Roughness Penalties on Finite Domains

Joseph A. O'Sullivan, *Senior Member, IEEE*

***Abstract*— A class of penalty functions for use in estimation and image regularization is proposed. These penalty functions are defined for vectors whose indexes are locations in a finite lattice as the discrepancy between the vector and a shifted version of itself. After motivating this class of penalty functions, their relationship to Markov random field priors is explored. One of the penalty functions proposed, a divergence roughness penalty, is shown to be a discretization of a penalty proposed by Good and Gaskins for use in density estimation. One potential use in estimation problems is explored. An iterative algorithm that takes advantage of induced neighborhood structures is proposed and convergence of the algorithm is proven under specified conditions. Examples in emission tomographic imaging and radar imaging are given.**

I. Introduction

The use of Markov random field models in image estimation and regularization problems has become common [2], [3], [11], [14], [20], [34]. While these models have been applied to nonnegative-valued images, there does not appear to be a standard or natural model for these cases in the same sense that Gauss–Markov random fields are natural for real valued images. The penalty methods proposed in this paper form a possible alternative approach. While penalties may be considered to be equivalent to priors (in an obvious sense to be made precise in Section III), their use may be better motivated by taking the viewpoint proposed here.

The goal of penalties is to include prior information in estimation problems. When this information concerns the smoothness of the estimates, a roughness penalty results. As discussed below, the roughness penalties are built from two quantities: shifts on the lattice and discrepancy measures. For simplicity, this paper restricts attention to periodic shifts on the lattice, resulting in what are commonly referred to as periodic boundary conditions. The penalty is determined by the discrepancy between the original vector and the shifted version of itself. The class of discrepancy measures studied is motivated by Csiszár [8] and Jones and Byrne [19]. In those papers, axiomatic derivations of least squares and I-divergence discrepancy measures are presented. The examples in this paper use those measures, but the results hold for a wide class of discrepancy measures.

Several methods for regularization in estimation problems have been proposed. The majority of them may be classified as penalty, constraint, prior probability, and stopping criterion methods. Stopping criteria have been proposed for iterative algorithms such as the expectation-maximization (EM) algorithm. These methods are based on the observation that image estimates converge to rough images as iterations proceed. By prematurely stopping the algorithms prior to convergence, the images should be smoother than if the algorithms were allowed to converge. Prior probability methods (such as Markov random field priors) may be used when a stochastic model for the data is appropriate. These methods are very similar (equivalent in many cases) to penalty methods. Hierarchical priors have been used to successively put neighborhood structures in regions then segment the estimate into separate regions [11], [20], [24]. Miller *et al.* [24] (see also the references of [24]) relate some constraint and prior methods, demonstrating an equivalence between them. Constraint methods include Grenander's method of sieves [7], [26], [32] where the estimate is constrained to be in a subset of parameter space. The subset is indexed by a quantity called the mesh size; as the number of data points gets large, the mesh size gets small, and the subset converges to the entire parameter space. These methods are important for studying properties such as consistency of the estimates. Penalty methods may be classified into two categories: those penalizing the discrepancy with a prior guess and those penalizing the roughness of the estimate. Our approach is the latter. The former are discussed in recent papers by Byrne [5], [6], they motivate Csiszár's results [8] and Jones' results [18], [19], and they include maximum entropy penalties. Lange [21] (see also references in [21]) uses penalties that are special cases of those derived here.

A major issue is the relationship between the proposed roughness penalties for finite lattices and penalties on functions defined on continuous domains. The divergence penalty introduced in the following section is shown to be a discretization of an information-theoretic penalty due to Good and Gaskins [13], [35], giving further evidence of its importance in estimation problems.

Applications of the results in this paper to two problems of interest are presented. The first is emission tomographic imaging. In this case, the equivalence between the divergence penalty and a standard implementation of a penalty due to Good and Gaskins [4], [23] is shown. The second is radar imaging, in which a very noisy image consisting of exponentially distributed random variables is given. The divergence penalty is applied to the means of the pixels and estimates are

Manuscript received October 29, 1993; revised September 6, 1994. This work was supported in part by Office of Naval Research contract N00014-92-J-1418 and in part by Rome Laboratory contract F30602-92-C-0004. The associate editor coordinating the review of this paper and approving it for publication was Prof. Xinhua Zhuang.

The author is with the Electronic Systems and Signals Research Laboratory, Department of Electrical Engineering, Washington University, St. Louis, MO 63130 USA.

IEEE Log Number 9413724.

obtained. A quantitative study of the smoothness obtained by different weights on the penalties is included.

The roughness penalties are derived and important properties such as convexity of the penalties are studied in Section II. Section III discusses how neighborhood structures are generated by the penalties and the use of these penalties in estimation problems. The divergence penalty is shown to be a discrete approximation of a continuous penalty due to Good and Gaskins in Section IV. The penalties are further extended to incorporate linear constraints in Section V. The applications are presented in Section VI, and the conclusions are in Section VII.

II. Roughness Penalties

Let $\mathbf{R}$ be the real line, $\mathbf{R}_+ = \{x \in \mathbf{R}: x > 0\}$, and $\mathbf{C}$ the complex numbers. Let $\mathbf{0}$ and $\mathbf{1}$ denote n-vectors of all zeros and ones, respectively, where the dimension n is clear from context. Let $V \in \{\mathbf{R}, \mathbf{R}_+, \mathbf{C}\}$. Scalars are in $\mathbf{R}$ or $\mathbf{C}$ as appropriate. If $\mathbf{x} \in V^n$, then $\mathbf{x}^\dagger$ equals $\mathbf{x}$ transposed if $V = \mathbf{R}$ or $\mathbf{R}_+$ and $\mathbf{x}^\dagger$ equals $\mathbf{x}$ complex conjugate transposed if $V = \mathbf{C}$. If it is clear that $V \neq \mathbf{C}$, then $\mathbf{x}^T$ may also be used for $\mathbf{x}$ transposed.

Definition 1: Let π_i be a permutation on n letters. A *shift* $S_i: V^n \to V^n$ (generated by the permutation π_i) is defined by

$$[S_i(\mathbf{x})]_k = x_{\pi_i(k)}. \tag{1}$$

The $n \times n$ permutation matrix $\mathbf{S}_i$ is defined by $\mathbf{S}_i\mathbf{x} = S_i(\mathbf{x})$.

Example 1: Let S_l be the left (circular) shift

$$[S_l(\mathbf{x})]_k = x_{(k+1) \bmod n}. \tag{2}$$

Example 2: Let S_r be the right (circular) shift

$$[S_r(\mathbf{x})]_k = x_{(k-1) \bmod n}. \tag{3}$$

The shifts are circular shifts in that no vector element is shifted off the lattice. This is clear in the two examples. In the following, all subscripts are assumed taken modulo n and the right side of (2) is written simply as x_{k+1}. There exists an integer $k(i)$ such that $\pi_i^{k(i)}$ is the identity map. The smallest such $k(i) > 0$ is the order of the shift. Clearly the two examples are of order n.

If instead of viewing V^n as being defined on the integers $\{1, 2, \cdots, n\}, V^n$ is viewed as being defined on a higher dimensional lattice, the shifts may be viewed as circular shifts on that lattice. An application to images is discussed later.

Definition 2: A function $d: V^n \times V^n \to \mathbf{R}_+ \cup \{0\}$ is called a *discrepancy measure* if $d(\mathbf{x}, \xi) = 0$ if and only if $\mathbf{x} = \xi$.

The types of discrepancy measures studied are motivated by the work of Csiszár [8] and Jones [18].

Example 3: Let $V = \mathbf{R}$. The least squares measure is defined by

$$d(\mathbf{x}, \xi) = \sum_{k=1}^{n} (x_k - \xi_k)^2. \tag{4}$$

Example 4: Let $V = \mathbf{R}_+$. The I-divergence is defined by

$$I(\mathbf{x}, \xi) = \sum_{k=1}^{n} \left[x_k \log \frac{x_k}{\xi_k} - x_k + \xi_k \right]. \tag{5}$$

Example 5: Let $V = \mathbf{R}_+$. The Itakura–Saito distance is defined by

$$d(\mathbf{x}, \xi) = \sum_{k=1}^{n} \left[-\log \frac{x_k}{\xi_k} - 1 + \frac{x_k}{\xi_k} \right]. \tag{6}$$

The roughness penalty is constructed from shifts and a discrepancy measure.

Definition 3: A *roughness penalty* with respect to the shifts $S = \{S_1, S_2, \cdots, S_I\}$ is a mapping $\Phi: V^n \to \mathbf{R}_+ \cup \{0\}$ defined by

$$\Phi(\mathbf{x}) = \sum_{i=1}^{I} w_i d(\mathbf{x}, \mathbf{S}_i\mathbf{x}) \tag{7}$$

where $w_i > 0$ is the ith weight.

