Algorithms for Constructing Min-Max Partitions of the Parameter Space for MDL Inference

Adriana Vasilache, Ioan Tăbuș, and Jorma Rissanen

Tampere University of Technology, Tampere, Finland

Abstract. In this paper we present several algorithms for the construction of min-max optimal partitions of the parameter space. Two interpretations of the problem lead to two families of practical algorithms that are tested and compared.

1 Introduction

The problem of partitioning the parameter space of a given class of models arises as a necessary step in statistical inference based on the minimum description length (MDL) principle. This is because the model cost is linked to the code-length needed to represent a quantized version of the parameters, either in implicit or explicit manner. The exact construction of the partition is not always needed, and establishing bounds for the average or min-max distortions may suffice, for instance when one wants to find only the structure of the model by the MDL principle. However, when the optimal MDL model including the parameter values themselves are needed, for instance when the models are used for classification or prediction, it is necessary to have an algorithm to give the actual quantization of the parameters.

For a class of models defined by a parametric probability distribution $f = f(y^n; \theta)$, the additional increase in code-length incurred by quantizing the parameter vector θ to the center of a partition cell θ^i can be measured by the Kullback-Leibler (K-L) distance between the un-quantized model $f = f(y^n; \theta)$ and the quantized one $f_j = f(y^n; \theta^j)$. It is therefore of interest to design partitions of the parameter space such that either the average or the min-max loss is minimized. A related problem was discussed in [1], where a preliminary step in the construction of the Kolmogorov structure function was that of obtaining a partition of the parameter space by minimizing the min-max K-L distance between the models $f_i = f(y^n; \theta^i)$ and $f_j = f(y^n; \theta^j)$, defined by the centers of two adjacent partition cells, $\theta^i = \theta(i)$ and $\theta^j = \theta(j)$.

In this paper we propose several algorithmic approaches to the construction of min-max optimal partitions. To ensure the practicality of the algorithms we decide to allocate a model $f = f(y^n; \theta)$ to the center $f_j = f(y^n; \theta^j)$ by the nearest neighbor rule; i.e. the smallest Euclidean norm, $\|\theta - \theta^j\|^2$. With this rule, the parameter space is partitioned into polygonal cells $V(\theta^i)$; i.e. the Voronoi cells corresponding to the centers θ^j . As in universal coding, the relevant criteria are of a min-max nature, since usually there is no apriori distribution on the

parameter space for defining meaningful averages. If we allow $\{V(\theta^i)\}$ to be a covering instead of a partition, the optimal min-max problem for the K-L distance $D(\cdot\|\cdot)$

$$\min_{\{\theta^i\}_{i=1}^N} \max_i \max_{\theta \in V(\theta^i)} D(f(y^n; \theta) \| f(y^n; \theta^i)) \quad (1)$$

has a simple solution if one can find the set of centers $\{\theta^i\}_{i=1}^N$ such that

$$\max_{\theta \in V(\theta^i)} D(f(y^n; \theta) \| f(y^n; \theta^i)) = d' \quad (2)$$

is constant. In the same way the slightly different min-max problem

$$\min_{\{\theta^i\}_{i=1}^N} \max_i \max_{j: (i,j) \text{ are neighbors}} D(f(y^n; \theta^j) \| f(y^n; \theta^i)) \quad (3)$$

has a simple solution when

$$D(f(y^n; \theta^j) \| f(y^n; \theta^i)) = d' \quad (4)$$

is constant for all neighbors i and j .

By the Taylor expansion the K-L distance between two adjacent models can be approximated as [2]

$$D(f_i \| f_j) = \frac{n}{2} (\theta^j - \theta^i)^T J(\tilde{\theta}) (\theta^j - \theta^i), \quad (5)$$

where $\tilde{\theta}$ is a point between θ^i and θ^j and $J(\tilde{\theta})$ is the Fisher information matrix [3] calculated at $\tilde{\theta}$.

Assuming further that the information matrix varies smoothly so that it can be calculated at θ^i instead of $\tilde{\theta}$, the goal of optimization expressed by Eq. (4) can be redefined as one of aiming at the partition of the parameter space such that the weighted distance

$$d_F(\theta^i, \theta^j) = (\theta^j - \theta^i)^T J(\theta^i) (\theta^j - \theta^i) \quad (6)$$

between each two adjacent centers θ^i and θ^j is the same.

In the quantization literature weighted metrics like $d(x, y) = (x - y)^T B_y (x - y)$, where B_y is a positive definite matrix depending on y , have been studied in view of creating a partition of the space such that the average weighted metric is minimized [4]. Also in the clustering literature the discriminative clustering using a mutual information criterion was shown to be asymptotically equivalent to vector quantization using Fisher weighted distance [5].

