Number of possible of Voronoi Partitions on the Feature Manifolds

Mary Inaba*and Hiroshi Imai*

Abstract

In this paper, we analyze possible number of the k Voronoi partitions of n points on the feature manifold.

In information geometry [1, 2, 8], differential geometric properties of probabilistic distribution have been studied. In the fields such as language recognition or learning theory, clustering of probability distributions, called "*distributional clustering*" minimizing sum of Kullback-Leibler divergence plays an important role. We take geometric approach for this distributional clustering problem and extend our result in the Euclidean space. We consider k-clustering problem of n distributions by divergence in general form on the manifold with dually flat property, and the generalized primary shatter functions for this Voronoi diagram is evaluated, which directly leads to a polynomial-time exact algorithm when the number of clusters and the dimension are regarded as constants.

1 Introduction

In information geometry [1, 2, 8], differential geometric properties of probabilistic distribution have been studied and a lot of fruitful results are reported. A set of parametrized probability distributions form a Riemannian manifold by their parameters, and, exponential family of probabilistic distribution, which contains Poisson, finite discrete, normal, and exponential distribution, has a nice geometric property "dually flat". Divergence in general form is defined on it, which is a distance-like function between two probability distributions and contains both Kullback-Leibler divergence and squared Euclidean distance as its special cases [1, 2]. Divergence can be linearized using its potential functions and their tangent hyperplanes. Using these nice properties, Onishi and Imai studied Voronoi diagram in the dually flat space [9, 10, 11].

Clustering problem is to group similar objects under some criteria, and, in general it is NP-hard. Geometric k-clustering problem is to find a good partition, called a k-clustering, of the given set S of n points $p_i = (\boldsymbol{x}_i)$ $(i = 1,\ldots,n)$ in the d-dimensional space into k disjoint nonempty subsets $S_1,\ldots,S_k$, and, this problem can be considered as a space subdivision problem, and solved efficiently using its geometric properties.

In the fields such as language recognition or learning theory, clustering of probability distributions, called "distributional clustering", minimizing Kullback-Leibler divergence, plays an important role.

In this paper, we consider k-clustering problem of n distributions by divergence in general form on the manifold with dually flat property, which we call "*feature manifold*", whose special case is the distributional clustering problem above. We take geometric approach for this problem by extending our results in the Euclidean space[6, 7].

First, we briefly explain exponential family, feature manifold, divergence in general form, and its geometrical structure with relationship of statistical inference.

Then we introduce weighted Voronoi diagram by divergence on the feature manifold and analyze its complexity. This generalized Voronoi diagram share nice properties with the Euclidean diagrams. Then, the generalized primary shatter functions for Voronoi diagrams is evaluated, which directly leads to a polynomial-time exact algorithm for the distributional clustering problem when the number

*Faculty of Science, University of Tokyo, Hongo, Bunkyo-ku, Tokyo, 113-0033 Japan {mary,imai}@is.s.u-tokyo.ac.jp

of clusters and the dimension are regarded as constants.

2 Properties of exponential family

For the statistical inference, first we assume a type of distribution such as normal distribution or Poisson distribution, then, given a set of observed data, we estimate parameters of distribution of that type, such as mean or deviation in the normal distribution case. In this sense, once a type of distribution is specified, statistical inference can be regarded as the estimation of parameters of the distribution. We regard a statistical distribution characterized by d parameters as a point in the d-dimensional parametric space, and investigate its geometric properties.

A set of parametrized probability distributions form a Riemannian manifold $\mathcal{M}$ by their parameters. For example, a class of one-dimensional normal distribution with mean μ and standard deviation σ form a manifold $\mathcal{M} = \{[\mu, \sigma] \mid \sigma > 0\}$, the upper half plane. This section describes fundamental properties of the manifold for exponential family. Since we will use two dual coordinates, θ-coordinate and η-coordinate, which generalizes the polarity with respect to a paraboloid, we will use the tensor notation.