As a direct consequence of the properties of the discrepancy measure, $\Phi(\mathbf{x}) \geq 0$ and $\Phi(\mathbf{x}) = 0$ if and only if

$$\mathbf{S}_i\mathbf{x} = \mathbf{x}, \quad \text{for all } i = 1, 2, \cdots, I.$$

Example 6: Let $V = \mathbf{R}$. A roughness penalty with respect to S_l is

$$\Phi_l(\mathbf{x}) = \sum_{m=1}^{n} (x_m - x_{m+1})^2. \tag{8}$$

Example 7: Let $V = \mathbf{R}_+$. A roughness penalty with respect to S_l and S_r is

$$\begin{aligned}\Phi_I(\mathbf{x}) &= I(\mathbf{x}, \mathbf{S}_l\mathbf{x}) + I(\mathbf{x}, \mathbf{S}_r\mathbf{x}) \\ &= \sum_{m=1}^{n} \Bigg(x_m \log \frac{x_m}{x_{m-1}} - x_m + x_{m-1} \\ &\quad + x_m \log \frac{x_m}{x_{m+1}} - x_m + x_{m+1} \Bigg).\end{aligned} \tag{9}$$

This penalty is called a divergence penalty and plays a central role in the simulations. This penalty has the nice feature that it is defined only for $\mathbf{x} \in \mathbf{R}_+^n$ and it arises naturally in terms of shifts and the I-divergence. Since the shifts are circular, it may be rewritten in several ways including

$$\begin{aligned}\Phi_I(\mathbf{x}) = -\sum_{m=1}^{n} x_m[(\log x_{m+1} - \log x_m) \\ - (\log x_m - \log x_{m-1})].\end{aligned} \tag{10}$$

In this form, it is a weighted second difference of the logarithms. This may also be viewed as a discretization of the penalty due to Good and Gaskins as described in Section IV.

Example 8: The previous example may be extended to images by defining vertical and horizontal shifts. Let $\mathbf{x} \in \mathbf{R}_+^{NM}$. Define the vertical shift S_V by

$$[S_V(\mathbf{x})]_{l,m} = x_{l-1,m}, \qquad 1 \geq l \geq N, 1 \geq m \geq M. \tag{11}$$

Here and in (12), the two subscripts are taken modulo N and M, respectively. Similarly, the horizontal shift S_H is

$$[S_H(\mathbf{x})]_{l,m} = x_{l,m-1}, \qquad 1 \geq l \geq N, 1 \geq m \geq M. \tag{12}$$

Then, the divergence penalty with respect to shifts $S = \{S_V, S_V^{-1}, S_H, S_H^{-1}\}$ is

$$\begin{aligned}\Phi_I(\mathbf{x}) = I(\mathbf{x}, S_V(\mathbf{x})) + I(\mathbf{x}, S_V^{-1}(\mathbf{x})) + I(\mathbf{x}, S_H(\mathbf{x})) \\ + I(\mathbf{x}, S_H^{-1}(\mathbf{x})).\end{aligned} \tag{13}$$

An important property of penalties when used in estimation problems is convexity. The penalty $\Phi(\mathbf{x})$ is convex on V^n if

$$\Phi(\alpha\mathbf{x} + (1-\alpha)\xi) \geq \alpha\Phi(\mathbf{x}) + (1-\alpha)\Phi(\xi) \tag{14}$$

for all $\mathbf{x}, \xi \in V^n$ and all $0 \geq \alpha \geq 1$. A sufficient condition for Φ to be convex is that the Hessian of Φ

$$\mathbf{H}_\Phi(\mathbf{x}) = \nabla_{xx}^2 \Phi(\mathbf{x})$$

is nonnegative definite; in turn, this Hessian is nonnegative definite if the matrix of second partial derivatives of d

$$\mathbf{H}_d(\mathbf{x}, \xi) = \begin{bmatrix} \nabla_{xx}^2\, d(\mathbf{x}, \xi) & \nabla_{\xi x}^2\, d(\mathbf{x}, \xi) \\ \nabla_{x\xi}^2\, d(\mathbf{x}, \xi) & \nabla_{\xi\xi}^2\, d(\mathbf{x}, \xi) \end{bmatrix} \tag{15}$$

is nonnegative definite. Note that

$$\mathbf{H}_\Phi(\mathbf{x}) = \sum_{i=1}^{I} w_i [\mathbf{I}\,\mathbf{S}_i^T] \mathbf{H}_d(\mathbf{x}, \mathbf{S}_i \mathbf{x}) \begin{bmatrix} \mathbf{I} \\ \mathbf{S}_i \end{bmatrix} \tag{16}$$

so $\mathbf{H}_d$ being nonnegative definite is just a sufficient condition. Suppose that $d(\mathbf{x}, \xi) = d_1(\mathbf{x} - \xi)$; then

$$\mathbf{H}_d(\mathbf{x}, \xi) = \begin{bmatrix} \mathbf{I} \\ -\mathbf{I} \end{bmatrix} \mathbf{H}_{d_1}(\mathbf{x} - \xi) [\mathbf{I} \ -\mathbf{I}]. \tag{17}$$

In this case, $\mathbf{H}_{d_1}$ should be nonnegative definite. Throughout most of this paper, a special form for d is assumed.

Definition 4: A discrepancy measure d is *generated by* the scalar discrepancy measure h if

$$d(\mathbf{x}, \xi) = \sum_{m=1}^{n} h(x_m, \xi_m). \tag{18}$$

Each of the discrepancy measures discussed in this paper is generated by a scalar discrepancy measure. Penalties based on such measures mesh well with neighborhood structure in estimation problems, as discussed in the next section. The second derivatives of h then determine whether Φ is convex. Lange [21] uses penalties generated by $h(x, \xi) = v(x - \xi)$ for functions $v(\cdot)$ that have specific properties including convexity.

Lemma 1: Let $V = \mathbf{R}_+$ and let d be generated by h. If $h(x, \xi) = \xi f(x/\xi)$ for some function f that is twice differentiable, then Φ is convex if f is convex.

Proof: By direct computation, the second partial derivatives of h give

$$\mathbf{H}_h = \frac{1}{\xi} \ddot{f}(x/\xi) \begin{bmatrix} 1 \\ -x/\xi \end{bmatrix} [1 - x/\xi]. \tag{19}$$

Then, $\mathbf{H}_h$ is nonnegative definite as long as $\ddot{f}$ is nonnegative.□

An example of this lemma is given by the I-divergence. There, $f(x) = x \log x + 1 - x$, and $\ddot{f} = 1/x > 0$. Discrepancy functions of the form given in the lemma play an important role in information theory (see [1], [8], [9]). The Itakura–Saito distance [17] is not in this form and is not recommended. If $\mathbf{H}_h$ is the Hessian for

$$h(x, \xi) = -\log \frac{x}{\xi} - 1 + \frac{x}{\xi} \tag{20}$$

then one eigenvalue of $\mathbf{H}_h$ is positive, and one is negative. In fact, a more general result is the following.

Lemma 2: Let $V = \mathbf{R}_+$. If $h(x, \xi) = f(x/\xi)$ for some f that is twice differentiable with $\ddot{f} \neq 0$, then the matrix $\mathbf{H}_h$ has one positive and one negative eigenvalue.

The proof is straightforward and omitted. This lemma is the primary motivation for ruling out the use of discrepancy measures like the Itakura–Saito distance to define roughness penalties.

III. Neighborhood Structures and Penalized Estimation

The penalties described above induce neighborhood structures in the same way as Markov random field priors. The terminology presented next follows the work of Besag [2], [3]. Throughout this section, assume that d is generated by h, that Φ is a roughness penalty with respect to the shifts $S = \{S_i\}$, and that π_i is the permutation on n letters corresponding to the shift S_i. When thought of as a lattice, the nth component of $\mathbf{x}$ may be referred to as a lattice site.

Definition 5: The *neighborhood* of site k is the set $N(k) = \{l : l \neq k, l = \pi_i^{\pm 1}(k), i = 1, \cdots, I\}$. The *neighbors* of x_k are the entries in the set $\{x_l : l \in N(k)\}$.

Thus, the neighborhood consists of all sites that the kth site may be mapped to or that are mapped to the kth site by the shifts $\{S_i\}$ (except the kth site itself). The neighbors are the components of $\mathbf{x}$ in the neighborhood of k.