We give first the basic definitions and notations, after which we propose several methods for obtaining the desired partition. The methods will be applied to a multinomial source. Also, theoretical considerations concerning the significance of having the neighbors on the hyper-ellipsoids are presented.

2 Definitions and Notations

We consider the parameter space to be finite. There are then, say N , models in this space, whose neighborhood relation is determined by the Euclidean metric (to simplify the calculations). The distance between two neighboring models is in turn evaluated by the Fisher weighted distance (6). The matrix $J(\theta^i)$ is the parametric Fisher information matrix. If $J(\theta^i)$ is decomposed into its the singular values

$$J(\theta^i) = V(\theta^i)\Lambda(\theta^i)V^T(\theta^i), \quad (7)$$

the Fisher weighted distance is in fact an Euclidean weighted distance in a rotated space, the rotation matrix being $V(\theta^i)$ and the weights being given by the singular values of $J(\theta^i)$.

2.1 Neighborhood Relationship

Gabriel neighbors. Given a set of points, two points are Gabriel neighbors if they share a facet of their Voronoi cells and if the line passing through the two points intersects that facet. In order to check if two points are Gabriel neighbors their midpoint is considered. If the midpoint has as a nearest neighbor none of the two points in discussion, then they are not Gabriel neighbors. If the midpoint has as a nearest neighbor one of the two points then one more check should be performed to establish that the two points are Gabriel neighbors. Two more points, equally distanced from the midpoint and situated on the line bisecting the line passing through the two initial points should be considered (see Fig. 1). If either one of the two initial points is their nearest neighbor then the initial points are Gabriel neighbors. This procedure implies the use of one parameter defining the distance of the secondary check points from the midpoint.

Voronoi neighbors. Two centers are Voronoi neighbors if they share a facet of their partition cells. To check if two centers are Voronoi neighbors one should check first if they are Gabriel neighbors. If they are not and if they have one common Gabriel neighbor, it should be subsequently checked if the points of the line bisecting the line connecting the two centers are ever quantized to one of the two centers.

3 Optimal Partition

From a practical point of view, Eq. (2) is interpreted as the realization of a partition having the vertices of each partition cell on an ellipse defined by Eq. (8) and Eq. (4) is interpreted as the realization of a partition having the neighbors of each center on an ellipse defined by

$$(\theta^j - \theta^i)^T J(\theta^i)(\theta^j - \theta^i) = \frac{d'}{n} = d. \quad (8)$$

Fig. 1. Neighborhood relationship definition.

Consider the general problem of making a partition in which certain points defining the partition should be positioned on ellipses whose shape and size vary in space. A practical measure to be minimized is the average distance between each of such points and the ellipses on which they should be. We denote this distance to the ellipses by (DE).

3.1 Enforcing the Neighbors of Each Center on an Ellipse (NE)

The practical criterion to be used for the set of points $P = \{\theta^i\}_{i=1}^N$ is mathematically expressed as

$$D_1(P) = \sum_{i=1}^N \sum_{j=1}^{\ell_i} \|\theta^i - \text{Pr}_j(\theta^i)\|^2, \quad (9)$$

where $\text{Pr}_j(\theta^i)$ is the projection of the point θ^i on the ellipse centered in θ^j , and ℓ_i is the number of neighbors of the center θ^i . The ellipse centered in θ^j is given by Eq. (8). The measure used to evaluate the performance of the algorithm will be an average of the estimates of d , obtained by fitting an ellipse to the Gabriel neighbors of each point in the set that generates the partition. The orientation of the ellipse at each point is given by the information matrix calculated at that point. This is virtually equivalent to calculating the average Fisher weighted distance for the neighbors of each point.

Uniform partitioning (UP). Before trying more complex algorithms a reasonable starting point is the evaluation of a uniform partitioning of the parameter space. We consider a lattice Z_n of partitions with different scaling factors. The set of points P in the parameter space is then defined as

$$P = \{s \cdot v | s \in \mathbb{R}^+, v \in \mathbb{Z}^n\}, \quad (10)$$

where s is the scaling factor. The optimization of this uniform partition is performed on the scaling factor with respect to DE for a given d .

Centroid assignment (CA). Consider a given center (point in the parameter space) and its Gabriel neighbors. In conformity with our goal this center should lie on all the ellipses centered at its neighbors. In one iteration each codevector is assigned to the centroid of its projections on the ellipses corresponding to its Gabriel neighbors, which then minimizes the DE. Note that in this case we are using the Euclidean distance. All the points are updated in each iteration step.