2.1 Exponential Family

A probability distribution parametrized by $\theta = [\theta^i]$ belongs to the exponential family if its probability density function $f(x;\theta)$ with probability variable (vector) x is expressed as

$$f(x;\theta) = \exp[C(x) + \sum_i \theta^i F_i(x) - \psi(\theta)].$$

Since $\int f(x;\theta)\mathrm{d}x = 1$, ψ is given by

$$\psi(\theta) = \log \int \exp[C(x) + \sum_i \theta^i F_i(x)]\mathrm{d}x$$

For this $\theta = [\theta^i]$, we define $\eta = [\eta_i]$ by

$$\eta_i = \int F_i(x) f(x;\theta)\mathrm{d}x.$$

θ and η are two coordinate systems on the manifold $\mathcal{M}$ of parameters of the distributions in the exponential family. η is also given by

$$\eta_i = \frac{\partial \psi(\theta)}{\partial \theta^i}$$

In the case of the exponential family, the dual potential function $\varphi(\eta)$ is defined in the η-coordinate system by

$$\varphi(\eta) = \int f(x;\theta)(\log f(x;\theta) - C(x))\mathrm{d}x$$

where θ in the right-hand side is that corresponding to η in the left-hand side. Note that when $C(x) \equiv 0$, this potential function φ becomes the minus of entropy of distribution,

$$\varphi(\theta) = \int f(x;\theta) \log f(x;\theta) dx = -H(f_\theta).$$

θ is then given by

$$\theta^i = \frac{\partial \varphi}{\eta_i}.$$

In fact, $\theta = \theta(p)$ and $\eta = \eta(p)$ give two coordinate systems on the manifold $\mathcal{M}$ of points p.

2.2 Properties of the divergence

In the sequel, we adopt the Einstein's notation.

$$\theta^i \eta_i \equiv \sum_i \theta^i \eta_i$$

The manifold $\mathcal{M}$ for the exponential family has good property "dually flat"; roughly speaking, locally, tangent vector and inner products are defined, and any tangent vector can be represented as the linear combination of basis tangent vectors, and, globally, two flat connections is induced where parallel is well defined, and, inner products of tangent vectors are invariant by either connection. Each connection corresponds to an affine coordinate system, and these can be transformed to each other by the Legendre transformation, which has dual structure (e.g., hyperplane corresponds to a point). We call this manifold with dually flat property, "*feature manifold*". One interesting point of this exponential family is one coordinate system directly corresponds to the definition of probabilistic distribution such as θ above, called Canonical Parameters or Natural Parameters, and the other coordinate system directly corresponds to the expectation value, which is η and called Expectation Parameters. We treate the θ-coordinate and the η-coordinate of the manifold $\mathcal{M}$ for the exponential family; $\theta(p)$ and $\eta(p)$ denote the θ- and η-coordinate values for a point p on $\mathcal{M}$, that is, $\theta(p) = [\theta^1(p), \ldots, \theta^d(p)]$, and $\eta(p) = [\eta_1(p), \ldots, \eta_d(p)]$.

2.3 Divergence

We can define a distance-like function *divergence* between two points p and q on $\mathcal{M}$.

Definition 1 (Divergence) *Consider the two potential functions $\psi, \varphi : \mathcal{M} \to \mathbf{R}$ for the exponential family. For two points $p, q \in \mathcal{M}$, define the divergence $D(p\|q)$ by*

$$D(p\|q) = \psi(p) + \varphi(q) - \theta^i(p)\eta_i(q)$$

The pair of potential functions are connected via the Legendre transformation, that is,

$$\theta^i = \frac{\partial \varphi}{\partial \eta_i}, \qquad \eta_i = \frac{\partial \psi}{\partial \theta^i}$$

ψ, φ are strictly convex, and

$$\varphi(q) = \max_{p \in S}\{\theta^i(p)\eta_i(q) - \psi(p)\}$$

$$\psi(p) = \max_{q \in S}\{\theta^i(p)\eta_i(q) - \varphi(q)\}$$

Hence, $D(p\|q) \geq 0$, and $D(p\|q) = 0$ iff $p = q$.

$$D(p\|p) = \psi(p) + \varphi(p) - \theta^i(p)\eta_i(p) = 0$$

But, unlike the distance, $D(p\|q) \neq D(q\|p)$, in general.