Definition 6: A *coding set* C is a set of sites such that no two sites in the set are neighbors. If a family of coding sets $\{C_1, C_2, \cdots, C_J\}$ forms a partition of $\{1, 2, \cdots, n\}$, the labeling of sites by the integers $\{1, 2, \cdots, J\}$ according to their coding sets is called a *coloring*. A *minimal coloring* is a coloring with J being the minimum possible.

One motivation for this study is the use of these penalties in maximum-likelihood estimation problems. An iterative algorithm based on coding sets that is a generalized EM algorithm is presented.

Suppose that $\mathbf{y} \in V^M$ is measured. Given $\mathbf{x}$, the probability density function for $\mathbf{y}$ is $f(\mathbf{y}|\mathbf{x})$. Assume that for all $\mathbf{y}$ there is an $\mathbf{x} \in V^n$ such that $f(\mathbf{y}|\mathbf{x}) > 0$. Then, the maximum penalized likelihood problem is to find the $\mathbf{x} \in V^n$ that maximizes

$$l(\mathbf{x}) = \log f(\mathbf{y}|\mathbf{x}) - \alpha\Phi(\mathbf{x}). \tag{21}$$

In this problem, α is the weight given to the penalty; larger values of α give higher weight to the penalty and induce more smoothing in the estimate.

The connection to prior probabilities follows from (21). If $f_x(\mathbf{x})$ is a prior probability density function on $\mathbf{x}$, then the log-likelihood function is

$$l_1(\mathbf{x}) = \log f(\mathbf{y}|\mathbf{x}) + \log f_x(\mathbf{x}). \tag{22}$$

It is clear from (22) that if

$$f_x(\mathbf{x}) = \frac{1}{Z}\exp\left[-\alpha\Phi(\mathbf{x})\right] \tag{23}$$

where

$$Z = \int \exp\left[-\alpha\Phi(\mathbf{x})\right] d\mathbf{x} \tag{24}$$

then the penalty is equivalent to the prior. For least squares discrepancy measures, the penalty is equivalent to a Gauss–Markov random field prior. For other discrepancy measures, the equivalent prior may not take on a common form. For this reason, it may be easier to motivate the use of (21) by using a penalty than by using a prior probability density function.

Let the sites be colored according to the coding sets $\{C_1, C_2, \cdots, C_J\}$. Let $P_j = \{x_k : k \in C_j\}$. Let $\mathbf{x}^0$ be an initial guess for the maximum penalized likelihood estimate. Then, we have the following iterative algorithm for finding the estimate based on the coloring. Each iteration of the algorithm has J steps. In the jth step of iteration p, fix all entries of $\mathbf{x}$ at their most recent estimates except those in P_j. Maximize $l(\mathbf{x})$ over the entries of $\mathbf{x}$ in P_j. If $j < J$, go to step $j+1$, otherwise go to step one of iteration $p+1$.

Note that this algorithm is based on the algorithm proposed by Besag [3]. It is convenient for implementation on parallel machines where a subvector of $\mathbf{x}$ may be updated all at once. Since the penalty neighborhood structure is used to determine the coding sets, the maximization of $-\alpha\Phi(\mathbf{x})$ over one coding set results in independent maximizations. That is

$$\Phi(\mathbf{x}) = \sum_{k\in C_j} g(x_k; x_l, l \in N(k)) + \text{other terms} \tag{25}$$

where

$$\begin{aligned} g(x_k; x_l, l \in N(k)) \\ = \sum_{i=1}^{I} w_i[h(x_k, x_{\pi_i(k)}) + h(x_{\pi_i^{-1}(k)}, x_k)]. \end{aligned} \tag{26}$$

The other terms in (25) do not depend on P_j (also recall that $N(k) \cap C_j = \phi$, the null set, for $k \in C_j$). If, in addition, $f(\mathbf{y}|\mathbf{x})$ decomposes into a product of terms in the components of $\mathbf{x}$, then the algorithm would be fully parallel. One way to obtain a parallel implementation is through an EM algorithm.

Suppose that for $\mathbf{z} \in \mathbf{R}^n$ there are two probability density functions f_1 and f_2 such that

$$f(\mathbf{y}|\mathbf{x}) = \int f_2(\mathbf{y}|\mathbf{z}) f_1(\mathbf{z}|\mathbf{x})\, d\mathbf{z}. \tag{27}$$

The vector $\mathbf{z}$ may be considered to be a random vector such that the joint probability density function for $\mathbf{y}$ and $\mathbf{z}$ given $\mathbf{x}$ is $f_2(\mathbf{y}|\mathbf{z})f_1(\mathbf{z}|\mathbf{x})$. Furthermore, assume that

$$f_1(\mathbf{z}|\mathbf{x}) = \prod_{m=1}^{n} f_{1m}(z_m|x_m). \tag{28}$$

The set of possible values of $\mathbf{z}$ is called the *complete data space*. The set of possible values of $\mathbf{y}$ is called the *incomplete data space*.

Central to the EM algorithm is the function

$$Q(\mathbf{x}|\mathbf{x}') = E[\log f_1(\mathbf{z}|\mathbf{x})|\mathbf{y}, \mathbf{x}'] \tag{29}$$

the expected value of the complete data log-likelihood given the incomplete data $\mathbf{y}$ and the estimate $\mathbf{x}'$ for $\mathbf{x}$. The EM algorithm has two steps. In the expectation (E-) step, the quantity $Q(\mathbf{x}|\mathbf{x}^p)$ is computed. In the maximization (M-) step, $Q(\mathbf{x}|\mathbf{x}^p)$ is maximized over $\mathbf{x}$ to get $\mathbf{x}^{p+1}$.

It is worth noting that the complete data space as defined here does not contain the usual many-to-one map to the incomplete data space [10], [25], [37]. That mapping is implicitly contained in the conditional density function $f_2(\mathbf{y}|\mathbf{z})$. For more on this observation, see [12] and [15].

The new suggestion is that the maximization not be performed directly, but using a modified form of the iterative algorithm above. This makes the algorithm a generalized EM (GEM) algorithm [10]. The GEM algorithm has the following iterations given an initial guess $\mathbf{x}^0$ with $p = 0$: first, do the E-step (29) to get $Q(\mathbf{x}|\mathbf{x}^p)$; second, let $j = p + 1 \bmod J$; third, for each $k \in C_j$, maximize over x_k

$$\begin{aligned} Q'_k(x_k|\mathbf{x}^p) = {} & -\alpha g(x_k; x_l^p, l \in N(k)) \\ & + E[\log f_1 k(z_k|x_k)|\mathbf{y}, \mathbf{x}^p] \end{aligned} \tag{30}$$

fourth, increment p and return to the E-step. Equation (30) emphasizes the parallel nature of the maximization step. For each element of the coding set C_j, the maximization proceeds independently of all other elements of that set since the neighbors of x_k are not in P_j. The coding sets should be chosen wisely, ideally corresponding to a minimal coloring.

A set of tools have developed in the literature for analyzing the convergence of EM algorithms [10], [16], [25], [33], [37]. The following analysis relies on those results and convergence of this GEM algorithm is proven under conditions that guarantee an EM algorithm would converge plus an additional condition. For this analysis, assume V is either $\mathbf{R}$ or $\mathbf{R}_+$.

Convergence Assumptions:

1) For $Q(\mathbf{x}|\xi)$, let $\nabla_x Q(\mathbf{x}|\xi)$ and $\nabla_\xi Q(\mathbf{x}|\xi)$ denote the vectors of first partial derivatives of Q, and $\nabla^2_{xx} Q(\mathbf{x}|\xi)$, and $\nabla^2_{x\xi} Q(\mathbf{x}|\xi)$ denote matrices of second derivatives of Q. Assume that all of these derivatives exist and are continuous.
2) The set $\Omega(\mathbf{x}^0) = \{\mathbf{x} \in V^n : l(\mathbf{x}) \geq l(\mathbf{x}^0)\}$ is compact for $l(\mathbf{x}^0) > -\infty$.
3) Assume that $H_\Phi(\mathbf{x})$ exists and is continuous for $\mathbf{x} \in \Omega(\mathbf{x}^0)$.
4) $l(\mathbf{x})$ is continuously differentiable for $\mathbf{x} \in \Omega(\mathbf{x}^0)$.

5) Assume that for any sequence $\{\mathbf{x}^k\} \in \Omega(\mathbf{x}^0)$

$$\lim_{k\to\infty} Q(\mathbf{x}^{k+1}|\mathbf{x}^k) - \alpha\Phi(\mathbf{x}^{k+1}) - Q(\mathbf{x}^k|\mathbf{x}^k) + \alpha\Phi(\mathbf{x}^k) = 0 \quad (31)$$

implies

$$\lim_{k\to\infty} ||\mathbf{x}^{k+1} - \mathbf{x}^k|| = 0$$

for $||\cdot||$ an appropriate norm.

