DIVERGENCE PENALTY FOR IMAGE REGULARIZATION

Joseph A. O'Sullivan

Department of Electrical Engineering, Campus Box 1127, Washington University, St. Louis, MO 63130 USA.
(314) 935-4173. Fax: 935-4842. E-mail: *jao@saturn.wustl.edu*

ABSTRACT

This paper discusses a new roughness penalty for use in estimation problems including image estimation problems. It is one of a new class of penalty functions for use in estimation and image regularization that has recently been proposed. These functions penalize the discrepancy between an image and a shifted version of itself; here the discrepancy measure is the I-divergence. This penalty is closely related to a penalty used for density estimation that was introduced by Good and Gaskins. Roughness penalty methods form an attractive alternative to Markov random field priors, achieving many of the same properties including the introduction of neighborhood structures. An example of the use of this new penalty for radar imaging using real radar data has been examined.

1. INTRODUCTION

Markov random field models are commonly used in image estimation and regularization [1,2,8]. However, there does not appear to be a standard or natural model for nonnegative-valued images in the same sense that Gauss-Markov random fields are natural for real-valued images. The penalty proposed here, based on the I-divergence between an image and a shifted version of itself, is an attractive alternative. While penalties may be considered to be equivalent to priors, their use may be better motivated by taking the viewpoint proposed here.

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, and is developed in detail in [13]. The former are discussed in recent papers by Byrne [3,4], they motivate Csiszar's results [5] and Jones' results [10,11], and they include maximum entropy penalties. Lange [12] (see also references in [12]) uses penalties that are special cases of those derived in [13]. In [13], the I-divergence penalty is shown to be a discretization of an information theoretic penalty due to Good and Gaskins [7], giving further evidence of its importance in estimation problems.

2. ROUGHNESS PENALTIES

Let $\mathbf{R}$ be the real line, $\mathbf{R}_+ = \{x \in \mathbf{R} : x > 0\}$. Let $V \in \{\mathbf{R}, \mathbf{R}_+\}$. 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 [13]

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

These shifts are circular in that no entries are shifted off the lattice. Examples include left and right shifts for sequences, and vertical and horizontal shifts for images. A function $d : V^n \times V^n \to \mathbf{R}_+ \cup \{0\}$ is called a *discrepancy measure* if $d(x, \xi) = 0$ if and only if $x = \xi$. The types of discrepancy measures studied are motivated by the work of Csiszar [5] and Jones [10]. These include squared error and divergence. Let $V = \mathbf{R}_+$. The I-divergence is defined by

$$I(\mathbf{x}, \xi) = \sum_{k=1}^{n} \left[x_k \log \frac{x_k}{\xi_k} - x_k + \xi_k \right] . \tag{2}$$

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

$$\Phi(\mathbf{x}) = \sum_{i=1}^{I} w_i d(\mathbf{x}, \mathbf{S}_i \mathbf{x}) \; , \tag{3}$$

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

Let $L = \mathbf{R}_+^n$. The divergence penalty with respect to left and right shifts is

$$\Phi_l(\mathbf{x}) = I(\mathbf{x}, \mathbf{S}_l \mathbf{x}) + I(\mathbf{x}, \mathbf{S}_r \mathbf{x}) \tag{4}$$

$$= \sum_{m=1}^{n} (x_m \log \frac{x_m}{x_{m-1}} - x_m + x_{m-1} + x_m \log \frac{x_m}{x_{m+1}} - x_m + x_{m+1}) \; .$$

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

$$\Phi_l(\mathbf{x}) = - \sum_{m=1}^{n} x_m[(\log x_{m+1} - \log x_m)$$

0-7803-1775-0/94 $3.00 © 1994 IEEE

$$- (\log x_m - \log x_{m-1})] \,. \qquad (5)$$

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 [7] (see [13] for details).

The divergence penalty may be extended to images by defining vertical and horizontal shifts. Let $\mathbf{x} \in \mathbf{R}_+^{NM}$. Define the vertical shift S_V and horizontal shift S_H in the obvious (circular) ways. Then the divergence penalty with respect to shifts $S = \{S_V, S_V^{-1}, S_H, S_H^{-1}\}$ is

$$\Phi_I(\mathbf{x}) = I(\mathbf{x}, S_V(\mathbf{x})) + I(\mathbf{x}, S_V^{-1}(\mathbf{x}))$$
$$+ I(\mathbf{x}, S_H(\mathbf{x})) + I(\mathbf{x}, S_H^{-1}(\mathbf{x})) \,. \qquad (6)$$

An important property of penalties when used in estimation problems is convexity. A sufficient condition for Φ to be convex is that the Hessian of Φ, $\mathbf{H}_\Phi = \nabla^2_{\mathbf{xx}}\Phi(\mathbf{x})$, is nonnegative definite; in turn, this Hessian is nonnegative definite if the matrix of second partial derivatives of d is nonnegative definite. Throughout most of this paper, a special form for d is assumed. A discrepancy measure d is *generated by* the scalar discrepancy measure h if

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

Each of the discrepancy measures discussed in this paper is generated by a scalar discrepancy measure. The second derivatives of h then determine whether Φ is convex.