Next, we consider the relation of $D(p\|q)$ with the potential function φ and a tangent hyperplane. Add a new coordinate z, corresponding to the height, to the η-coordinate system, and consider the graph $z = \varphi$ in the $[\eta, z]$-space.

For $p \in \mathcal{M}$, lift it up to the graph $(\eta_1(p), \eta_2(p), \ldots, \eta_d(p), \varphi(p))$, and consider the tangent hyperplane

$$z - \varphi(p) = \frac{\partial \varphi}{\partial \eta_i}(p)(\eta_i - \eta_i(p)) = \theta^i(p)(\eta_i - \eta_i(p))$$

$$\text{(i.e.,) } z = \theta^i(p)\eta_i - \psi(p)$$

Then, for a point $q \in \mathcal{M}$, the height difference of a point lifted to the graph $z = \varphi(\eta)$

$$(\eta_1(q), \eta_2(q), \ldots, \eta_d(q), \varphi(q))$$

to a point lifted to the above tangent hyperplane

$$(\eta_1(q), \eta_2(q), \ldots, \eta_d(q), \theta^i(p)(\eta_i(q) - \eta_i(p)) + \varphi(p))$$

is given by

$$\varphi(q) - \theta^i(p)\eta_i(q) + \theta^i(p)\eta_i(p) - \varphi(p)$$
$$= \psi(p) + \varphi(q) - \theta^i(p)\eta_i(q) = D(p\|q)$$

And, by the symmetric duality of the definition of divergence, this linearlization technique can be also applied in the θ-coordinate system; namely, the divergence $D(p\|q)$ is also the difference of the height at the point p between the potential function ψ and tangent hyperplane on ψ on the point q in the θ-coordinate system.

The divergence has such a nice and natural meaning, which was used to analyze the ∇^*-Voronoi diagram as stated and cited in Theorem 1.

2.4 Maximum likelihood method

For a parametrized probability distribution $f(x;\theta)$, suppose we are given a set S_x of n observation $\{x^{(1)}, \ldots, x^{(n)}\}$. For these data, the likelihood function is defined as

$$L(\theta) = \prod_{l=1}^{n} f(x^{(l)}; \theta)$$

and the maximum likelihood method finds θ that maximizes $L(\theta)$.

For the exponential family, we can consider log likelihood. Let $l(x^{(l)};\theta) = \log f(x^{(l)};\theta)$, then, $L(\theta)$ is maximized when

$$\hat{L}(\theta) = \sum_{l=1}^{n} l(x^{(l)};\theta)$$
$$= \sum_{l=1}^{n} C(x^{(l)}) + \sum \theta^i F_i(x^{(l)}) - \psi(\theta).$$

By partial differentiation by θ^i

$$\sum_{l=1}^{n} F_i(x^{(l)}) - \eta_i(\theta)$$

$L(\theta)$ is maximized when $\eta_i(\theta) = \frac{1}{n}\sum_{l=1}^{n} F_i(x^{(l)})$. Recall the definition $\eta_i \equiv \int F_i(x) f(x;\theta) \mathrm{d}x$, the maximum likelihood estimator is nothing but the centroid of the set S in the η-coordinate system. Consequently, given a set S_p of n probability distribution $\{p^{(1)}, \ldots, p^{(n)}\}$, the centroid of the set S_p in the η-coordinate system also becomes a maximum likelihood estimator of whole distribution.

Suppose, given a set S_p of n probability distribution in the exponential family, $\{p^{(1)}, \ldots, p^{(n)}\}$, the centroid of the set S_p in the η- coordinate system is $\eta_i(\bar{p}) = \frac{1}{n}\sum_{j=1}^{n} \eta_i(p^{(j)})$. Divergence $D(p\|p^{(l)})$ is

$$D(p\|p^{(l)}) = \psi(\theta(p)) + \varphi(p^{(l)}) - \sum_i \theta^i(p)\eta_i(p^{(l)}).$$

The lemma below generalizes a well-known formula for the case of sum of squared Euclidean distances, or variance. A case for the Kullback-Leibler divergence was known. For the general divergence, this lemma is shown in

[10].

