data-acquisition sensor being used. For radar sensors, an additive Gaussian model is a reasonable first choice, and for focal-plane arrays, such as a CCD camera, a Poisson model or a Poisson–Gaussian-mixture model is an appropriate initial choice. In stringent applications where high performance is sought, more refined models that account for significant effects present in a sensor must be formulated and used, so that, for example, nonuniformity of response and offset in focal plane arrays usually needs to be taken into account in scientific applications.

In the presence of additive Gaussian noise, the discrete inverse problem given by (119) becomes

$$r(y) = \sum_x h(y,x)c(x) + w(y), \qquad y \in Y$$

where $w(\cdot)$ is white with mean zero and variance σ^2, and the image recovery problem is to estimate $c(\cdot)$ given a realization of $r(\cdot)$. For describing photoconversion electrons in a focal-plane array, $a(\cdot)$ in (119) becomes a Poisson process $n(\cdot)$ with mean-value function $\sum_x h(y,x)c(x)$, and the restoration problem is to estimate $c(\cdot)$ from a realization of the Poisson process. If nonuniformity of response, offset, and thermoelectrons are significant, then the Poisson process $n(\cdot)$ modeling photoconversions has intensity $\beta(y)\sum_x h(y,x)c(x) + \mu_0(y)$. Here, $\beta(\cdot)$ and $\mu_0(\cdot)$ are functions that account for nonuniformity and offset, respectively; these functions are routinely determined in calibration measurements using a flat field and a dark field exposure of the focal-plane array. If read-out noise is a significant factor in a focal-plane-array sensor, then (119) becomes

$$r(y) = n(y) + w(y), \qquad y \in Y \tag{124}$$

where $n(\cdot)$ is a Poisson process modeling photoconversions and offset, and $w(\cdot)$ is an independent, white, Gaussian process modeling read-out noise. The mean-value function of $n(\cdot)$ is

$$E[n(y)] = \beta(y)\sum_x h(y,x)c(x) + \mu_0(y) \tag{125}$$

and the mean and variance of $w(y)$ are m and σ^2.

The data log-likelihoods for each of these models is a functional of $c(\cdot)$ that is fundamental to the problem of estimating $c(\cdot)$ from the available data. For the additive Gaussian noise model, the data log-likelihood (when reduced to only terms that are c-dependent) is

$$L(c) = \frac{1}{N_0}\mathrm{Re}\left[\sum_x \sum_y r^*(y)h(y,x)c(x)\right] - \frac{1}{2N_0}\sum_y \left|\sum_x h(y,x)c(x)\right|^2. \tag{126}$$

For the Poisson model, it is

$$L(c) = -\sum_y \sum_x \beta(y)h(y,x)c(x) + \sum_y \log\left[\sum_x h(y,x)c(x) + \mu_0(y)\right] n(y). \tag{127}$$

And, for the Poisson–Gaussian-mixture model

$$L(c) = \sum_y \log\left[\sum_{n(j)} \frac{1}{n(j)!}\mu^{n(j)}(j) e^{-\mu(j)} e^{[r(j)-n(j)-m]^2/2\sigma^2}\right] \tag{128}$$

where

$$\mu(j) = \beta(y)\sum_x h(y,x)c(x) + \mu_0(y). \tag{129}$$

The purpose of imagation is to recover or estimate the object $c(\cdot)$ given the data available. The method of maximum-likelihood estimation can be applied to this problem, and if there are constraints on the form of $c(\cdot)$ or if $c(\cdot)$ is a random process with a prior distribution, the method of maximum *a*

posteriori probability estimation can be used. A closed-form solution is well known for the additive Gaussian model without constraints or a prior, which is $\hat{c} = (\boldsymbol{H}^T\boldsymbol{H})^{\#}\boldsymbol{H}^T\boldsymbol{a}$. However, a closed-form solution is not possible for the Poisson and Poisson–Gaussian-mixture models, and numerical solutions such as those discussed in the next section must be employed.

Regularization is often necessary in order to obtain acceptable restorations. This is because the stochastic inverse problems are usually ill-posed and numerically unstable. In some cases, discretization is imposed by the sensor used to acquire data, such as with a charge-coupled-device camera and other focal-plane arrays. Discretization of continuous data is one form of regularization, but this alone can lead to the problem of dimensional instability described by Tapia and Thompson [95]. Grenander sieves [42] can be used to introduce regularization as was done by Snyder and Miller [89]. With this method, estimates are restricted to a subset of the function space C supporting $c(\cdot)$. The size of this subset is controlled by the amount of data available to perform the estimation, such that the subset grows as the amount of data increases, but the rate of growth is controlled so that the estimate of $c(\cdot)$ converges in a stable manner. Alternatively, regularization can be introduced via a penalty function $\Phi(c)$ that enforces smoothness (see O'Sullivan [69]). With penalty methods, the estimate maximizes the penalized log-likelihood $L(c) + \alpha\Phi(c)$, where α is a Lagrange multiplier that controls the emphasis given to the data log-likelihood and the penalty function when selecting the maximizer. When $c(\cdot)$ is a random process with a prior $p(c)$, the MAP estimate of $c(\cdot)$ is obtained by maximizing $L(c) + \log p(c)$. It is evident that many penalized maximum-likelihood estimation problems are equivalent to MAP estimation problems by defining $p(c) = \frac{1}{Z}e^{\alpha\Phi(c)}$, where Z is a normalization constant; for this equivalence to hold, Z must be finite, so that the prior defined in this way is proper.