6) Let Q'_k be given by (30), let $k \in C_j$, and define the $n \times 1$ vector $\eta_k(\mathbf{x}, \mathbf{x}')$ to have lth entry equal to[1]

$$[\eta_k(\mathbf{x}, \mathbf{x}')]_l = \begin{cases} x_l, & \text{if } l \in C_m, \text{ for } m < j, \\ x'_l, & \text{if } l \in C_m, \text{ for } m \ge j. \end{cases}$$

Define $\mathbf{G}(\mathbf{x}, \mathbf{x}')$ to be an $n \times 1$ vector with kth entry

$$\frac{\partial Q'_k}{\partial x_k}(x_k|\eta_k(\mathbf{x}, \mathbf{x}')).$$

Note that by definition of the algorithm and if Assumptions 1–3 hold, $\mathbf{G}(\mathbf{x}^{(m+1)J}, \mathbf{x}^{mJ}) = 0$. Let $\mathbf{x}^*$ be such that $\mathbf{G}(\mathbf{x}^*, \mathbf{x}^*) = 0$. Assume that $\nabla_x \mathbf{G}(\mathbf{x}^*, \mathbf{x}^*)$ is negative definite and that the magnitude of the largest eigenvalue of

$$\nabla_x \mathbf{G}(\mathbf{x}^*, \mathbf{x}^*)^{-1} \nabla_{x'} \mathbf{G}(\mathbf{x}^*, \mathbf{x}^*) \quad (32)$$

(denote this eigenvalue by ρ) is less than one.

7) From the definition of $\mathbf{G}$ in Assumption 6, the GEM algorithm proposed becomes both an A-algorithm in Hero and Fessler's terminology [16] and a one-step stationary method in Ortega and Rheinholdt's terminology [27, p. 299]. Sufficient conditions for an iteration started at an arbitrary point in a fixed region to converge are given in [16, Theorem 1]. Let $||\cdot||: \mathbf{R}^n \to \mathbf{R}_+$ be a vector norm and define the induced matrix norm by $||\mathbf{A}|| = \sup_x (||\mathbf{A}\mathbf{x}||/||\mathbf{x}||)$. Let $\Theta \subset \Omega(\mathbf{x}^0)$ for some $\mathbf{x}^0$ such that $l(\mathbf{x}^0) > -\infty$ be defined as an arbitrary convex open set containing $\mathbf{x}^*$ from Assumption 6 such that:

   a) For all $\mathbf{x} \in \Theta$, if $\mathbf{x}^p = \mathbf{x}$ for any p in the GEM algorithm, then $\mathbf{x}^{p+1} \in \Theta$.

   b) For any $\mathbf{x}, \tilde{\mathbf{x}} \in \Theta$ such that $\mathbf{G}(\tilde{\mathbf{x}}, \mathbf{x}) = 0$, define $\Delta(\tilde{\mathbf{x}}, \mathbf{x})$ as the box in $\mathbf{R}^{2n}$ with corners at $(\mathbf{x}^*, \mathbf{x}^*)$ and $(\tilde{\mathbf{x}}, \mathbf{x})$ and with sides parallel to the $2n$ axes; by the mean value theorem, there exists $(\mathbf{x}_1, \mathbf{x}_2) \in \Delta(\tilde{\mathbf{x}}, \mathbf{x})$ such that

$$\mathbf{x}^* - \tilde{\mathbf{x}} = -\nabla_x \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)^{-1} \nabla_{x'} \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)(\mathbf{x}^* - \mathbf{x}) \quad (33)$$

   then assume $\nabla_x \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)$ is negative definite and $||\nabla_x \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)^{-1} \nabla_{x'} \mathbf{G}(\mathbf{x}_1, \mathbf{x}_2)|| \le \alpha^2 < 1$ for all $\mathbf{x} \in \Theta$.

Lemma 3: Let $V \in \{\mathbf{R}, \mathbf{R}_+\}$. Let $\mathbf{x}^0$ be such that $l(\mathbf{x}^0) > -\infty$ and suppose the sequence $\{\mathbf{x}^p\}$ is given by the GEM algorithm described above. Under the convergence Assumptions 1–4,

$$l(\mathbf{x}^{p+1}) \ge l(\mathbf{x}^p) \quad (34)$$

with equality if and only if x_k^p maximizes (30) for each $k \in C_j, j = p + 1 \bmod J$. If $\lim_{p\to\infty} ||\mathbf{x}^{p+1} - \mathbf{x}^p|| = 0$ (Assumption 5 implies this), then the limit points of the algorithm form a connected and compact set and limit points are stationary points of l. Under convergence Assumptions 1–4 and 6, there is an open ball $B(\mathbf{x}^*)$ in V^n containing x^* such that all GEM sequences starting at $\mathbf{x}_0 \in B(\mathbf{x}^*)$ converge to x^* at linear rate. The convergence factor for the subsequence $\{\mathbf{x}^{mJ}\}$ is ρ. If in addition Assumption 7 holds, there is only one limit point for any $\mathbf{x}^0 \in \Theta$, and the GEM algorithm converges to that point at linear rate.

The proof is in the appendix. The condition on the matrix (32) is the additional requirement needed for convergence of this algorithm. Note that $\mathbf{G}$ has a block triangular structure in its dependence on $\mathbf{x}$ and $\mathbf{x}'$. This is due to the GEM algorithm only updating subblocks of the estimate for x at each iteration. The convergence is established for the subsequence of the GEM algorithm consisting of every Jth element through the use of $\mathbf{G}$. The convergence of the entire sequence then follows from the fact that the iterations monotonically increase $l(\mathbf{x})$. The convergence factor is smaller $(\rho^{1/J})$ than for the traditional EM algorithm (ρ) due to only updating subblocks of x. The convergence depends on the negative definiteness of the matrix $\nabla_x G(\cdot, \cdot)$ that has entries that depend on the second derivatives of Q and of Φ. Thus, while the convexity of Φ is not used directly in the proof, it makes the satisfaction of this negative definiteness condition easier.

IV. Relationship to Good's Roughness

The divergence penalty as presented in (13) and (16) is closely related to a penalty proposed by Good and Gaskins [13]. Let $x(t) \in \mathbf{L}_1(\mathbf{R})$ be a probability density function to be estimated. Then, the penalty proposed in [13] is given by

$$\Phi_c(x) = \int x\left(\frac{\partial \log x}{\partial t}\right)^2 dt = \int \frac{1}{x}\left(\frac{\partial x}{\partial t}\right)^2 dt. \quad (35)$$

Defining $\gamma(t) = \sqrt{x(t)}$, this penalty may be rewritten as

$$\Phi_c(\gamma^2) = 4\int \left(\frac{\partial \gamma}{\partial t}\right)^2 dt. \quad (36)$$

This forms the basis for the implementation in [4], [23] for images, which is used in Section VI for comparison to the proposed implementation. In that implementation, the integral in (36) is approximated by a discrete sum and the optimization is performed over a discrete version of $\gamma(t)$ rather than directly over $x(t)$.

[1] Notice that if k and k' are both in C_j, then $\eta_k(\mathbf{x}, \mathbf{x}') = \eta_{k'}(\mathbf{x}, \mathbf{x}')$. Also, if $k \in C_1$, then $\eta_k(\mathbf{x}, \mathbf{x}') = \mathbf{x}'$.

As noted in [13], $\Phi_c(x)$ is the Fisher information for estimating the mean of an otherwise known density function. Thus, using the same arguments that lead to the conventional derivation of the Fisher information [36, p. 66], Φ_c may be rewritten as

$$\Phi_c(x) = -\int x\left(\frac{\partial^2 \log x}{\partial t^2}\right) dt. \tag{37}$$

This motivates the use of a discretization of (37) directly. Approximate $x(t)$ by a piecewise constant function equal to x_k over intervals $[t_0 + (k-1)\Delta, t_0 + k\Delta), n = 1, 2, \cdots, n$. Approximate the second derivative by

$$\frac{\partial^2 \log x}{\partial t^2} \approx \frac{1}{\Delta^2}[(\log x_{m+1} - \log x_m) - (\log x_m - \log x_{m-1})]. \tag{38}$$

Then, the divergence penalty as written in (13) equals Δ times a discretization of (37) using (38). This approach avoids the need to use $\gamma(t)$. The positivity of the estimates for x_k is guaranteed by the divergence only being defined for positive arguments.