Lemma 1 [13]. Let $V = \mathbf{R}_+$ and let d be generated by h. If $h(x, \xi) = xf(x/\xi)$ for some function f that is twice differentiable, then Φ is convex if $2\dot{f}(x) + x\ddot{f}(x) \geq 0$.

An example of this lemma is given by the I-divergence. There, $f(x) = \log x + 1/x - 1$, and $2\dot{f} + x\ddot{f} = 1/x > 0$. Discrepancy functions of the form given in the lemma play an important role in information theory (see [5,6]). The Itakuro-Saito distance is not in this form and is not recommended [13].

These penalties induce neighborhood structures in the same way as Markov random field priors if d is generated by h. Following Besag [1,2], the *neighborhood* of site k is the set $N(k) = \{\pi_1^{\pm1}(k), \ldots, \pi_I^{\pm1}(k)\}$. The *neighbors* of x_k are the entries in the set $\{x_l : l \in N(k)\}$. 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 $\{C_1, C_2, \ldots, C_J\}$ forms a partition of $\{1, 2, \ldots, n\}$, the labeling of sites by the integers $\{1, 2, \ldots, J\}$ according to their coding sets is called a *coloring*.

3. PENALIZED ESTIMATION

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 penalized maximum likelihood problem is to find the $\mathbf{x} \in V^n$ that maximizes

$$l(\mathbf{x}) = \log f(\mathbf{y}|\mathbf{x}) - \alpha\Phi(\mathbf{x}) \,. \qquad (7)$$

In this problem, α is the weight given to the penalty and it controls the amount of smoothing. Larger values of α give higher weight to the penalty and induce more smoothing in the estimate.

The connection to prior probabilities is clear from (7). If $f_x(\mathbf{x})$ is a prior probability density function on $\mathbf{x}$, then the loglikelihood function for estimating $\mathbf{x}$ is $l_1(\mathbf{x}) = \log f(\mathbf{y}|\mathbf{x}) + \log f_x(\mathbf{x})$. Clearly, if $f_x(\mathbf{x}) = Z^{-1}\exp[-\alpha\Phi(\mathbf{x})]$, where $Z = \int \exp[-\alpha\Phi(\mathbf{x})]d\mathbf{x}$, then the penalty is equivalent to the prior. For least squares discrepancy measures, the prior is Gaussian. For other discrepancy measures such as the I-divergence, the equivalent prior may not take on a common form. For this reason, it may be easier to justify the use of (7) from the penalty view than from the prior probability view.

The neighborhood structure may be used to derive an expectation-maximization (EM) algorithm based on coding sets [13]. Assume there exists $\mathbf{z} \in V^n$ such that (i) given $\mathbf{z}$ the entries of $\mathbf{y}$ are independent of $\mathbf{x}$ and (ii) the conditional density function of the kth entry of $\mathbf{z}$, z_k, given $\mathbf{x}$ depends only on x_k. The set of values of $\mathbf{z}$ are called the complete data space, the set of values of $\mathbf{y}$ the incomplete data space. We may write

$$f(\mathbf{y}|\mathbf{x}) = \int f_2(\mathbf{y}|\mathbf{z}) f_1(\mathbf{z}|\mathbf{x}) dz \,, \qquad (8)$$

where

$$f_1(\mathbf{z}|\mathbf{x}) = \prod_{m=1}^{n} f_{1m}(z_m|x_m) \,. \qquad (9)$$

Let a coloring be based on the coding sets $C = \{C_1, C_2, \cdots, C_J\}$. The EM algorithm from [13] based on the coding sets C has the following steps:

1. Obtain an initial estimate $\mathbf{x}_0$ for $\mathbf{x}$; let $p = 0$.
2. (E-step) Let $j = p + 1 \bmod J$; for each $k \in C_j$, compute

$$Q'_k(x_k|\mathbf{x}^p) = E[\log f_{1k}(z_k|x_k)|\mathbf{y}, \mathbf{x}^p] \,. \qquad (10)$$

3. (M-step) For each $k \in C_j$ maximize $Q'_k(x_k|\mathbf{x}^p)$ over x_k to get x_k^{p+1}; for $l \notin C_j$, $x_l^{p+1} = x_l^p$.
4. If converged, stop, else $p = p + 1$ go to step 2.