C. An Application Modeled by Gaussian Data

For sensors that exhibit additive Gaussian noise $w(\cdot)$, (119) becomes

$$\int_X h(y,x)c(x)\,dx + w(y) = a(y), \qquad y \in Y \tag{130}$$

for which the discrete version, analogous to (130) is

$$\boldsymbol{Hc} + \boldsymbol{w} = \boldsymbol{a}. \tag{131}$$

If $\boldsymbol{c}$ is deterministic and $\boldsymbol{w}$ has zero mean, the data log-likelihood is

$$L(\boldsymbol{c}) = 2\mathrm{Re}\,(\boldsymbol{r}^\dagger\boldsymbol{W}^{-1}\boldsymbol{Hc}) - \boldsymbol{c}^\dagger\boldsymbol{H}^\dagger\boldsymbol{W}^{-1}\boldsymbol{Hc} \tag{132}$$

when terms that do not involve $\boldsymbol{c}$ are neglected, where the superscript † denotes the Hermitian transpose.

A model used for spectrum estimation and radar imaging arises when $\boldsymbol{c}$ has a prior distribution that is Gaussian with zero mean and diagonal covariance $\boldsymbol{\Sigma}$ [67], [71], [91]. The ij element, σ_{ij}^2, of $\boldsymbol{\Sigma}$ corresponds to the power gain of the signal reflected from the ij pixel in the pixelized representation of the object's scattering function in delay-Doppler coordinates. In this case, the data a are Gaussian-distributed with zero mean and covariance $\boldsymbol{K}_a = \boldsymbol{H}^\dagger\boldsymbol{\Sigma}\boldsymbol{H} + N_0 I$, assuming that the noise $\boldsymbol{w}$ is white Gaussian with zero mean and covariance $\boldsymbol{W} = N_0 I$, so the probability density of $\boldsymbol{a}$ is

$$p(\boldsymbol{r} : \boldsymbol{\Sigma}) = \pi^{-N}(\det \boldsymbol{K}_a)^{-1}\exp\,(-\boldsymbol{a}^\dagger\boldsymbol{K}_a^{-1}\boldsymbol{a})$$

where N is the dimension of $\boldsymbol{a}$. The problem of forming the scattering-function image is that of estimating the diagonal matrix $\boldsymbol{\Sigma}$ from some given data set $\boldsymbol{r}$. The log-likelihood function is

$$L(\boldsymbol{\Sigma}) = -\log\det\,(\boldsymbol{H}^\dagger\boldsymbol{\Sigma}\boldsymbol{H} + N_0 I) - \boldsymbol{a}^\dagger(\boldsymbol{H}^\dagger\boldsymbol{\Sigma}\boldsymbol{H} + N_0 I)^{-1}\boldsymbol{a}. \tag{133}$$

The maximization of $L(\boldsymbol{\Sigma})$ with respect to $\boldsymbol{\Sigma}$ is in the class of problems studied by Burg, Luenberger, and Wenger [10] for spectrum estimation and in [67], [71], [91] for radar imaging. The following algorithm, derived using the EM method, was used in the radar imaging context by Snyder, O'Sullivan, and Miller [91].

Step 0. Choose an initial estimate $\hat{\Sigma}^{(0)}$, set $k = 0$;
Step 1. Evaluate $\hat{c}^{(k)}$ and $\hat{\Sigma}^{(k+1)}$ according to

$$\hat{\boldsymbol{c}}^{(k)} = \hat{\Sigma}^{(k)}\boldsymbol{H}(\boldsymbol{H}^\dagger\hat{\Sigma}^{(k)}\boldsymbol{H} + N_0 I)^{-1}\boldsymbol{a} \tag{134}$$

$$\hat{\Sigma}^{(k+1)} = \hat{\Sigma}^{(k)} - \hat{\Sigma}^{(k)}\boldsymbol{H}(\boldsymbol{H}^\dagger\hat{\Sigma}^{(k)}\boldsymbol{H} + N_0 I)^{-1}\boldsymbol{H}^\dagger\hat{\Sigma}^{(k)\dagger} + \hat{\boldsymbol{c}}^{(k)}\hat{\boldsymbol{c}}^{(k)\dagger}; \tag{135}$$

Step 2. $k \leftarrow k + 1$;
Step 3. Repeat Step 1 until done.

D. An Application Modeled by Poisson Data

Scintillation detectors are used to sense photons emanating from radioactive decays in a radionuclide. Some radionuclides emit a single photon in each decay, as occurs in SPECT (single-photon-emission computed tomography) systems used in nuclear medicine. A decay in other radionuclides results in a positron, which interacts quickly with a nearby electron, resulting in two annihilation photons that propagate in nearly opposite directions away from the annihilation site, as in PET (positron-emission tomography) systems [90, Ch. 3]. A decay or an annihilation is called an *event*. Through the measurement of single-photon events or annihilation-pair events (typically, for PET, about 10^6 events per planar section, acquired in a time interval on the order of 10 to 20 min), the objective is to form an image displaying an estimate of the spatial distribution or concentration of the radionuclide. Three-dimensional, volumetric imaging is sought. This is usually accomplished by means of a sequence of planar images spanning the volume of interest, with each planar image being treated independently of others; however, direct volumetric imaging that accounts for intravolume dependencies has been demonstrated to be more accurate [60].