Note that while the derivation of (37) given above assumes that $x(t)$ is a probability density function, the derivation carries through under more general conditions. Expanding the integrand in (37) yields $\ddot{x} - (1/x)\dot{x}^2$. Assume that $x(t)$ is nonnegative and integrable. Then, as long as $\ddot{x}$ exists and is integrable, its integral equals 0 and (37) equals (35).[2] This justifies the use of the divergence penalty as a discretization of the penalty from Good and Gaskins. Further motivation for the continuous penalty (35) is that it is the Fisher information for position estimation for a Poisson process with integrable intensity function $x(t)$, assuming count-record data [31, p. 80].

V. Linear Constraints

In this section, the estimate for $\mathbf{x}$ is further constrained to satisfy linear equations. Denote by L the subset of V^n

$$L = \{\mathbf{x} \in V^n : \mathbf{Bx} = \mathbf{b}\}, \tag{39}$$

where $\mathbf{B} \in \mathbf{R}^{k \times n}$ (resp., $\mathbf{C}^{k \times n}$) and $\mathbf{b} \in \mathbf{R}^k$ (resp., $\mathbf{C}^k$). Throughout, we assume that L is nonempty, $k < n$, and $\mathbf{B}$ is of full rank. We may also denote L by $L(\mathbf{B}, \mathbf{b})$.

Example 9: Let $V = \mathbf{R}_+$. The set $L = \{\mathbf{x} \in \mathbf{R}_+^n : \mathbf{1}^T\mathbf{x} = 1\}$ is the set of probability vectors.

The shift S_i is *compatible with* L if $S_i(L) = L$. Note that S_i is compatible with L if and only if $L(\mathbf{B}, \mathbf{b}) = L(\mathbf{B}S_i, \mathbf{b}) = L(\mathbf{B}S_i^m, \mathbf{b})$, for all m. In order to explore the implications of this further, the structure of L is further defined. L is a linear variety [22, p. 16] (or the intersection of a linear variety and $\mathbf{R}_+^n$ if $V = \mathbf{R}_+$). In fact

$$L = \{\mathbf{x} + N(\mathbf{B})\} \cap V^n \tag{40}$$

where $\mathbf{x}$ is an arbitrary vector from L and $N(\mathbf{B})$ is the null space of $\mathbf{B}$ (in $\mathbf{R}^n$ or $\mathbf{C}^n$). Since $S_iL = L, \mathbf{S}_i\mathbf{x} \in L$, and the null space of $\mathbf{B}$ is an invariant subspace for $\mathbf{S}_i$. By the projection theorem [22, p. 51], there is a unique element $\mathbf{x}_o$ of $\mathbf{R}^n$ (or $\mathbf{C}^n$) of minimum Euclidean norm such that (40) is satisfied with $\mathbf{x} = \mathbf{x}_o$. Furthermore, $\mathbf{x}_o$ is orthogonal to all elements of $N(\mathbf{B}); \mathbf{x}_o = \mathbf{0}$ if and only if $\mathbf{b} = \mathbf{0}$. In that case, the rows of $\mathbf{B}$ are linear combinations of k eigenvectors of $\mathbf{S}_i$. The remaining $n - k$ eigenvectors of $\mathbf{S}_i$ span the null space of $\mathbf{B}$ (no distinction is made between the left and right eigenvectors of $\mathbf{S}_i$ since they are complex conjugate transposes of each other). Now suppose $\mathbf{x}_o \neq \mathbf{0}$. Then it is still true that $n - k$ eigenvectors of $\mathbf{S}_i$ span the null space of $\mathbf{B}$. The vector $\mathbf{x}_o$ is a linear combination of the other k eigenvectors of $\mathbf{S}_i$. Now the claim is that $\mathbf{x}_o$ is a multiple of an eigenvector of $\mathbf{S}_i$ with eigenvalue one. This holds because $\mathbf{S}_i\mathbf{x}_o$ has the same norm as $\mathbf{x}_o$ and $\mathbf{S}_i\mathbf{x}_o \in \{\mathbf{x}_o + N(\mathbf{B})\}$. If $\mathbf{S}_i\mathbf{x}_o \neq \mathbf{x}_o$, then this contradicts the uniqueness of the minimum norm element from the projection theorem. Note that if $V = \mathbf{R}_+, \mathbf{x}_o$ is not necessarily in L.

Definition 8: The pair $(\mathbf{B}, \mathbf{b})$ is said to be in *standard form* if either

1) $\mathbf{b} = [1 \quad 0 \quad \cdots \quad 0]^T$; the rows of $\mathbf{B}$ are orthogonal to each other; and all but the first row of $\mathbf{B}$ have unit Euclidean norm or
2) $\mathbf{b} = \mathbf{0}$ and the rows of $\mathbf{B}$ are orthonormal.

Lemma 4: $L(\mathbf{B}, \mathbf{b})$ is equivalent to $L(\mathbf{B}_s, \mathbf{b}_s)$, where $(\mathbf{B}_s, \mathbf{b}_s)$ is in standard form.

Proof: If $\mathbf{U}$ is an invertible $k \times k$ matrix, then $L(\mathbf{B}, \mathbf{b}) = L(\mathbf{UB}, \mathbf{Ub})$. If $\mathbf{b} = 0$, select $\mathbf{U}$ (through elementary row operations) to make the rows of $\mathbf{B}$ orthonormal. If $\mathbf{b} \neq 0$, let $\mathbf{U} = \mathbf{U}_1\mathbf{U}_2\mathbf{U}_3$. Select $\mathbf{U}_3$ to transform $\mathbf{b}$ to standard form. Select $\mathbf{U}_2$ (using elementary row operations) to remove the linear dependence of the first row on the remaining rows. Select $\mathbf{U}_1$ to make the remaining $k - 1$ rows orthonormal. □

From the discussion above we have the following lemma.

Lemma 5: If $(\mathbf{B}, \mathbf{b})$ is in standard form with $\mathbf{b} \neq 0$, and S_i is compatible with $\mathbf{B}$, then the first row of $\mathbf{B}$ is an eigenvector of $\mathbf{S}_i$ with eigenvalue one. Let $(i_1, i_2, \cdots, i_j)$ be a cycle in the decomposition of π_i into disjoint cycles. Then, the $i_1, i_2, \cdots$, and i_j entries of the first row of $\mathbf{B}$ are equal.

Proof: Let the first row of $\mathbf{B}$ be denoted β. The minimum norm solution to $\mathbf{Bx} = \mathbf{b}$ is given by

$$\mathbf{x}_o = \mathbf{B}^\dagger(\mathbf{BB}^\dagger)^{-1}\mathbf{b}. \tag{41}$$

Since the rows of $\mathbf{B}$ are orthogonal, $\mathbf{BB}^\dagger$ is diagonal implying that $\mathbf{x}_o$ is proportional to $\beta^\dagger$. This means that β is an eigenvector of $\mathbf{S}_i$ with eigenvalue one. The fact that the row elements are equal within one cycle follows by noticing that $\mathbf{S}_i$ may be decomposed as a direct sum of matrices corresponding to each cycle. Then, the eigenvector for the matrix corresponding to a given cycle with eigenvalue one has entries equal over that cycle. □

The use of linear constraints in estimation problems complicates the GEM algorithm presented in the previous section. In particular, the coding sets are no longer decoupled, being coupled by the linear constraints. For the algorithm above to be useful an additional step must be included at the end of an iteration to promote mixing between the coding sets.

[2] As pointed out by an anonymous reviewer, an alternate condition for equality of (35) and (37) is that $\dot{x}(a) = \dot{x}(b)$, and that the limits on the integral are from a to b.