At each iteration, this algorithm updates only that part of the vector $\mathbf{x}$ corresponding to one coding set. The coding sets are updated cyclically. The M-step is straightforward and parallelizable due to the decoupling that results from this choice of complete data space. The computation of the E-step may be the hard part. For many problems, however, the decoupling above is natural and the E-step may be no more complicated than a conventional EM algorithm.

In [13], this algorithm is shown to converge under certain conditions. Essentially, the conditions are that a conventional EM algorithm would converge plus a condition that guarantees that the cyclic updating converges to the same point. The condition depends on viewing the update from step p to step $p + J$ as one step, then showing that this algorithm converges. The proof is based in part on recent work by Hero and Fessler [9].

4. APPLICATION TO RADAR IMAGING

As described in [14,15], diffuse radar targets are commonly modeled as having a reflectivity density that is an uncorrelated complex Gaussian random process whose intensity is called the scattering function. Experimental data have been collected from a rough rotating sphere placed on a pedestal in a compact radar range [14]. Under the assumptions in [14], after preprocessing the data are z_{ij}, an image of 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 receiver noise intensity (measured as 0.0023, see [14]). Since the data are independent, there is no need to use an iterative algorithm and the loglikelihood minus the penalty (7) is minimized for various values of α to attempt to measure the amount of smoothing introduced by the penalty. Shown in Figures 1-4 are four images for a data set with total signal energy to total noise energy ratio 0 dB. Figure 1 shows the image for a very small weight on the penalty, and the remaining figures show logarithmically increasing values of the weights. Note that in Figure 4 the image is blurred significantly due to the large weight, but the image is smooth and the background noise is reduced. In Figure 3, the seemingly increased noise level is due solely to the displays being self-normalized. In [13], a performance curve that quantifies the tradeoff between roughness and likelihood parameterized by α is introduced and studied.

5. CONCLUSIONS

A new penalty for image regularization based on the I-divergence is introduced. This penalty is a special case of a class of penalties introduced in [13]. When used in estimation problems, these penalties trade off smoothness for likelihood. The penalties are useful in other cases as well, including deblurring.

ACKNOWLEDGMENT

This work was sponsored by the US Office of Naval Research under contract 59922.

REFERENCES

[1] 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.

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

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

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

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

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

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

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

[9] A. O. Hero and J. F. Fessler, "Asymptotic convergence properties of EM-type algorithms," Comm. and Signal Proc. Lab. Tech. Report 282, EECS Dept., University of Michigan, Ann Arbor, April 1993.

[10] 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.

[11] 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.

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

[13] J. A. O'Sullivan, "Roughness penalties on finite domains," submitted for publication, 1993.

[14] J. A. O'Sullivan, P. Moulin, D. L. Snyder, and D. G. Porter, "An application of splines to maximum likelihood radar imaging," *International Journal on Imaging Systems and Technology*, vol. 4, pp. 256-264, 1992.

[15] 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.

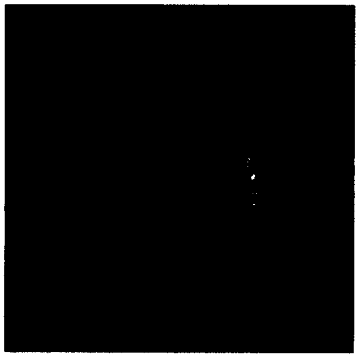

Figure 1

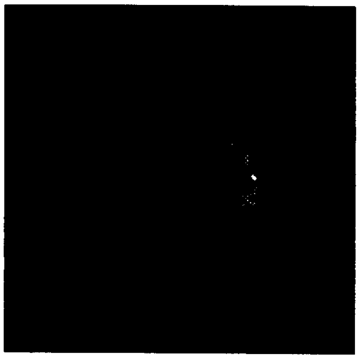

Figure 2

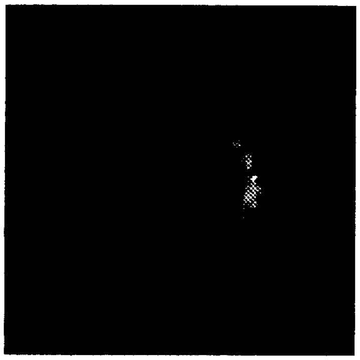

Figure 3

Figure 4