Lemma 1
$\sum_{l=1}^{n} D(p||p^{(l)}) = nD(p||\overline{p}) + \sum_{l=1}^{n} D(\overline{p}||p^{(l)})$

Proof:
$\sum_{l=1}^{n} D(p||p^{(l)})$
$= \sum_{l=1}^{n} (\psi(p) + \varphi(p^{(l)}) - \theta^i(p)\eta_i(p^{(l)}))$
$= \sum_{l=1}^{n} [(\psi(p) + \varphi(\overline{p}) - \theta^i(p)\eta_i(\overline{p}))$
$+ (-\varphi(\overline{p}) + \varphi(p^{(l)}))]$
$= nD(\theta||\overline{\theta}) + \sum_{l=1}^{n} (\psi(\overline{p}) - \theta(\overline{p}^i)\eta_i(\overline{p}) + \varphi(p^{(l)}))$
(since $\psi(\overline{p}) + \varphi(\overline{p}) - \theta(\overline{p})^i\eta_i(\overline{p}) = D(\overline{p}||\overline{p}) = 0$)
$= nD(p||\overline{p}) + \sum_{l=1}^{n} (\psi(\overline{p}) - \theta(\overline{p})^i\eta_i(p^{(l)}) + \varphi(p^{(l)}))$
$= nD(p||\overline{p}) + \sum_{l=1}^{n} D(\overline{p}||p^{(l)})$
□

The following also holds.

Lemma 2
$\sum_{l=1}^{n} D(\overline{p}||p^{(l)}) = \sum_{l=1}^{n} \varphi(p^{(l)}) - n\varphi(\overline{p})$

Proof:
$\sum_{l=1}^{n} D(\overline{p}||p^{(l)})$
$= n\psi(\overline{p}) + \sum_{l=1}^{n} \varphi(p^{(l)} - \sum_{l=1}^{n} \theta(\overline{p})^i\eta_i(p^{(l)})$
$= n\psi(\overline{p}) + \sum_{l=1}^{n} \varphi(p^{(l)}) - n\theta(\overline{p})^i\eta_i(\overline{p})$
$= \sum_{l=1}^{n} \varphi(p^{(l)}) + n(\psi(\overline{p}) - \theta(\overline{p})^i\eta_i(\overline{p}))$
$= \sum_{l=1}^{n} \varphi(p^{(l)}) - n\varphi(\overline{p})$
□

Since the divergence of two identical points is 0 and the divergence of two distinct points is positive, it is seen that minimizing sum of divergences gives the η-coordinate, namely, the maximum likelihood estimator.

3 Weighted Voronoi diagrams by divergence

The Voronoi diagram by the divergence is investigated in [9, 10]. In extending the all-pair sum of squared Euclidean distances to the divergence case, multiplicatively and additively weighted Voronoi diagrams are useful. Hence, this section investigates such weighted diagrams. As will be seen, the weighted diagram has similar structures as the weighted Euclidean diagram [3], and this result may be viewed as an extension of [3].

3.1 ∇^*-Voronoi diagrams by divergence

First, we consider the locus of points q equidistant, in the sense of the divergence, from two points p and p' on the manifold $\mathcal{M}$. Expanding $D(p||q) = D(p'||q)$, we have
$\psi(p)+\varphi(q)-\theta^i(p)\eta_i(q) = \psi(p')+\varphi(q)-\theta^i(p')\eta_i(q)$.
This is a hyperplane $(\theta^i(p) - \theta^i(p'))\eta_i(q) = \psi(p) - \psi(p')$ in the η-coordinate.

Definition 2 (∇^*-Voronoi diagram) *For k generator points $r^{(j)}$ $(j = 1,\dots,k)$, the ∇^*-Voronoi diagram consists of Voronoi regions $V(r^{(j)})$ defined as follows in [10].*

$$V(r^{(j)}) = \bigcap_{j'\neq j} \{p \mid D(p^{(j)}||p) < D(p^{(j')}||p)\}$$

For the ∇^*-Voronoi diagram, the following holds.

Theorem 1 (Onishi, Imai [10])
The ∇^*-Voronoi diagram can be obtained as the projection to the manifold $\mathcal{M}$ of the upper envelope of hyperplanes which are tangent hyperplanes in the $[\eta, z]$-coordinate of the graph $z = \varphi(p)$ at $[\eta(p), \varphi(p)]$.