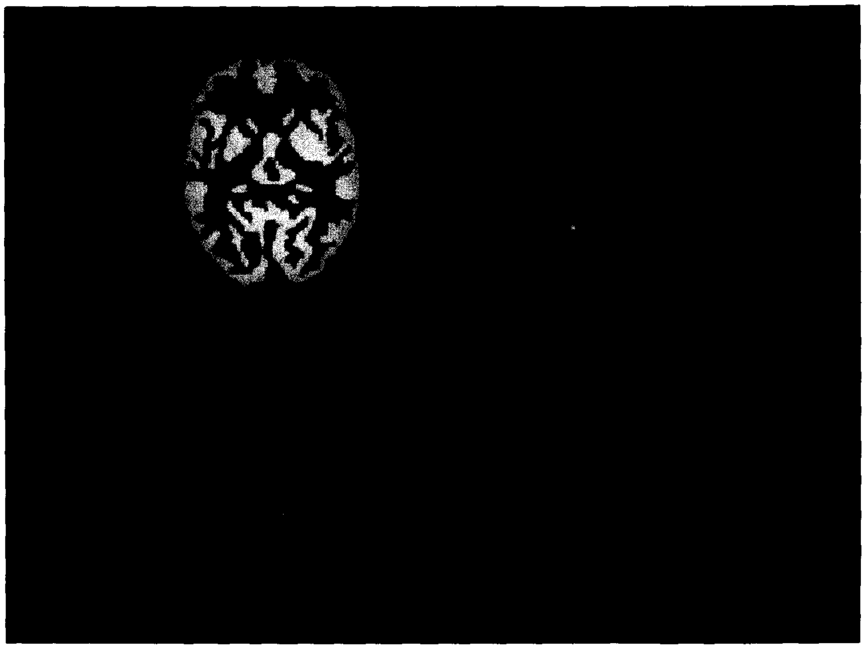


Fig. 1. Photograph of the display on an AMT DAP 510 for the emission tomography simulation. The upper left is a digitized Hoffman brain phantom. The upper right shows the image obtained using the divergence penalty discussed in the paper with $\alpha = 0.0025$. The lower left shows the unpenalized estimate. The lower right shows a cross section through the images.

VI. Applications

A. Application to Emission Tomographic Imaging

The problem of estimating a radioactivity distribution in the presence of photon attenuation and background radiation is presented by Snyder and Miller in [31]. The derivation here assumes familiarity with the basic concepts in [31]. Our emphasis is on a discrete version of the problem.

Let $N(j)$ be the available data at location j. $N(j)$ is Poisson distributed with mean $\mu(j)$

$$\mu(j) = \mu_0(j) + \sum_{i=1}^{I} \beta(j|i)p(j|i)\lambda_i. \tag{42}$$

The term $\mu_0(j)$ accounts for background intensity, $\beta(j|i)$ is a spatially dependent attenuation, and $p(j|i)$ is the probability that an event at location i in the input space is measured at location j in the output space. The imaging problem is to recover the intensities $\{\lambda_i, i = 1, 2, \cdots, I\}$ from the data $\{N(j), j = 1, 2, \cdots, J\}$. The log-likelihood function is given by

$$\begin{aligned} L(\lambda) = &-\sum_{i=1}^{I} \overline{\beta}(i)\lambda_i \\ &+ \sum_{j=1}^{J} \ln \left[\sum_{i=1}^{I} \beta(j|i)p(j|i)\lambda_i + \mu_0(j)\right] N(j) \end{aligned} \tag{43}$$

where $\overline{\beta}(i) = \Sigma_{j=1}^{J} \beta(j|i)p(j|i)$. The penalized log-likelihood is maximized over λ_i using the GEM algorithm, with the divergence penalty (13).

Simulations were run on an AMT DAP 510. We compared our penalty method with existing Good's roughness methods from [4], [23], [29]. The factor α was chosen to be equivalent for the two methods. From Figs. 1 and 2 and other simulations our method is seen to yield almost identical results to the implementations from [4], [23], [29]. This is expected since they are two implementations of the same penalty.

B. Application to Radar Imaging

As described in [26], [28], and [32], diffuse radar targets are commonly modeled as having a reflectivity density that is an uncorrelated complex Gaussian random process. The scattering function for the reflectivity density models the intensity of the reflections. If the estimate is piecewise constant, the scattering function estimation problem is one of estimating the variances of the vector of reflectivities. Experimental data have been collected from a rough rotating sphere placed on a pedestal in a compact radar range [28]. Under the assumptions in [28], a noisy estimate of the reflectivities may be obtained. More precisely, the measured data are modeled by

$$\mathbf{y} = \Gamma^{\dagger}\mathbf{c} + \mathbf{w} \tag{44}$$

where $\mathbf{w}$ is a noise vector of independent and identically distributed zero-mean complex Gaussian random variables; $\mathbf{c}$ is a vector of independent complex-valued Gaussian random

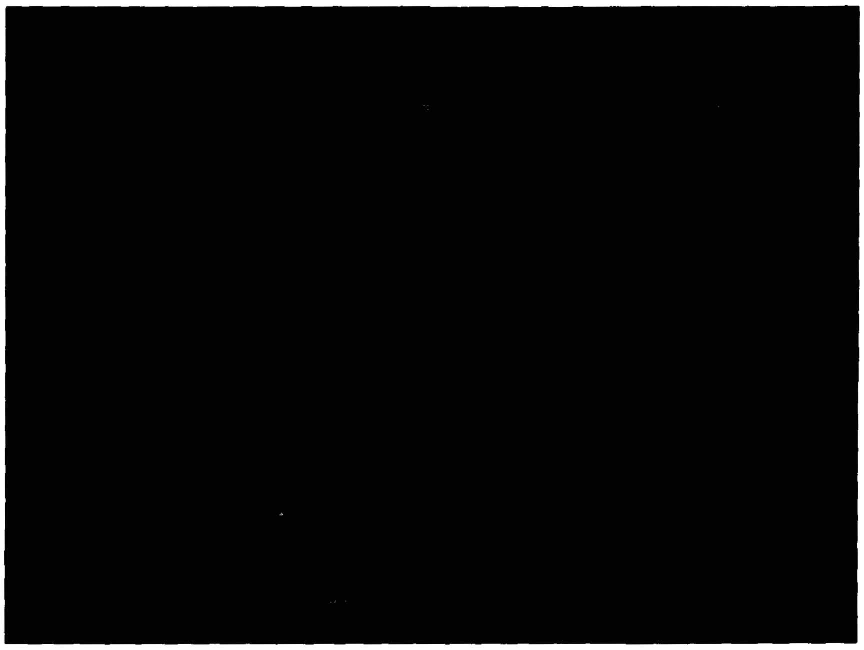


Fig. 2. The upper right and lower left are the same as in Fig. 1. The upper left shows the image obtained using the implementation of Good's penalty as presented in [4] and [23]. The two upper images are almost identical and in the lower right, the cross sections lie on top of each other.

variables representing the reflectivities; and Γ is a matrix of samples of the complex envelope of the transmitted signal times complex exponentials. The matrix Γ is assumed to be unitary so $\Gamma\Gamma^\dagger = \mathbf{I}$. Thus

$$\Gamma \mathbf{y} = \mathbf{c} + \tilde{\mathbf{w}} \tag{45}$$

where $\tilde{\mathbf{w}} = \Gamma\mathbf{w}$ has the same distribution as $\mathbf{w}$. A sufficient statistic consists of the entries of $\Gamma\mathbf{y}$ magnitude squared, which are denoted by z_{ij}.

The z_{ij} are then independent, exponentially distributed random variables with means $x_{ij} + N_0$, where x_{ij} is the discrete approximation to the scattering function and N_0 is the white noise intensity (measured as 0.0023, see [28]). Since the data are independent, there is no need to use the EM algorithm and the log-likelihood minus the penalty (21) is minimized for various values of α to attempt to measure the amount of smoothing introduced by the penalty. The divergence penalty (13) is used. Shown in Fig. 3 is a set of images produced for a range of values of α for one of the experimental data sets. The signal-to-noise-ratio (total signal energy to total noise energy) was estimated to be 0 dB [28]. As can be seen in Fig. 3, the images become smoother for larger values of α. Fig. 3(b) and (c) are taken from near the corner of the tradeoff curve in Fig. 4.

Fig. 4 shows the tradeoff curve for the penalized estimation problem. This curve plots the value of the log-likelihood at the maximum penalized likelihood estimate versus the value of the penalty at that estimate as the weight α varies. This curve quantifies the tradeoff in several senses. First, maximizing likelihood subject to a constraint on the value of the roughness penalty corresponds to finding the values of x_{ij} corresponding to the point on the curve for that penalty value. Second, all possible estimation procedures yield likelihoods and penalty values below the curve. Third, for a fixed value of likelihood, the curve defines the smallest possible penalty value. Fourth, the parametric nature of the curve can be examined. Let the value of the log-likelihood for a given α be $l(\alpha)$, the penalty value be $\phi(\alpha)$, and the penalized likelihood value be $v(\alpha) = l(\alpha) - \alpha\phi(\alpha)$. Then, it is easy to show that