A major problem consists of the points near the boundary of the domain since they do not have neighbors toward the border, therefore they can only be pushed inside the domain. If it happens that after an iteration a point gets out of the domain, it will be eliminated. This means that some of the initial points get eliminated since no procedure exists with which extra points could be inserted.

Grid neighborhood structure (GNS). The partition centers are let free to move, but the neighborhood relation between them is predefined by the original rectangular grid. Only one point is considered in each iteration step, and its neighbors given by the grid are updated by moving them toward the ellipse of the center. In geometrical interpretation this is similar to the preservation of the structure in the self-organizing maps [6], but here the goal of the optimization is completely different. The algorithm is parameterized by a factor indicating how far each neighbor is moved toward its projection on the ellipse. Compared to the centroid assignment method a supplementary feature is the possibility to add centers in the domain by allowing a larger initial set of centers in the grid. Out of this set of centers only those that are inside the admissible domain are considered for the calculation of the information matrix and the corresponding ellipse, but their neighbors are moved when called for by the algorithm, even if they lie outside the domain.

3.2 Enforcing the Vertices of Each Cell on an Ellipse (VE)

The practical criterion to be optimized in this case is

$$D_2(P) = \frac{1}{\sum_{i=1}^N nv_i} \sum_{i=1}^N \sum_{j=1}^{nv_i} \|v_i^j - \text{Pr}_i(v_i^j)\|^2, \quad (11)$$

where nv_i is the number of vertices of the partition cell of the center i and $v_i^j, j = \overline{1, nv_i}$ are the vertices of the partition cell for the center θ^i . The measure used to evaluate the performance of the algorithm will be an average of the estimates of d , obtained by fitting an ellipse to the partition cell vertices of each center.

Uniform partition of space. A uniform partition of the parameter space can be used as a reference and/or as initialization for further algorithms. A scaled Z_n lattice can be used, the optimization being this time the average distance

between the partition set vertices and their projections on the ellipses on which they should lie.

Centroid assignment. In this case the centroid assignment algorithm is related to the minimization of the average distance of the partition set vertices to the ellipses on which they should lie. Therefore one vertex is replaced at each iteration by the average of its projections on the ellipses on which it should lie. All the vertices are updated at each iteration, and the new centers are updated to be the centroids of the polyhedral volume defined by a set of vertices. The delimitation of a partition set is considered to be realized by the same vertices even if the vertex positions change from an iteration to another.

Minimization of the maximum Fisher (MMF) weighted distance. Another method of putting the partition set vertices on the ellipses is by selecting the center of the partition as the point that minimizes the maximum Fisher weighted distance between the center and the vertices. The consistency of this approach with the declared goal will be proved in the next section. The minimization of the maximum Fisher weighted distance is done numerically. There are two possible approaches for updating the set of centers, first by updating all the centers at one iteration (MMF1), and secondly by updating only one center at one iteration (MMF2). The second approach implies supplementary calculations of the vertices of the partition sets.

4 Optimal Location of Points on an Ellipse

We have presented two approaches with goal to equalize the distance between pairs of adjacent models. In this section we check if the methods might have any other significance as well. More precisely we show that, under certain conditions when some neighboring points in the parameter space are on the ellipse whose center defines a model, the center minimizes the maximum K-L distance to the neighboring points as well.

The following theorems are given for the 2-dimensional case, but the generalization to higher dimensions is straightforward. The proofs are left out due to lack of space.

Theorem 1. *If more than three points $\{x_i\}$ are located on a circle $C(R, c)$, but not on the same semicircle, then the center of the circle c is the point for which $\max_i d(c, x_i)$ is minimized, where $d(c, x_i)$ is the Euclidean distance between x_i and c .*

If we consider the ellipse given by the equation $\frac{\lambda_1}{d}y_1^2 + \frac{\lambda_2}{d}y_2^2 = 1$, then the transformation given by $T = \begin{pmatrix} \sqrt{\frac{\lambda_1}{\lambda_2}} & 0 \\ 0 & \sqrt{\frac{\lambda_2}{\lambda_1}} \end{pmatrix}$ with $\lambda_1 \neq 0, \lambda_2 \neq 0$, transforms the ellipse into a circle. The transform is invertible.

Theorem 2. *If more than three points $\{x_i\}$ are located on an ellipse $E(a, b, c)$, but not on the same semi-ellipse, then the center of the ellipse c is the point for which $\max_i d_{a,b}(c, x_i)$ is minimized, where $d_{a,b}(c, x_i)$ is the weighted Euclidean distance between x_i and c , having the weights a and b on the first and the second dimensions respectively. Here a and b are the semi-axes of the ellipse.*

For the points in the parameter space excluding the border of the domain, the neighbors as well as the vertices of the corresponding partition sets, are around the point in the sense of the Theorem 2. Therefore, Theorem 2 insures the equivalence of the minimization of the maximum Fisher weighted distance from a point c to a set of points $\{x_i\}$ and the placement of the points on the ellipse centered in c .