To obtain estimates of radionuclide concentrations, models for scintillation data must account for the photon-fluctuation statistics of radioactive decay and for the effects that occur when photons propagate through a scattering medium to reach detectors. The models that are used account only approximately for some effects and neglect others altogether. For

example, photon scattering (that is, deviation of a photon's flight path from a straight line due to Compton and photoelastic scattering) is usually only roughly accommodated using an attenuation function, an additive and independent "photon" noise in PET, and a point-response function that is broader than would be predicted by the finite size and geometry of scintillation detectors alone. Photons that are undetected due to finite recovery time in a scintillation detector are neglected. While these effects can be significant in practice, they are usually neglected to keep data models tractable.

A source-channel model for event detections is a useful conceptual framework for formulating the problem of estimating the radionuclide distribution. It can be formulated as follows. The source produces points representing random locations of radioactive decays or positron–electron annihilations in the region containing a radionuclide. Let $\mathcal{X}$ be the source-output space. This is the space where events occur; an individual event occurs as a point at position $x \in \mathcal{X}$. This source space can be a subset of $\boldsymbol{R}^2$ (planar SPECT and PET), $\boldsymbol{R}^3$ (volumetric SPECT and PET), $\boldsymbol{R}^2 \times \boldsymbol{R}_+$ (PET in which the differential time-of-flight of the annihilation photon pair is measured [93]), and perhaps other parameters, depending on the sensor configuration. The channel, representing the sensor system, produces outputs that are points in a channel-output space Y. A detected event occurs as a point at a random position $y \in \mathcal{Y}$. The elements of y depend on the sensor configuration. In SPECT, for example, $y = (p_1, p_2, \theta)$ and $Y = \boldsymbol{R}^2 \times [0, 2\pi)$, where (p_1, p_2) are the measured positions of the detection event in the scintillation crystal of the Anger camera, and θ is the angle of the camera in its orbit. In PET, y parameterizes the flight line of annihilation photons and, in time-of-flight PET, the flight-line parameters along with the differential propagation time. The channel can map a source point at x into channel-output point at y, or it can delete the source point (corresponding to an absorbed photon), and it can add extraneous points (accounting in part for photon scatter).

A reasonable model for the source, based on the physics of radioactive decay, is that the source produces points as an inhomogeneous Poisson process, denoted by $\{N(A), A \in \sigma(X)\}$, having an intensity function that is proportional to the concentration of the radionuclide. Let $\{\lambda(x), x \in \mathcal{X}\}$ denote the intensity function of the source.

We assume that the channel action on individual source points is independent from point to point. Let $\{p(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$ denote the transition probability-density of the channel; given that the source produces a point at x and that this point is detected, this is the density of the random location of the detection in the channel-output space. This transition density of the channel is the normalized point-spread function of the sensor. Let $\beta(y|x)$ denote the probability that a source point at x that is headed towards an output location y is detected (this is the photon survival probability), and let $\alpha(y|x) = 1 - \beta(y \mid x)$ be the probability that the source point is undetected (this is the photon-absorption probability). Finally, we assume that the channel can introduce extraneous (noise) points into its output and that these occur as an independent, inhomogeneous Poisson process with intensity $\{\mu_0(y), y \in \mathcal{Y}\}$. It follows from these assumptions, as was discussed by Miller and Snyder [62], that the channel output is also an inhomogeneous Poisson process, denoted by $\{M(B), B \in \sigma(\mathcal{Y})\}$, with intensity function $\{\mu(u), y \in \mathcal{Y}\}$, where

$$\mu(y) = \int_{\mathcal{X}} \beta(y \mid x) p(y \mid x) \lambda(x)\, dx + \mu_0(y). \tag{136}$$

Thus the log-likelihood functional of the channel-output process is given by (see Snyder and Miller [90, Chs. 2 and 3] for further discussion of this point)

$$l(\lambda) = -\int_{\mathcal{Y}} \mu(y)\, dy + \int_{\mathcal{Y}} \log[\mu(y)]\, M(dy). \tag{137}$$

The problem is to estimate the source intensity given the measured channel output points and the source-channel model. This problem was first formulated for emission tomography by Rockmore and Macovski [76] in 1976 using maximum-likelihood estimation, but their direct formulation proved to be intractable for producing maximum-likelihood estimates. It was not made computationally tractable until Shepp and Vardi [84] and Lange and Carson [58] applied the EM method to this estimation problem. Following this work, many subsequent publications have extended the approach. Recognizing that the EM algorithm will be implemented computationally, the first step is to discretize the source-output space and the channel-output space into pixels or voxels, then let $\{N(x), x \in \mathcal{X}\}$ and $\{M(y), y \in \mathcal{Y}\}$ denote the source-output and channel-output Poisson processes on the discrete spaces, where $N(x)$ is the number of single or annihilation-pair photons occurring in pixel x, and $M(y)$ is the number of detection events in pixel y. The log-likelihood functional of the channel-output process becomes

$$l(\lambda) = \sum_{y \in \mathcal{Y}} \mu(y) + \sum_{y \in \mathcal{Y}} \log[\mu(y)] M(y) \tag{138}$$

where

$$\mu(y) = \sum_{x \in \mathcal{X}} \beta(y \mid x) p(y \mid x) \lambda(x) + \mu_0(y). \tag{139}$$