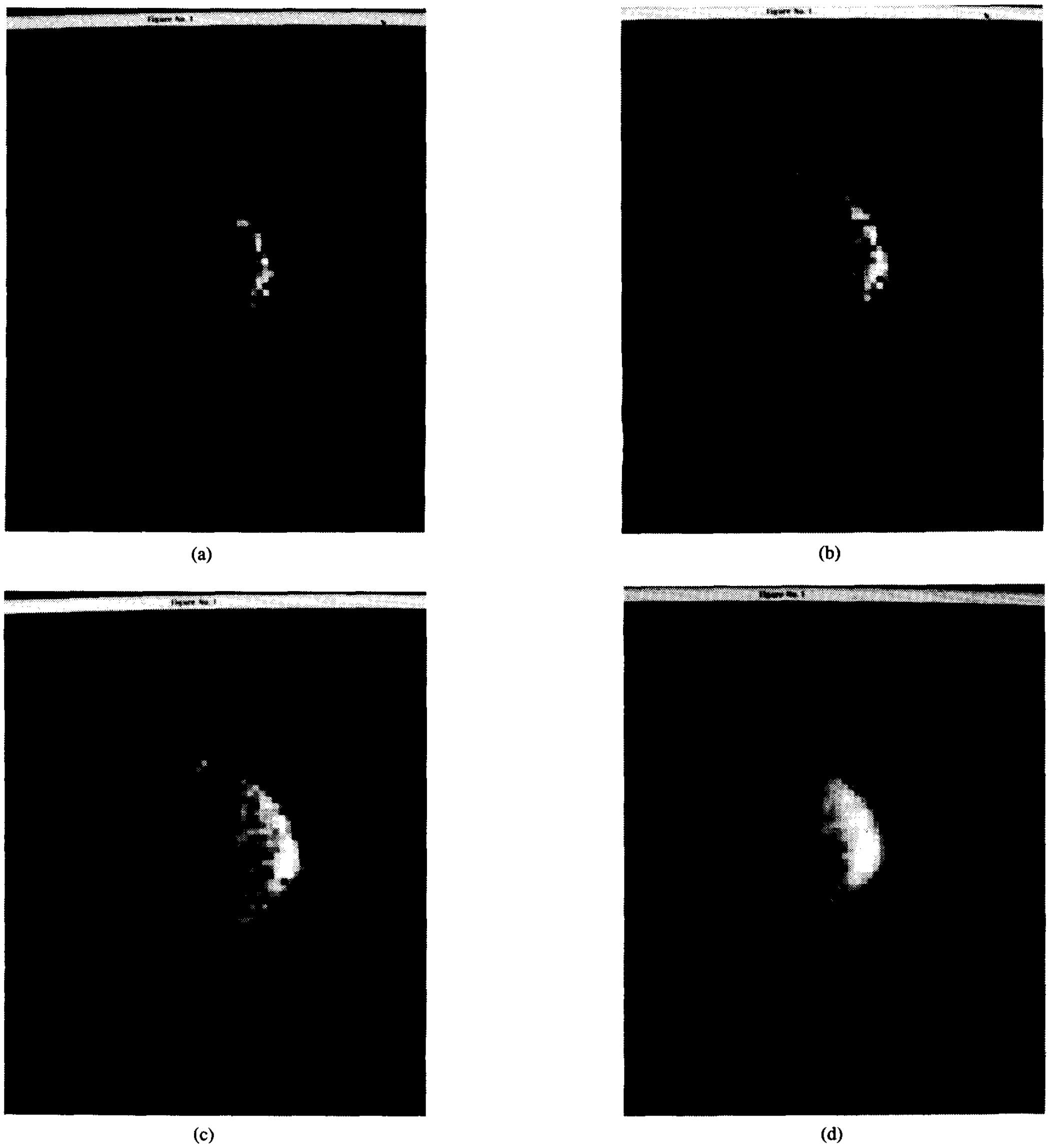
$$\frac{dl}{d\Phi}(\alpha) = \alpha \tag{46}$$

and $dv/d\alpha = -\phi(\alpha) < 0$. Thus, the penalized likelihood decreases as the weight on the penalty increases. The value of the likelihood and the value of the penalty decrease as the weight on the penalty increases. The tradeoff curve is concave because the increasing value of α for smaller $l(\alpha)$ implies an increasing slope.

VII. Conclusions

A class of roughness penalties has been proposed for finite dimensional vector spaces. The penalties are formed from two quantities: shifts and discrepancy measures. Discrepancies between the vectors and shifted copies of themselves are penalized. If the coordinates in the vector correspond to lattice locations on a finite lattice, then the shifts correspond to cyclic shifts of the lattice. This is analogous to periodic boundary conditions on the lattice.

An iterative algorithm has been proposed along with its use in an EM algorithm. This algorithm is based on the use of


(a)

(b)

(c)

(d)

Fig. 3. Shown are four penalized estimates of images of the rotating sphere using real data. Each image is 64 × 64 and displayed on a dB scale with the peak value shifted to the highest value on a linear gray scale. The weights in the images are: 0.0512 for (a), 0.4096 for (b), 3.277 for (c), and 26.21 for (d). The corresponding values of penalty and likelihood are shown in Fig. 4.

coding sets to update only part of the estimated vector at each step. Convergence is proven under certain assumptions. In the process, a recent generalization of the concept of a complete data space was used (see [12], [15]). The penalty may be thought of as a discretization of a penalty on function spaces proposed by Good and Gaskins [13]. It avoids the need to use the square root of the function to be estimated, enforcing positivity in a natural way.

The extension to the nonperiodic case is straightforward, but slightly less mathematically pleasing. There are lattice sites shifted off the lattice and the discrepancy measure is defined on the reduced dimensional vector representing the shifted sites that overlap the original lattice. The use of the EM algorithm goes through for this case with the modification that not all g functions are the same.

Appendix

Proof of Lemma 3: Taking the usual approach [10], [16], [25], write

$$\log f(\mathbf{y}|\mathbf{x}) = -\log f(\mathbf{z}|\mathbf{y},\mathbf{x}) + \log f_2(\mathbf{y}|\mathbf{z}) + \log f_1(\mathbf{z}|\mathbf{x}). \tag{A1}$$

Since the left hand side does not depend on $\mathbf{z}$, multiplying by $f(\mathbf{z}|\mathbf{y},\mathbf{x}^p)$ and integrating out $\mathbf{z}$ yields

$$\log f(\mathbf{y}|\mathbf{x}) = H(\mathbf{x}|\mathbf{x}^p) + Q(\mathbf{x}|\mathbf{x}^p) + E[\log f(\mathbf{y}|\mathbf{z})|\mathbf{y},\mathbf{x}^p] \tag{A2}$$

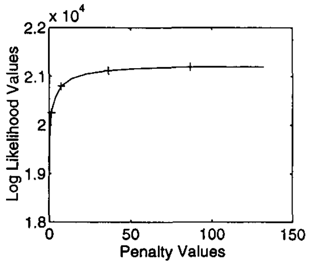


Fig. 4. Tradeoff curve for penalized estimation. This curve plots log likelihood versus penalty values as the weight α varies. Each point on the curve is equivalently obtained by maximizing likelihood subject to a maximum value for the penalty. The "+" marks on the curve correspond to the images in Fig. 3.

where

$$H(\mathbf{x}|\mathbf{x}^p) = E[-\log f(\mathbf{z}|\mathbf{y},\mathbf{x})|\mathbf{y},\mathbf{x}^p]. \tag{A3}$$

Then

$$\begin{aligned} l(\mathbf{x}^{p+1}) - l(\mathbf{x}^p) &= Q(\mathbf{x}^{p+1}|\mathbf{x}^p) - Q(\mathbf{x}^p|\mathbf{x}^p) - \alpha\Phi(\mathbf{x}^{p+1}) \\ &\quad + \alpha\Phi(\mathbf{x}^p) + D(\mathbf{x}^{p+1}|\mathbf{x}^p) \end{aligned} \tag{A4}$$

where

$$\begin{aligned} D(\mathbf{x}^{p+1}|\mathbf{x}^p) &= H(\mathbf{x}^{p+1}|\mathbf{x}^p) - H(\mathbf{x}^p|\mathbf{x}^p) \\ &= E\left[\log \frac{f(\mathbf{z}|\mathbf{y},\mathbf{x}^p)}{f(\mathbf{z}|\mathbf{y},\mathbf{x}^{p+1})}|\mathbf{y},\mathbf{x}^p\right] \end{aligned} \tag{A5}$$

is the divergence between the conditional densities of $\mathbf{z}$ given $\mathbf{x}^p$ and $\mathbf{x}^{p+1}$. Note that

$$D(\mathbf{x}^{p+1}|\mathbf{x}^p) \geq 0 \tag{A6}$$

with equality if and only if

$$f(\mathbf{z}|\mathbf{y},\mathbf{x}^p) = f(\mathbf{z}|\mathbf{y},\mathbf{x}^{p+1}) \tag{A7}$$

almost everywhere. Thus, to show (34) it is sufficient to show the remaining terms in (A4) are nonnegative. This follows from the maximization (30):

$$\begin{aligned} Q(\mathbf{x}^{p+1}|\mathbf{x}^p) &- Q(\mathbf{x}^p|\mathbf{x}^p) - \alpha\Phi(\mathbf{x}^{p+1}) + \alpha\Phi(\mathbf{x}^p) \\ &= \sum_{k\in C_j} Q'_k(x_k^{p+1}|\mathbf{x}^p) - Q'_k(x_k^p|\mathbf{x}^p) \end{aligned} \tag{A8}$$

since (30) is maximized for each x_k, each term on the right side of (A8) is nonnegative. (A8) equals 0 only if x_k^p maximizes $Q'_k(x_k|\mathbf{x}^p)$.