By this theorem, the combinatorial complexity of the ∇^*-Voronoi diagram can be bounded by the upper bound theorem for convex polytopes.

3.2 Weighted Voronoi diagram by the divergence

Definition 3 (∇^*-circle) *A ∇^*-circle with center $c \in \mathcal{M}$ and radius of divergence $\rho \in \mathbf{R}_+$, which is defined by*
$\{ q \mid D(c||q) = \rho \} = \{ q \mid \psi(c)+\varphi(q)-\theta^i(c)\eta_i(q) = \rho \}$

Now, we consider the locus of points equidistant from two weighted points. First, we consider a case with multiplicative weight. Suppose the relative weight ratio is $w > 1$ for p to p', and $D(p||q) = wD(p'||q)$. This is expanded to

$$\varphi(q)+\frac{(\psi(p) - w\psi(p'))}{1-w}-\frac{\theta^i(p) - w\theta^i(p')}{1-w}\eta_i(q) = 0$$

Here, it should be noted that $\frac{\theta^i(p)-w\theta^i(p')}{1-w}$ may not be on $\mathcal{M}$. In case it is a point c on $\mathcal{M}$, due to the strict convexity of ψ,

$$\frac{(\psi(p) - w\psi(p'))}{1-w} = \psi(c) - \rho, \qquad \rho > 0$$

and we have the following.

Lemma 3 *The locus of weighted equidistant points from p and p' whose relative weight ratio*

$w > 1$ *is a* ∇^**-circle when* $\frac{\theta^i(p) - w\theta^i(p')}{1-w}$ *is on* $\mathcal{M}$.

This generalizes the Apollonius' circle in the Euclidean case.

When additive weights are further considered, the locus may become empty (this is also the case in the Euclidean space), but, when it exists, it satisfies nice properties like above.

Now, move to the weighted Voronoi diagram by the divergence. Suppose that, for each k points $r^{(j)}$, a multiplicative weight $w^{(j)}$ and an additive weight $\tilde{w}^{(j)}$ are given. The Voronoi region of point $r^{(j)}$ is given by

$$V(r^{(j)}) = \bigcap_{j' \neq j} \{p \mid w^{(j)} D(r^{(j)} \| p) + \tilde{w}^{(j)} < w^{(j')} D(r^{(j')} \| p) + \tilde{w}^{(j')}\}$$
$$= \bigcap_{j' \neq j} \{p \mid H < (-w^{(j)} + w^{(j')}) \varphi(p)\}$$

where

$$H \equiv \sum_i (w^{(j)} \theta^i(r^{(j)}) - w^{(j')} \theta^i(r^{(j')})) \eta_i(p) + (w^{(j)} \psi(r^{(j)}) - w^{(j')} \psi(r^{(j')})) + (\tilde{w}^{(j)} - \tilde{w}^{(j')})$$

Considering the $(d+1)$-th coordinate axis z to form the $(d+1)$-dimensional space $[\eta, z]$, we define a polytope P defined by

$$P = \{[\eta, z] \mid z > \frac{1}{-w^{(j)} + w^{(j')}} H,\ -w^{(j)} + w^{(j')} > 0\}$$
$$\cap \{[\eta, z] \mid z < \frac{1}{-w^{(j)} + w^{(j')}} H,\ -w^{(j)} + w^{(j')} < 0\}$$
$$\cap \{[\eta, z] \mid 0 > H,\ -w^{(j)} + w^{(j')} = 0\}$$

and further consider the projection of

$$P \cap \{[\eta, z] \mid z = \varphi(\eta(p))\}$$

to the original d-dimensional $[\eta^i]$ space is the Voronoi region $V(r^{(j)})$. Since $\varphi(\theta)$ is a convex function, the intersection with P does not increase the combinatorial complexity in magnitude. The combinatorial complexity of P is bounded by $O(n^{\lfloor \frac{d+1}{2} \rfloor})$ using the upper bound theorem for convex polytopes, and this is the complexity bound for each Voronoi region, so that we obtain the following.