Depending on the choice of complete data, Politte and Snyder [74] identify the choice of two algorithms formed by the EM method. The algorithm formed by the EM method will be either

$$\hat{\lambda}^{(k+1)}(x) = \hat{\lambda}^{(k)}(x) \left\{ \bar{\alpha}(x) + \sum_{y \in \mathcal{Y}} \left[\frac{\beta(y \mid x) p(y \mid x)}{\hat{\mu}^{(k)}(y)} \right] M(y) \right\} \tag{140}$$

or

$$\hat{\lambda}^{(k+1)}(x) = \hat{\lambda}^{(k)}(x) \frac{1}{\bar{\beta}(x)} \sum_{y \in \mathcal{Y}} \left[\frac{\beta(y \mid x) p(y \mid x)}{\hat{\mu}^{(k)}(y)} \right] M(y) \tag{141}$$

depending on the choice of complete data, where

$$\bar{\beta}(x) = 1 - \bar{\alpha}(x) = \sum_{y \in Y} \beta(y \mid x) p(y \mid x)$$

and where

$$\hat{\mu}^{(k)}(y) = \int_X \beta(y \mid x) p(y \mid x) \hat{\lambda}^{(k)}(x)\, dx + \mu_0(y). \quad (142)$$

While these two EM algorithms converge towards the same limit point, their convergence rates differ, with the second one converging more rapidly [74]. This shows that the choice of complete data does influence algorithm behavior.

The SPECT and PET inverse problems are ill-posed, so that regularization to stabilize solutions is needed. Sieves and roughness penalties have been used for this purpose [61], [74], [89].

Data acquired in optical imaging systems are also often modeled as Poisson-distributed. One important area where such models along with information-based image recovery has been used effectively is in addressing the long-standing and difficult problem faced by astronomers of forming images of objects seen through clear-air atmospheric turbulence. Roggemann and Welsh [77] review the classic methods of Labeyrie (recovery from Fourier modulus), of Knox and Thompson (recovery from squared Fourier modulus or second-order correlations), and of Weigelt (recovery from third-order correlations) developed and used effectively by astronomers for this problem. A new method of recovery of an object's image from known second-order and higher order correlation functions of the image has been developed by Snyder and Schulz [79], [80], [86], based on a Poisson data model and the use of maximum-likelihood estimation. Paxman, Schulz, and Fienup [72] and Seldin and Paxman [81] have introduced a new data-collection approach in which multiple, phase-diverse snapshots of an object seen through turbulence are used with a Poisson data model and constrained maximum-likelihood estimation to produce substantially improved object images.

E. An Application Modeled by Poisson–Gaussian Data

The following source-channel model is a useful framework for characterizing a wide variety of applications when a charge-couple-device (CCD) camera is used to image scenes in the visible and infrared portions of the spectrum. A discrete model is used because a CCD camera produces data from a pixel array and, also, because an EM algorithm will be used to perform imagation. We envision a scene that emits incoherent radiation that propagates towards a CCD camera. Light falling onto the focal plane of the camera has an intensity given by

$$i(y) = \sum_{x \in \mathcal{X}} h(y \mid x) \lambda(x), \qquad y \in \mathcal{Y} \quad (143)$$

where $\{h(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$ is the point-spread function of the camera, $\{\lambda(x), x \in \mathcal{X}\}$ is the radiance of the scene, and are the source-output and channel-output spaces, respectively. For light that propagates through free space or through short paths in the atmosphere, the point-spread function is determined by the configuration of optical elements in the camera [39], [40], including pupil shape, obscurations, and any aberrations that are present. The number $M(y)$ of photoelectron conversions occurring during a T-second exposure interval in a pixel at y in the CCD array is Poisson-distributed with mean $\mu(y) = T\beta(y)i(y)$, where $\beta(y)$ accounts for nonuniform quantum efficiency, pattern noise, bad pixels, and charge-transfer inefficiency. By assumption, the number of photoconversions is independent from pixel to pixel. The process of reading out the pixel values results in the channel-output process

$$R(y) = M(y) + M_0(y) + G(y), \qquad y \in \mathcal{Y} \quad (144)$$

where $\{M_0(y), y \in \mathcal{Y}\}$ is a Poisson-distributed process that accounts for extraneous thermoelectrons and for offset bias in the CCD array, and $\{G(y), y \in \mathcal{Y}\}$ is a Gaussian-distributed process accounting for noise in the readout amplifier integrated into the CCD array circuit [88]. The processes $M(\cdot)$, $M_0(\cdot)$, and $G(\cdot)$ are mutually independent and independent from pixel to pixel. The mean-value function for $\{M_0(y), y \in \mathcal{Y}\}$ is assumed to be the known function $\{\mu_0(y), y \in \mathcal{Y}\}$, and $\{G(y), y \in \mathcal{Y}\}$ is assumed to have a constant mean m and variance σ^2.