$l(\mathbf{x})$ is bounded above because it is a continuous function on a compact set. Since $l(\mathbf{x})$ is bounded above and $l(\mathbf{x}^p)$ is nondecreasing, there is a limiting value, $l^* = \lim_{p\to\infty} l(\mathbf{x}^p)$. That the limit points form a connected and compact set and are stationary points follows from Wu [37]. The existence of an open ball such that all sequences starting in that ball converge and the derivation of the convergence rate follow from [27, Theorem 10.3.5]. The convergence for all sequences starting from $\mathbf{x}^0 \in \Theta$ follows from the fact that $||\mathbf{x}^{(m+1)J} - \mathbf{x}^*|| < \alpha^2||\mathbf{x}^{mJ} - \mathbf{x}^*||$. For m large enough, $\mathbf{x}^{mJ} \in B(\mathbf{x}^*)$ and the sequence converges at a linear rate determined by ρ from [27, Theorem 10.3.5]. See also [16, Theorem 1]. □

Acknowledgment

Some simulations were run by Asif Chinwalla, a graduate student. The data for the radar images were obtained from McDonnell Douglas. The simulations using the penalty from [4], [23] shown in Figs. 1 and 2 were implemented by M. I. Miller and coworkers. The paper has benefited from the insightful comments of reviewers.

References

[1] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *J. Royal Stat. Soc. Ser. B*, vol. 28, pp. 131-142, 1966.

[2] J. Besag, "Spatial interaction and the statistical analysis of lattice systems (with discussion)," *J. Royal Stat. Soc. Ser. B*, vol. 36, pp. 192-236, 1974.

[3] ——, "On the statistical analysis of dirty pictures (with discussion)," *J. Royal Stat. Soc. Ser. B*, vol. 48, pp. 259-302, 1986.

[4] C. S. Butler and M. I. Miller, "Maximum *a posteriori* estimation for SPECT using regularization techniques on massively parallel computers," *IEEE Trans. Med. Imag.*, vol. 12, pp. 84-89, 1993.

[5] C. L. Byrne, "Iterative image reconstruction algorithms based on cross-entropy minimization," *IEEE Trans. Image Processing*, vol. 2, pp. 96-103, Jan. 1993.

[6] C. L. Byrne and J. Graham-Eagle, "Iterative image reconstruction algorithms based on cross-entropy minimization," in *Proc. SPIE Conf. Inverse Problems in Scattering and Imaging*, San Diego, July 1992.

[7] Y. Chow and U. Grenander, "A sieve method for the spectral density," *Ann. Stat.*, vol. 13, no. 3, pp. 998-1010, 1985.

[8] I. Csiszár, "Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems," *Annals Statist.*, vol. 14, no. 4, pp. 2032-2066, 1991.

[9] ——, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299-318, 1967.

[10] A. D. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *J. Royal Stat. Soc. Ser. B*, vol. 39, no. 1, pp. 1-37, 1977.

[11] H. Derin, P. A. Kelly, G. Vezina, and S. G. Labitt, "Modeling and segmentation of speckled images using complex data," *IEEE Trans. Geosci. Remote Sensing*, vol. 28, no. 1, pp. 76-87, 1990.

[12] J. A. Fessler and A. O. Hero, "Complete-data spaces and generalized EM algorithms," in *Proc. 1993 ICASSP*, Minneapolis, Apr. 1993.

[13] I. J. Good and R. A. Gaskins, "Nonparametric roughness penalties for probability densities," *Biometrika*, vol. 58, pp. 255-277, 1971.

[14] J. D. Gorman and B. J. Thelen, "A Markov random field model for complex-valued radar imagery," in *Proc. 1993 IEEE Int. Symp. Inform. Theory*, San Antonio, Jan. 1993, p. 136.

[15] A. O. Hero and J. A. Fessler, "A recursive algorithm for computing CR-type bounds on estimator covariance," *IEEE Trans. Inform. Theory*, vol. 40, no. 4, pp. 1205-1210, 1994.

[16] ——, "Asymptotic convergence properties of EM-type algorithms," Comm. and Signal Proc. Lab. Tech. Report 282, EECS Dept., University of Michigan, Ann Arbor, Apr. 1993.

[17] F. Itakura and S. Saito, "Analysis-synthesis telephony based on the maximum likelihood method," in *Proc. 6th Int. Conf. Acous. C*, Tokyo, Japan, 1968, pp. 17-20.

[18] L. K. Jones, "Approximation-theoretic derivation of logarithmic entropy principles for inverse problems and unique extension of the maximum entropy method to incorporate prior knowledge," *SIAM J. Appl. Math.*, vol. 49, no. 2, pp. 650-661, 1989.

[19] 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, no. 1, Jan. 1990.

[20] P. A. Kelly, H. Derin, and K. D. Hartt, "Adaptive segmentation of speckled images using a hierarchical random field model," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, no. 10, pp. 1628-1641, Oct. 1988.

[21] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Med. Imag.*, vol. 9, no. 4, pp. 439-446, Dec. 1990.

[22] D. G. Luenberger, *Optimization by Vector Space Methods.* New York: Wiley, 1969.

[23] M. I. Miller and B. Roysam, "Bayesian image reconstruction for emission tomography incorporating Good's roughness prior on massively parallel processors," in *Proc. Natl. Acad. Sci. USA,* vol. 88, Apr. 1991, pp. 3223–3227.

[24] M. I. Miller, B. Roysam, K. R. Smith, and J. A. O'Sullivan, "Representing and computing regular languages on massively parallel networks," *IEEE Trans. Neural Networks,* vol. 2, no. 1, pp. 56–72, Jan. 1991.

[25] 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 constrained covariances," *Proc. IEEE,* vol. 75, pp. 892–907, July 1987.

[26] 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, no. 2, pp. 801–813, Mar. 1992.

[27] J. M. Ortega and W. C. Rheinholdt, *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic, 1970.

[28] J. A. O'Sullivan, P. Moulin, D. L. Snyder, and D. G. Porter, "An application of splines to maximum likelihood radar imaging," *Int. J. Imaging Syst. Technol.,* vol. 4, pp. 256–264, 1992.

[29] D. L. Snyder, A. D. Lanterman, and M. I. Miller, "Regularizing images in emission tomography via an extension of Good's roughness penalty," Electronic Systems and Signals Research Laboratory Monograph 92-17, Washington University, St. Louis, MO, 1992.

[30] 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.

[31] ——, *Random Point Processes in Time and Space.* New York: Springer-Verlag, 1991.

[32] 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, May 1989.

[33] D. L. Snyder, T. J. Schulz, and J. A. O'Sullivan, "Deblurring subject to nonnegativity constraints," *IEEE Trans. Signal Processing,* vol. 40, no. 5, pp. 1143–1150, May 1992.

[34] B. J. Thelen and J. D. Gorman, "A non-Gaussian MRF model for nonnegative imagery," in *Proc. 26th Conf. on Information Sci. and Systems,* Princeton, NJ, Mar. 1992, pp. 363–368.

[35] J. R. Thompson and R. A. Tapia, *Nonparametric Function Estimation, Modeling, and Simulation.* Philadelphia: SIAM, 1990.

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

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

Joseph A. O'Sullivan (S'83–M'85–SM'92) was born in St. Louis, MO, on January 7, 1960. He received the B.S, M.S., and Ph.D. degrees in electrical engineering from the University of Notre Dame in 1982, 1984, and 1986, respectively.

In 1986, he was appointed Visiting Assistant Professor in the Department of Electrical Engineering at Washington University; in 1987, Assistant Professor; and in 1994, Associate Professor. He has also been a Research Associate in the Electronic Systems and Signals Research Laboratory at Washington University since 1986. Prof. O'Sullivan is currently the Publications Editor for the IEEE TRANSACTIONS ON INFORMATION THEORY. He was chairman of the St. Louis Section of the IEEE in 1994. His research interests include information theory with applications in magnetic recording and formal languages, estimation theory including applications in radar, and image processing.