Theorem 2 *The combinatorial complexity of the weighted divergence Voronoi diagram is* $O(n^{\lfloor \frac{d+3}{2} \rfloor})$.

4 Clustering by divergence

For a given set S of n points $p^{(l)}$ $(l = 1, \ldots, n)$ on the manifold $\mathcal{M}$, a k-clustering is a partition of S into nonempty k disjoint subsets $S_1, \ldots, S_k$ whose union is S.

Problem 1 (Divergence-sum clustering)

$$\min_{r^{(j)},\ S_j\ (j=1,\ldots,k)} \sum_{j=1}^{k} \sum_{p^{(l)} \in S_j} D(r^{(j)} \| p^{(l)})$$

Here, $r^{(j)}$ is a representative point for S_j, and, since the sum of divergence is minimized at the centroid, $r^{(j)}$ is simply set to the centroid of S_j in the η-coordinate.

This clustering criterion corresponds to maximizing the Classification Maximum Likelihood (CLM) for the exponential family [4].

Problem 2 (All-pair divergence-sum clustering)

$$\min \sum_{j=1}^{k} \sum_{p^{(l)}, p^{(l')} \in S_j} D(p^{(l)} \| p^{(l')})$$

Generally, the divergence does not satisfy the symmetricity, but in this all-pair version the symmetric property inside each cluster holds since the divergences in both directions are included.

Theorem 3 *An optimal clustering for the divergence-sum clustering problem is identical with a partition by the* ∇^**-Voronoi diagram generated by the centroids of clusters.*

Proof: Suppose contrarily that there is an optimal clustering $\{S_1, \ldots, S_k\}$ which is not a partition by the ∇^*-Voronoi diagram generated by the centroids of clusters. From Lemma 1, the centroid of each cluster $\tilde{r^{(1)}}, \ldots, \tilde{r^{(k)}}$ is the representative of each cluster. Construct the ∇^*-Voronoi diagram generated by these centroids $\tilde{r^{(1)}}, \ldots, \tilde{r^{(k)}}$. Since it is not a ∇^*-Voronoi partition, there exist at least one point p which belongs cluster $\exists j S_j$ but belong to the ∇^*-Voronoi cell $\mathrm{Vor}(\tilde{r^{(k)}})$ for $k \neq j$. Then, moving p from cluster S_j to S_k, the total sum of divergences with respect to the current centroids strictly decreases. Furthermore, for the updated clusters S_j and S_k, recomputing their centroids further reduces the total sum of divergence, due to the convexity of the divergence. This contradicts the original clustering is optimal, and we obtain the lemma. □

Theorem 4
For the all-pair divergence-sum clustering, an optimal clustering $S_1, \ldots, S_k$ is identical with a partition by the weighted ∇-Voronoi diagram generated by centroids $\bar{r}^{(j)}$ of cluster S_j with multiplicative weight $|S_j|$ and additive weight $\sum_{p^{(l)} \in S_j} D(r^{(j)} \| p^{(l)})$ for cluster S_j.

5 Generalized primary shatter function of Voronoi partition by divergence

By Theorem 3, k-clustering problem by divergence can be solved by enumerating all the partitions of n points induced by the corresponding Voronoi diagram generated by k points, and finding a partition with minimum one. We call a partition of n points induced by such a Voronoi diagram a Voronoi partition. The number of all possible Voronoi partitions by k generators corresponds to evaluation of the generalized primary shatter function [6] for a label space induced by the Voronoi diagrams. That is, k generators are numbered from 0 to $k-1$, and, each of n points is labeled by the label of a generator whose Voronoi region contains the point. The generalized primary shatter function of this label space is the number of all possible partitions.

In this section, utilizing the dual structure between the η- and θ-coordinate system, we evaluate the generalized primary shatter function $\pi_S(m)$ for the label space $S = (X, \mathcal{L})$ defined for the ∇^*-Voronoi diagram, where X is a set of infinite points on the d-dimensional statistical manifold, and $\mathcal{L}$ is a set of functions from X to $\{0, \ldots, k-1\}$. The function for the weighted ∇^*-Voronoi diagram can be evaluated in a similar way. We evaluate $\pi_S(m)$ by counting the number of cells of an arrangement of hyperplanes in the $(d+1)k$-dimensional representative space.