For imagation, it is convenient to embed the scene in a hypothetical stochastic process, which can be regarded as the output of the source in the source-channel model. Thus we imagine a Poisson-distributed process $\{N(x), x \in \mathcal{X}\}$ having intensity $\{\lambda(x), x \in \mathcal{X}\}$; one can regard the points of this process as "photons" emanating from the scene. The source output is the set of points (or counts in the discrete model) of this hypothetical process in the source space $\mathcal{X}$. This contrived source model is legitimate because the photoconversion process $\{M(y), y \in \mathcal{Y}\}$ will be a Poisson process with mean function $\{\mu(y), y \in \mathcal{Y}\}$ when the source output $\{N(y), y \in \mathcal{Y}\}$ is a Poisson process with mean function $\{\lambda(y), y \in \mathcal{Y}\}$ and

$$\mu(y) = T\beta(y) \sum_{x \in \mathcal{X}} h(y \mid x) \lambda(x), \qquad y \in \mathcal{Y}. \quad (145)$$

The channel, representing the camera, maps the output of the source into the channel-output process $\{R(y), y \in \mathcal{Y}\}$, which is a Poisson–Gaussian mixture. The log-likelihood functional for the channel output is

$$\ell(\lambda) = \sum_{y \in \mathcal{Y}} \log\left(\frac{1}{n(y)!} [\mu(y) + \mu_0(y)]^{n(y)} \exp\{-[\mu(y) + \mu_0(y)]\} \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-[R(y) - n(y) - m]^2 / 2\sigma^2\} \right). \quad (146)$$

Selecting complete data and applying the EM algorithm yields [88]

$$\hat{\lambda}^{(k+1)}(x) = \hat{\lambda}^{(k)}(x) \frac{1}{\bar{\beta}(x)} \sum_{y \in \mathcal{Y}} \left[\frac{\beta(y) p(y \mid x)}{\hat{\mu}^{(k)}(y)} \right] f[R(y), \hat{\mu}^{(k)}, \sigma] \quad (147)$$

where

$$f[r, \mu, \sigma] = \frac{\sum_{n=0}^{\infty} (n/n!) \exp[-(r - n - m)^2 / 2\sigma^2]}{\sum_{n=0}^{\infty} (1/n!) \mu^n \exp v[-(r - n - m)^2 / 2\sigma^2]} \quad (148)$$

and

$$\hat{\mu}^{(k)}(y) = T\beta(y) \sum_{x \in \mathcal{X}} h(y \mid x) \hat{\lambda}^{(k)}(x) + \mu_0(y), \qquad y \in \mathcal{Y}. \tag{149}$$

Evaluation of the function $f[\cdots]$ through the use of saddle-point integration and approximations, with applications to Hubble Space Telescope imagery, is discussed by Snyder, Helstrom, Lanterman, Faisal, and White [87].

X. Conclusions and Future Directions

An information-theoretic framework for imaging is in the earliest stages of development but can already be seen as the basis for data models, performance metrics, and processing strategies for treating image formation problems. An all-encompassing model has yet to be formulated that can play as powerful a role for imaging as Shannon's source-channel model plays for communications. Nonetheless, the importance that information-theoretic concepts already play leads us to predict that such a formal model will eventually emerge.

Image formation often involves optimization of metrics rooted in information theory, such as likelihood, divergence, discrimination, and entropy. For such methods, it is not only a requirement but also a strength that accurate models must be available for scenes, for the environment between scenes and sensors, and for the image-related data produced by sensors. These models need to account generally for the way that the underlying physics governs the production of the observed data at each stage along the way. Deterministic and stochastic models may appear different on the surface, but image-formation methods based on the optimization of information-theoretic metrics of discrimination and likelihood share many common features. Scenes exhibit great complexity and variability; methods for modeling scenes are evolving rapidly and are already sophisticated mathematically, but in many respects, available models are still too limited to accommodate effects that can have a pronounced influence on the performance of imaging systems, such as clutter that surrounds and often obscures objects to be identified in a scene. Propagation effects in optical imaging applications, such as scattering in turbid media and phase and amplitude fluctuations in turbulent media cannot be easilly modeled. Sensor technology is complicated and evolves rapidly so that models for sensor data often have limited accuracy. The future effectiveness of information-theoretic approaches to image formation will rely on addressing these modeling issues.

Acknowledgment

The authors wish to thank the editors for inviting them to participate in the fifty-year anniversary of Information Theory through this contribution. They are grateful to Dr. Pierre Moulin, to Dr. Alfred O. Hero III, to Dr. Aaron Lanterman, to Dr. G. David Forney, Jr., and to Dr. Bruce Hajek for reading drafts of the manuscript and providing many helpful comments.