5 Results

We consider a ternary independent source having parameters $\theta_0, \theta_1, \theta_2 = 1 - \theta_0 - \theta_1$ where the independent parameters are grouped in the vector $\theta = [\theta_0 \ \theta_1]'$. The information matrix for the probability density function $P(x; \theta) = \theta_0^{1_{x,0}} \theta_1^{1_{x,1}} (1 - \theta_0 - \theta_1)^{1_{x,2}}$ is given by

$$J_n(\theta) = nJ(\theta) = n \begin{bmatrix} \frac{1}{\theta_0} + \frac{1}{1-\theta_0-\theta_1} & \frac{1}{1-\theta_0-\theta_1} \\ \frac{1}{1-\theta_0-\theta_1} & \frac{1}{\theta_1} + \frac{1}{1-\theta_0-\theta_1} \end{bmatrix}, \quad (12)$$

and the matrix of interest is $J(\theta) = \lim \frac{1}{n} J_n(\theta)$.

5.1 Neighbors on Ellipses

First, the uniform partition has been tested with differently scaled versions of the Z_2 lattice. For a given target d , the number of points in the admissible domain results by setting the scale to the value minimizing the average distance from the centers to the ellipses centered in their neighbors. The results are presented in Fig. 3, the average d being generally larger than the target d . On the same plot the results obtained with the grid of the neighborhood structure are also presented. Starting with the optimal uniform partitions the GNS leads to the average d closer to the target values. The GNS algorithm is used without adding any supplementary centers. This feature is useful when initialized with a general partition or, at least with a non-optimal uniform one, which shows that the GNS method finds only a local optimum. Figure 2 exemplifies the resulting partition centers after CA together with the corresponding ellipses for a target d of 0.03. The effect of evading the borders can be observed along the axes. Due to this effect, it is not fair to have the graphics for the CA method on the same figure as for the UP and GNS methods.

5.2 Partition Set Vertices on Ellipses

Since the distance to ellipses in this case does not vary smoothly as a function of scale for a uniform partition the results for the UP method are only suitable

Fig. 2. Centroid assignment method for target $d = 0.03$

to be used as starting points for subsequent algorithms. The centroid approach gives better results than the uniform partition but, like in the NE case, it can only be used for a part of the domain since the centers tend to move away from the borders. The MMF method has the best results (see Tab. 1) and, generally, the update of one center per iteration gives better results. In all the methods the vertices at infinity have been neglected.

Table 1. Comparison of methods for vertices on ellipses. N is the number of points in the parameter space, d_{UP} , d_{MMF1} and d_{MMF2} are the average d for the corresponding methods.

N	d_{UP}	d_{MMF1}	d_{MMF2}
300	0.00582	0.00481	0.00484
528	0.00374	0.00308	0.00303
780	0.00293	0.00241	0.00236
1225	0.00205	0.00169	0.00166

Fig. 3. Method comparison for neighbors on ellipses case.

6 Conclusion

We have presented several methods for obtaining optimal partitions in the min-max sense. We compared them within each of two families, and found GNS to be the best in NE family and MMF in the VE family. The CA method gives promising results, but the resulting partition does not cover the entire space. The proposed methods are exemplified for a 2-dimensional case, but they can also be applied to higher dimensional spaces.

References

1. Rissanen, J., Tăbuș, I.: Kolmogorov's structure function in MDL theory and lossy data compression. In: *Advances in Minimum Description Length: Theory and Applications*. The MIT Press Cambridge (2004)
2. Rissanen, J.: *Lectures on statistical modeling theory*. Lecture notes (2003) Tampere University of Technology.
3. Kay, S.M.: *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall International, Inc. (1993)
4. Li, J., Chaddha, N., Gray, R.M.: Asymptotic performance of vector quantizers with a perceptual distortion measure. *IEEE Trans. on Information Theory* **45** (1999) 1082–1091
5. Sinkkonen, J., Kaski, S., Nikkilä, J.: Discriminative clustering: optimal contingency tables by learning metrics. In: Elomaa, T., Mannila, H., Toivonen, H., eds.: *Proceedings of the 13th European Conference on Machine Learning*. Lecture Notes in Artificial Intelligence (2002) 418–430
6. Kohonen, T.: *Self organization and associative memory*. 3 edn. Springer, Berlin (1989)