In the d-dimensional statistical manifold with a dually flat structure, and given k representative points for k-clustering, each of k points can be considered to move independently. Denote by R a set of k generator points $\{r^{(1)}, r^{(2)}, \ldots, r^{(k)}\}$, and, denote by X a set of n observed points which are partitioned. We will consider two spaces, one is the dk-dimensional space of

$$(\theta^1(r^{(1)}), \theta^2(r^{(1)}), \ldots, \theta^d(r^{(1)}), \ldots, \theta^1(r^{(k)}), \theta^2(r^{(k)}), \ldots, \theta^d(r^{(k)})),$$

which we call *representative space*, and the other is the $k(d+1)$-dimensional space,

$$(\theta^1(r^{(1)}), \ldots, \theta^d(r^{(1)}), \theta^{d+1}(r^{(1)}) = \psi(r^{(1)}), \ldots, \theta^1(r^{(k)}), \ldots, \theta^d(r^{(k)}), \theta^{d+1}(r^{(k)}) = \psi(r^{(k)})).$$

Definition 4 (Equivalent Relationship concerning to partitioning) *Given a set X of n points, and sets R and R' of k generator points. If the partitioning $\{X_1, \ldots, X_k\}$ and $\{X'_1, \ldots, X'_k\}$ induced by R and R', respectively are identical, R and R' are in* equivalence relationship *concerning to partitioning.*

By the definition of the ∇^*-Voronoi Diagram, this equivalent relationship changes only when $\exists x, r^{(j)}, r^{(l)}$, the sign of $D(r^{(j)} \| x) - D(r^{(l)} \| x)$ changes. Hence, consider a hypersurface in the dk-dimensional representative space:

$$D(r^{(j)} \| x) - D(r^{(l)} \| x) = \psi(r^{(j)}) - \psi(r^{(l)}) - (\theta^i(r^{(j)}) - \theta^i(r^{(l)})\eta_i(x)) = 0$$

This can be regarded as a hyperplane in the above-mentioned $(d+1)k$-dimensional space, and the total number of the hyperplanes is $n\binom{k}{2} = O(nk^2)$.

Note that in the Euclidean space, $\psi(\theta) = \frac{1}{2}\theta^2$, and $\theta_i(x) = \eta_i(x)$, and the formula above becomes $\psi(r_j) - \psi(r_l) - x(r_j - r_l) = \frac{1}{2}(r_j - r_l)(r_j + r_l - 2x) = 0$. This is a quadratic hypersurface in the dk-dimensional representative space, and the arrangements such hypersurfaces can be theoretically handled. However, in the case of general divergence, the hypersurface is not algebraic, etc., and we need some other technique to evaluate the number of cells of this hypersurface arrangement.

For this purpose, we consider the above-mentioned hyperplane arrangement in the $(d+1)k$-dimensional space. The target arrangement in the dk-dimensional representative space is obtained as the intersection of this arrangement with $\theta^{d+1}(r^{(j)}) = \psi(r^{(j)})$ $(j = 1, \ldots, k)$. This is the lower envelope of the intersection of the $(d+1)k$-dimensional hyperplane arrangement and $\theta^{d+1}(r^{(j)}) \geq \psi(r^{(j)})$ $(j = 1, \ldots, k)$. Then, due to the convexity of ψ, the combinatorial complexity of the lower envelope is bounded in magnitude with the complexity of the $(d+1)k$-dimensional arrangement of $n\binom{k}{2}$ hyperplanes.

This induces, either $r_i = r_j$, that is, r_i and r_j are same point, or, $x = 1/2(r_j + r_l)$, that is, x is the middle point of r_j and r_l.

And the number of cells can be considered as the upper bound of the primary shatter function $\pi_S(n)$ for the label space $S = (R, \mathcal{L})$, where R is $k(d+1)$-dimensional representative space and $\mathcal{L}$ is a set of functions from R to $\{0, \ldots, k-1\}$ according to Voronoi k-partitioning.