References

[1] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Roy. Statist. Soc. Ser. B*, vol. 28, pp. 131–142, 1966.
[2] S. Amari, *Differential-Geometrical Methods in Statistics* (Lecture Notes in Statistics, vol. 28). Berlin, Germany: Springer-Verlag, 1985.
[3] S. Arimoto, "An algorithm for computing the capacity of an arbitrary discrete memoryless channel," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 14–20, 1972.
[4] J. Besag, "Spatial interaction and the statistical analysis of lattice systems," *J. Roy. Statist. Soc.*, vol. 36, pp. 192–236, 1974.
[5] A. Blake and A. Zisserman, *Visual Reconstruction*. Cambridge, MA: MIT Press, 1987.
[6] R. E. Blahut, "Computation of channel capacity and rate distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, 1972.
[7] L. M. Bregmen, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *U.S.S.R. Comp. Math. and Math. Phys.*, vol. 7, pp. 200–217, 1967.
[8] J. Browne and A. R. De Pierro, "A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography," *IEEE Trans. Med. Imag.*, vol. 15, pp. 687–699, Oct. 1996.
[9] J. A. Bucklew, *Large Deviation Techniques in Decision, Simulation, and Estimation*. New York: Wiley, 1990.
[10] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices," *Proc. IEEE*, vol. 70, pp. 963–974, Sept. 1982.
[11] C. S. Butler and M. I. Miller, "Maximum a posteriori estimation for single photon emission computed tomography using regularization techniques on a massively parallel computer," *IEEE Trans. Med. Imag.*, vol. 12, pp. 84–89, Mar. 1993.
[12] C. L. Byrne, "Iterative image reconstruction algorithms based on cross-entropy minimization," *IEEE Trans. Image Processing*, vol. 2, pp. 96–103, Jan. 1993.
[13] R. Chellappa, *Markov Random Fields: Theory and Applications*. New York: Academic, 1993.
[14] P. A. Chou, M. Effros, and R. M. Gray, "A vector quantization approach to universal noiseless coding and quantization," *IEEE Trans. Inform. Theory*, vol. 42, pp. 1109–1138, July 1996.
[15] Y. Chow and U. Grenander, "A sieve method for the spectral density," *Ann. Stat.*, vol. 13, no. 3, pp. 998–1010, 1985.
[16] C. K. Chui, *Multivariate Splines*. Philadelphia, PA: SIAM, 1988.
[17] P. L. Combettes, "The foundation of set theoretic estimation," *Proc. IEEE*, vol. 81, pp. 182–208, 1993.
[18] T. M. Cover, "An algorithm for maximizing expected log investment return," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 369–373, 1984.
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[20] I. Csiszár, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.
[21] ———, "I-divergence geometry of probability distributions and minimization problems," *Ann. Prob.*, vol. 3, no. 1, pp. 146–158, 1975.
[22] ———, "A geometric interpretation of Darroch and Ratcliff's generalized iterative scaling," *Ann. Stat.*, vol. 17, no. 3, pp. 1409–1413, 1989.
[23] ———, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Ann. Stat.*, vol. 19, pp. 2032–2066, 1991.
[24] I. Csiszár and G. Tusnady, "Information geometry and alternating decisions," *Stat. Decisions*, Suppl. issue no. 1, pp. 205–207, 1984.
[25] J. N. Darroch and D. Ratcliff, "Generalized iterative scaling for log-linear models," *Ann. Math. Stat.*, vol. 43, no. 5, pp. 1470–1480, 1972.
[26] I. Daubechies, *Ten Lectures on Wavelets*. Philadelphia, PA: SIAM, 1992.
[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Stat. Soc. Ser. B*, vol. 39, pp. 1–37, 1977.
[28] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, 1994.
[29] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613–627, May 1995.
[30] D. L. Donoho *et al.* "Wavelet shrinkage: Asymptotia?" (with discussion), *J. Roy. Stat. Soc. Ser. B*, vol. 57, no. 2, pp. 301–369, 1995.
[31] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. New York: Springer-Verlag, 1985.
[32] J. A. Fessler and A. O. Hero, "Space alternating generalized expectation-maximization algorithm," *IEEE Trans. Signal Processing*, vol. 42, pp.

2664–2677, 1994.
[33] ——, "Penalized maximum likelihood image reconstruction using space alternating generalized EM algorithms," *IEEE Trans. Image Processing*, vol. 4, pp. 1417–1429, Oct. 1995.
[34] S. Geman and D. Geman, "Stochastic relaxation, Gibbs' distributions, and Bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-6, pp. 721–741, Nov. 1984.
[35] S. Geman and C.-R. Hwang, "Diffusions for global optimization," *SIAM J. Contr. Optimiz.*, vol. 24, pp. 1031–1043, 1987.
[36] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Boston, MA: Kluwer, 1992.
[37] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: Johns Hopkins Univ. Press, 1996.
[38] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities," *Biometrika*, vol. 58, pp. 255–277, 1971.
[39] J. W. Goodman, *Statistical Optics*. New York: Wiley-Interscience, 1986.
[40] ——, *Fourier Optics*. New York: McGraw-Hill, 1985.
[41] J. D. Gorman and A. O. Hero, "Lower bounds for parameter estimation with constraints," *IEEE Trans. Inform. Theory*, vol. 36, pp. 1285–1301, Nov. 1990.
[42] U. Grenander, *Abstract Inference*. New York: Wiley, 1981.
[43] U. Grenander and M. I. Miller, "Representations of knowledge in complex systems," *J. Roy. Stat. Soc., Ser. B*, vol. 56, pp. 549–603, 1994.
[44] U. Grenander, M. I. Miller, and A. Srivastava, "Hilbert–Schmidt lower bounds for estimators on matrix Lie groups," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.
[45] R. M. Haralick and L. G. Shapiro, *Computer and Robot Vision*, vols. 1 and 2. Reading, MA: Addison-Wesley, 1993.
[46] C. Heil, "Wavelets and frames," in *Signal Processing Part I: Signal Processing Theory*, L. Auslander *et al*., Eds. New York: Springer-Verlag, 1990.
[47] A. Hero and J. A. Fessler, "A recursive algorithm for computing Cramer–Rao type bounds on estimator covariance," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1205–1210, July 1994.
[48] A. O. Hero, M. Usman, A. C. Sauve, and J. A. Fessler, "Recursive algorithms for computing the Cramer–Rao bound," *IEEE Trans. Signal Processing*, vol. 45, pp. 803–807, Mar. 1997.
[49] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Trans. Med. Imag.*, vol. 13, pp. 601–609, Aug. 1994.
[50] F. Itakura and S. Saito, "Analysis synthesis telephony based on the maximum likelihood method," in *Repts. 6th Int. Congr. Acoustics*, Y. Kohasi, Ed. (Tokyo, Japan, 1968), pp. 17–20.
[51] E. T. Jaynes, "On the rationale of maximum entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, 1982.
[52] L. K. Jones and C. L. Byrne, "General entropy criteria for inverse problems, with applications to data compression, pattern classification and cluster analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 23–30, 1990.
[53] L. S. Joyce and W. L. Root, "Precision bounds in superresolution processing," *J. Opt. Soc. Amer. A*, vol. 1, pp. 149–168, 1984.
[54] A. Kirsch, "An introduction to the mathematical theory of inverse problems," *Applied Mathematical Sciences*, vol. 120. Berlin, Germany: Springer-Verlag, 1996.
[55] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959; Dover, 1968.
[56] S. Kullback and M. A. Khairat, "A note on minimum discrimination information," *Ann. Math. Stat.*, vol. 37, pp. 279–280, 1966.
[57] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Stat.*, vol. 22, pp. 79–86, 1951.
[58] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comp. Assist. Tomogr.*, vol. 8, no. 2, pp. 306–316, 1984.
[59] A. D. Lanterman, M. I. Miller, and D. L. Snyder, "General metropolis-hastings jump diffusions for automatic target recognition in infrared scenes," *Opt. Eng.*, vol. 36, pp. 1123–1137, 1997.
[60] M. I. Miller and C. S. Butler, "3-D maximum a posteriori estimation for single photon emission computed tomography on massively parallel computers," *IEEE Trans. Med. Imag.*, vol. 12, pp. 560–565, Sept. 1993.
[61] M. I. Miller and B. Roysam, "Bayesian image reconstruction for emission tomography: Implementation of the EM algorithm and good's roughness prior on massively parallel processors," *Proc. Nat. Acad. Sci.*, vol. 88, pp. 3223–3227, 1991.
[62] M. I. Miller and D. L. Snyder, "The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and toeplitz and constrained covariances," *Proc. IEEE*, vol. 75, pp. 892–907, 1987.
[63] M. I. Miller, A. Srivastava, and U. Grenander, "Conditional-mean estimation via jump-diffusion processes in multiple target tracking and recognition," *IEEE Trans. Signal Processing*, vol. 43, pp. 2678–2690, 1995.
[64] P. Moulin, "A method of sieves for radar imaging and spectrum estimation," D.Sc. dissertation, Dept. Elec. Eng. Washington Univ., St. Louis, MO, 1990.
[65] ——, "A multiscale relaxation technique for SNR maximization in nonorthogonal subband coding," *IEEE Trans. Image Processing*, vol. 4, pp. 1269–1281, Sept. 1995.
[66] P. Moulin and J. Liu, "Analysis of multiresolution image denoising schemes using generalized-gaussian and complexity priors," preprint, submitted for publication.
[67] P. Moulin, J. A. O'Sullivan, and D. L. Snyder, "A method of sieves for multiresolution spectrum estimation and radar imaging," *IEEE Trans. Inform. Theory*, vol. 38, pp. 801–813, 1992.
[68] M. Nikolova, J. Idier, and A. Mohammad-Djafari, "Inversion of large-support Ill-posed linear operators using a piecewise Gaussian MRF," *IEEE Trans. Image Processing*, vol. 7, pp. 571–585, Apr. 1998.
[69] J. A. O'Sullivan, "Roughness penalties on finite domains," *IEEE Trans. Image Processing*, vol. 4, pp. 1258–1268, Sept. 1995.
[70] ——, "Alternating minimization algorithms: From Blahut–Arimoto to expectation–maximization," in *Codes, Curves, and Signals: Common Threads in Communications*, A. Vardy, Ed. Norwell, MA: Kluwer, 1998.
[71] J. A. O'Sullivan, D. L. Snyder, D. G. Porter, and P. Moulin, "An application of splines to maximum likelihood radar imaging," *J. Imaging Syst. Technol.*, vol. 4, pp. 256–264, 1992.
[72] R. G. Paxman, T. J. Schulz, and J. R. Fienup, "Joint estimation of object and aberrations by using phase diversity," *J. Opt. Soc. Amer. A*, vol. 9, pp. 1072–11085, July 1992.
[73] D. G. Politte, "Reconstruction algorithms for time-of-flight assisted positron-emission tomography," M.S.E.E. thesis, Sch. Eng. Appl. Sci., Washington Univ., St. Louis, MO, 1983.
[74] D. G. Politte and D. L. Snyder, "Corrections for accidental coincidences in maximum-likelihood image reconstruction for position-emission tomography," *IEEE Trans. Med. Imag.*, vol. 10, pp. 82–89, 1991.
[75] D. D. Robertson, J. Yuan, G. Wang, and M. W. Vannier, "Total hip prosthesis metal-artifact suppression using iterative deblurring reconstruction," *J. Comp. Assist. Tomogr.*, vol. 21, pp. 293–298, 1997.
[76] A. Rockmore and A. Macovski, "A maximum likelihood approach to emission image reconstruction from projections," *IEEE Trans. Nucl. Sci.*, vol. NS-23, pp. 1428–1432, 1976.
[77] M. C. Roggemann and B. Welsh, *Imaging Through Turbulence*. New York: CRC Press, 1996.
[78] I. N. Sanov, "On the probability of large deviations of random variance," *Matem. Sbornik*, vol. 42, pp. 11–44, 1957.
[79] T. J. Schulz and D. L. Snyder, "Imaging a randomly moving object from quantum-limited data: Applications to image recovery from second- and third-order autocorrelations," *J. Opt. Soc. Amer. A*, vol. 8, pp. 801–807, May 1991.
[80] ——, "Image recovery from correlations," *J. Opt. Soc. Amer. A*, vol. 9, pp. 1266–1272, Aug. 1992.
[81] J. H. Seldin and R. G. Paxman, "Phase-diverse speckle reconstruction of solar data," in *Proc. SPIE Conf. 2302*, July 1994.
[82] M. I. Sezan and H. Stark, "Image restoration by the method of convex projections: Part 2—Applications and numerical results," *IEEE Trans. Med. Imag.*, vol. MI-1, pp. 95–101, 1982.
[83] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 1948.
[84] L. A. Shepp and Y. Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. MI-1, pp. 113–122, 1982.
[85] J. E. Shore and R. W. Johnson, "Axiomatic derivation of the principle of maximum entropy and the principle of minimum cros-entropy," *IEEE Trans, Inform. Theory*, vol. IT-26, pp. 26–37, Jan. 1980.
[86] D. L. Snyder and T. J. Schulz, "High-resolution imaging at low-light levels through weak turbulence," *J. Opt. Soc. Amer. A*, vol. 7, pp. 1251–1265, July 1990.
[87] D. L. Snyder, C. W. Helstrom, A. D. Lanterman, M. Faisal, and R. L. White, "Compensation for read-out noise in CCD images," *J. Opt. Soc. Amer. A*, vol. 12, pp. 272–283, 1995.
[88] D. L. Snyder, A. M. Hammoud, and R. L. White, "Image recovery from data acquired with a charge-coupled-device camera," *J. Opt. Soc. Amer. A*, vol. 10, pp. 1014–1023, 1993.
[89] D. L. Snyder and M. I. Miller, "The use of sieves to stabilize images produced with the EM algorithm for emission tomography," *IEEE Trans. Nucl. Sci.*, vol. NS-32, pp. 3864–3872, 1985.
[90] ——, *Random Point Processes in Time and Space*. New York: Springer-Verlag, 1991, ch. 3.