To count the cell of hyperplane arrangement, we use the following lemma [5]:

Lemma 4 *The number $N(n)$ of cells constructed from n hyperplanes in E^d is*

$$N(n) = \sum_{i=0}^{d} \binom{n}{i} = O(n^d). \quad (1)$$

And to enumerate the possible Voronoi partitions, we use topological sweep of the hyperplane arrangement in the representative space.

Lemma 5
Given an arrangement of a set of hyperplanes in the representative space. Let $C_{\mathcal{A}}$ denote the number of the cells of the arrangement $\mathcal{A}$. We can enumerate all possible Voronoi partitions in $O(nC_{\mathcal{A}})$ time.

Theorem 5 *The number of distinct partitions of n points induced by the ∇^*-Voronoi diagram generated by k points on $\mathcal{M}$ is bounded by $O(n^{(d+1)k})$.*

Proof:
From Theorem 1, we may simply bound the number of partitions of n points induced by the projection, to the original d-dimensional space, of the upper envelope of k hyperplanes in the $(d+1)$-dimensional space. This would give a little loose bound, but with this method we do not have to consider special structures of specific potential functions. With this method, the $O(n^{(d+1)k})$ bound can be obtained. □

6 Concluding Remarks

We need to consider the problem such that each point $p^{(l)}$ belongs to every cluster in some sense.

Problem 3 (Mixed divergence clustering) *Specifically, $\zeta(j,l)(\geq 0)$ denotes how much point p^l belongs to cluster S_j, where*

$$\sum_{j=1}^{k} \zeta(j,l) = 1 \text{ for each } l.$$

Then, the clustering problem is to find an optimal k-clustering for

$$\min \sum_{j=1}^{k} w\left(\sum_{l=1}^{n} \zeta(j,l)\right) \sum_{l=1}^{n} \zeta(j,l) D(\tilde{r}^{(j)} \| p^{(l)})$$

In this problem, $\tilde{r}^{(j)}$ is the weighted centroid defined by

$$\eta_i(\tilde{r}^{(j)}) = \frac{1}{\sum_{l=1}^{n} \zeta(j,l)} \sum_{l=1}^{n} \zeta(j,l) \eta_i(p^{(l)})$$

References

[1] S. Amari: *Differential Geometrical Methods in Statistics.* Lecture Notes in Statistics, Vol.28, Springer-Verlag, New York, 1985.

[2] S. Amari and H. Nagaoka: *Method of Information Geometry* (in Japanese). Iwanami Shoten, Tokyo, 1993.

[3] F. Aurenhammer and H. Edelsbrunner: An Optimal Algorithm for Constructing the Weighted Voronoi Diagram in the Plane. *Pattern Recognition,* Vol.17 (1984), 251–257.

[4] G. Celeux and G. Govaert: A Classification EM Algorithm for Clustering and Two Stochastic Versions. *Computational Statistics & Data Analysis,* Vol.14 (1992), pp.315–332.

[5] H. Edelsbrunner: Algorithms in Combinatorial Geometry. Springer-Verlag, 1987.

[6] S. Hasegawa, H. Imai and M. Ishiguro: ϵ-Approximations of k-Label Spaces. *Theoretical Computer Science,* Vol.137 (1995), pp.145–175.

[7] M. Inaba, N. Katoh and H. Imai: Applications of Weighted Voronoi Diagrams and Randomization to Variance-Based k-Clustering. *Proceedings of the 10th ACM Symposium on Computational Geometry,* 1994, pp.332–339.

[8] M. K. Murray and J. W. Rice: *Differential Geometry and Statistics.* Chapman & Hall, 1993.

[9] K. Onishi and H. Imai: Voronoi Diagrams for an Exponential Family of Probability Distributions in Information Geometry. *Japan-Korea Joint Workshop on Algorithms and Computation*, Fukuoka, 1997, pp.1–8.

[10] K. Onishi and H. Imai: Riemannian Computational Geometry — Voronoi Diagram and Delaunay-type Triangulation in Dually Flat Space, submitted for publication.

[11] K. Sadakane, H. Imai, K. Onishi, M. Inaba, F. Takeuchi and K. Imai: Voronoi Diagrams by Divergences with Additive Weights. *Proceedings of the 14th Annual ACM Symposium on Computational Geometry*, 1998, pp.403–404.