[91] D. L. Snyder, J. A. O'Sullivan, and M. I. Miller, "The use of maximum likelihood estimation for forming images of diffuse radar targets from delay-doppler data," *IEEE Trans. Inform. Theory*, vol. 35, pp. 536–548, 1989.

[92] D. L. Snyder, T. J. Schulz, and J. A. O'Sullivan, "Deblurring subject to nonnegativity constraints," *IEEE Trans. Signal Processing*, vol. 40, pp. 1143–1150, 1992.

[93] D. L. Snyder, L. J. Thomas Jr., and M. M. TerPogossian, "A mathematical model for positron-emission tomography systems having time-of-flight measurements," *IEEE Trans. Nucl. Sci.*, vol. NS-28, pp. 3575–3583, 1981.

[94] A. Srivastava, M. I. Miller, and U. Grenander, "Multiple target direction of arrival tracking," *IEEE Trans. Signal Processing*, vol. 43, pp. 1282–1285, 1995.

[95] R. A. Tapia and J. R. Thompson, *Nonparametric Probability Density Estimation*. Baltimore, MD: Johns Hopkins Univ. Press, 1978.

[96] A. N. Tikhonov and V. Y. Arsenin, *Solutions of Ill-Posed Problems*. Washington, DC: Winston, 1977.

[97] R. Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses," *IEEE J. Robot. Automat.*, vol. RA-3, no. 4, pp. 323–344, 1987.

[98] M. W. Vannier, M. I. Miller, and U. Grenander, "Modeling and data structure for registration to a brain atlas of multimodality images," in *Functional Neuroimaging—Technical Foundations*, R. W. Thatcher *et al*., Eds. New York: Academic, 1994, pp. 217–221.

[99] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, pt. I. New York: Wiley, 1968.

[100] Y. Vardi and D. Lee, "From image deblurring to optimal investments: Maximum-likelihood solutions to positive linear inverse problems," *J. Roy. Stat. Soc. Ser. B*, pp. 569–612, 1993.

[101] Y. Vardi, L. A. Shepp, and L. Kaufmann, "A statistical model for positron emission tomography," *J. Amer. Stat. Soc.*, vol. 80, pp. 8–35, 1985.

[102] G. Wahba, *Spline Models for Observational Data*. Philadelphia, PA: SIAM, 1990.

[103] G. Wang, D. L. Snyder, J. A. O'Sullivan, and M. W. Vannier, "Iterative deblurring for CT metal artifact reduction," *IEEE Trans. Med. Imag.*, vol. 15, pp. 657–664, 1996.

[104] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: Wiley, 1965.

[105] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Stat.*, vol. 11, pp. 95–103, 1983.

[106] D. C. Youla, "Generalized image restoration by the method of alternating orthogonal protections," *IEEE Trans. Circuits Syst.*, vol. CAS-25, pp. 694–702, 1978.

[107] ———, "Mathematical theory of image restoration by the method of convex projections," in *Image Recovery, Theory and Applications*, H. Stark, Ed. New York: Academic, 1987, ch. 2.

[108] D. C. Youla and H. Webb, "Image restoration by the method of convex projections: Part 1—Theory," *IEEE Trans. Med. Imag.*, vol. MI-1, no. 2, pp. 81–94, 1982.