Hypothesis testing, information divergence and computational geometry

Frank Nielsen

Sony Computer Science Laboratories Inc
Frank.Nielsen@acm.org
www.informationgeometry.org

Abstract. In this paper, we consider the Bayesian multiple hypothesis testing problem from the stance of computational geometry. We first recall that the probability of error of the optimal decision rule, the maximum *a posteriori* probability (MAP) criterion, is related to both the total variation and the Chernoff statistical distances. We then consider the exponential family manifolds, and show that the MAP rule amounts to a nearest neighbor classifier that can be implemented either by point locations in an additive Bregman Voronoi diagram or by nearest neighbor queries using various techniques of computational geometry. Finally, we show that computing the best error exponent upper bounding the probability of error, the Chernoff distance, amounts to (1) find a unique geodesic/bisector intersection point for binary hypothesis, (2) solve a closest Bregman pair problem for multiple hypothesis.

1 Introduction

Developing efficient techniques to determine the underlying *probability model* that produces the random observations is an endeavor of data analysts. In this paper, we consider the simpler *detection* task modeled by the *Multiple Hypothesis Problem* (MHT) where we are given a set of n hypothesis $H_1 : X \sim P_1$, ..., $H_n : X \sim P_n$ for a random variable X following one of the potential distributions $P_1, ..., P_n$, and one has to *decide* based on an independent and identically distributed sample observations $x_1, ..., x_m$ of X which hypothesis holds true. There are numerous applications in engineering fields; For example, in radar detection, we may distinguish airplane signatures using MHT. Let X^m denote the random vector $X_1 \sim X, ..., X_m \sim X$. This statistical classification problem is trivially deterministically solved when the distribution supports are mutually disjoint, but cannot be solved with certainty otherwise. That this, there is an inherent *probability of error*, denoted by P_e^m, of wrongly classifying (and a probability of correct classification $P_c^m = 1 - P_e^m$). We are interested in the asymptotic regime $\alpha = -\frac{1}{m} \log P_e^m$ of P_e^m when $m \to \infty$.

In the Bayesian setting of this n-ary hypothesis testing problem, we are further given *prior probabilities* $w_i = \Pr(X \sim P_i) > 0$, with $\sum_{i=1}^n w_i = 1$ that induce conditional probabilities $\Pr(X = x | X \sim P_i)$. Since observation x belongs to only one class, we have $\Pr(X = x) = \sum_{i=1}^n \Pr(X \sim P_i) \Pr(X =$

$x|X \sim P_i) = \sum_{i=1}^{n} w_i \Pr(X|P_i)$. Let $c_{i,j}$ denote the cost of making decision H_i when in fact H_j is true (matrix $[c_{ij}]$ is called the *design* matrix), and denote by $p_{i,j}(u)$ the probability of making this decision (depending on criterion r). The *Bayesian detector* seeks to minimize the *expected cost* $E_X[c(r(x))]$ with $c(r(x)) = \sum_i \left(w_i \sum_{j \neq i} c_{i,j} p_{i,j}(r(x)) \right)$. The probability of error P_e is a special case obtained for $c_{i,i} = 0$ and $c_{i,j} = 1$ for $i \neq j$: $P_e = E_X \left[\sum_i \left(w_i \sum_{j \neq i} p_{i,j}(r(x)) \right) \right]$. The maximum *a posteriori* probability (MAP) rule considers classifying x as follows: $\mathrm{MAP}(x) = \mathrm{argmax}_{i \in \{1,\ldots,n\}} w_i p_i(x)$, where $p_i(x) = \Pr(X = x|X \sim P_i)$ are the conditional probabilities. Among all potential rules [1], this MAP Bayesian detector minimizes *by construction* the probability of error P_e.

2 Probability of error and divergences

First, consider the case of the binary hypothesis setting with equal priors (ie., $w_1 = w_2 = \frac{1}{2}$). Let $(\mathcal{X}, \mathcal{E})$ be a measurable space with $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{E}$ a σ-algebra on the set $\mathcal{X}$. We denote by P_i the class-conditional probability measures with p_i their Radon-Nikodym densities with respect to a dominating measure[1] ν. We write the average probability of error as:

$$P_e = \int_{x \in \mathcal{X}} p(x) \min(\Pr(H_1|x), \Pr(H_2|x)) \mathrm{d}\nu(x). \quad (1)$$

Using Bayes' rule $\Pr(H_i|X = x) = \frac{\Pr(H_i)\Pr(X=x|H_i)}{\Pr(X=x)} = w_i p_i(x)/p(x)$, we get $P_e = \frac{1}{2} \int_{x \in \mathcal{X}} \min(p_1(x), p_2(x)) \mathrm{d}\nu(x)$. By using the following two tricks of the trades, we get mathematically rid of the min operator:

Trick 1. $\forall a, b \in \mathbb{R},\ \min(a, b) = \frac{a+b}{2} - \frac{|a-b|}{2}$,
Trick 2. $\forall a, b > 0,\ \min(a, b) \leq \min_{\alpha \in (0,1)} a^{\alpha} b^{1-\alpha}$,

and shall link the probability of error to *distances* between probability distributions, called *information divergences*. First, using the first mathematical rewriting trick, we get $P_e = \frac{1}{2} \int_{x \in \mathcal{X}} \left(\frac{p_1(x)+p_2(x)}{2} - \frac{|p_1(x)-p_2(x)|}{2} \right) \mathrm{d}\nu(x) = \frac{1}{2} \left(1 - \frac{1}{2} \int_{x \in \mathcal{X}} |p_1(x) - p_2(x)| \mathrm{d}\nu(x) \right) = \frac{1}{2}(1 - \mathrm{TV}(P_1, P_2))$, where $\mathrm{TV}(P, Q) = \frac{1}{2} \int_{x \in \mathcal{X}} |p(x) - q(x)| \mathrm{d}\nu(x)$ is the *total variation distance* between distributions P and Q. The total variation distance is a *metric* that satisfies both the symmetric and triangular inequality properties. In practice, computing the total variation is intractable[2] for two reasons:

- First, to get rid of the absolute value function, we split the support into two dominant regions: Namely, one region $\mathcal{R}_1$ where $p_1(x) \geq p_2(x)$ and

[1] For continuous distributions, ν is the Lebesgue measure on the Borel σ-algebra. For discrete distributions, ν is the counting measure on the power set σ-algebra.

[2] Except in simple cases. Even the case of univariate Gaussians yields a complex analytic expressions relying on the erf function.

the other region $\mathcal{R}_2$ where $p_2(x) > p_1(x)$. It follows that $\mathrm{TV}(P_1, P_2) = \frac{1}{2}\left(\int_{x\in\mathcal{R}_1}(p_1(x)-p_2(x))\mathrm{d}\nu(x) + \int_{x\in\mathcal{R}_2}(p_2(x)-p_1(x))\mathrm{d}\nu(x)\right)$. Finding explicitly these region borders is not trivial[3] in general, specially for multivariate distributions.

– Second, we need to compute explicitly both integrals. For d-variate distributions like Gaussians, this is again difficult as soon as $d > 1$.

Fortunately, the second trick [2] allows one to *upper bound* the probability of error. From the fact that $\min(a, b) \leq \min_{\alpha\in(0,1)} a^\alpha b^{1-\alpha}$ for $a, b > 0$, we get an *upper bound* on P_e:

$$P_e = \frac{1}{2}\int_{x\in\mathcal{X}} \min(p_1(x), p_2(x))\mathrm{d}\nu(x) \leq \frac{1}{2}\min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}\nu(x). \quad (2)$$

The integral on the rhs. is bounded by one. It measures the similarity between the distributions. For $\alpha = \frac{1}{2}$, it is commonly called the Bhattacharyya coefficient.Defining the Chernoff statistical distance [3] by:

$$C(P_1, P_2) = -\log\min_{\alpha\in(0,1)}\int_{x\in\mathcal{X}} p_1^\alpha(x)p_2^{1-\alpha}(x)\mathrm{d}\nu(x) \geq 0, \quad (3)$$

allows to link the *best exponent error* of P_e with the Chernoff distance [2]: $P_e \leq w_1^{\alpha^*}w_2^{1-\alpha^*}e^{-C(P_1,P_2)} \leq e^{-C(P_1,P_2)}$ (since $0 < w_1, w_2 < 1$), where α^* denote the best α value in Eq. 3. Note that the Bayesian error exponent does not depend on the prior probabilities as long as they are non-zero.

The Chernoff distance can be calculated in closed-form for families of probability distributions called *exponential families* [4]. An exponential family $\mathcal{F}_F$ is a set of probability measures $\mathcal{F}_F = \{P_\theta\}_\theta$ dominated by a measure $\mathrm{d}\mu$ having their Radon-Nikodym densities p_θ expressed canonically as $p_\theta(x) = \exp(t(x)^\top\theta - F(\theta))$, for θ belonging to the natural parameter space: $\Theta = \left\{\theta \in \mathbb{R}^D \middle| \int p_\theta(x)\mathrm{d}\mu(x) = 1\right\}$. Since $\log\int_{x\in\mathcal{X}} p_\theta(x)\mathrm{d}\nu(x) = \log 1 = 0$, it follows that we can express the normalizing function F as$F(\theta) = -\log\int\exp(x^\top\theta)\mathrm{d}\mu(x)$. We recognize the logarithm of the Laplace transform of the measure μ. For full regular families [4], it can be proved that function F is strictly convex and differentiable over the open convex set Θ. Function F is the cumulant function (also called partition function or log-normalizer), and characterizes the family. Parameter θ (natural parameter) defines the member P_θ of the family $\mathcal{F}_F$. Let $D = \dim(\Theta)$ denote the dimension of Θ, the *order* of the family. The term $t(x)$ is a measure mapping called the sufficient statistic [4]. Many usual families of probability distributions $\{P_\lambda \mid \lambda \in \Lambda\}$ are exponential families [4] in disguise once an invertible mapping $\theta(\lambda) : \Lambda \to \Theta$ is elucidated and the measure $\mathrm{d}\mu(x)$

[3] For exponential families (to be described next), this amounts to compute the $(D-1)$-dimensional hyperplane obtained from the intersection of two hyperplanes parameterized by $y = t(x)$. Here, the sufficient statistics $t(x)$ plays the role of "kernel mapping."

expressed as $e^{k(x)}\mathrm{d}\nu(x)$ where ν is the Lebesgue or counting measures. We refer to [4] for such decompositions for the Poisson, Gaussian, multinomial, Gamma, Beta, Dirichlet, etc. distributions.

3 Computational information geometry

In this section, we consider the manifold induced by the exponential family $\mathcal{F}$ and present several algorithmic techniques of computational geometry tailored to the dually flat differential structure [5] to perform the MAP decision rule and to compute the best error exponent (ie., the Chernoff distance).

3.1 MAP decision rule and additive Bregman Voronoi diagrams

A distribution P_θ of an exponential family $\mathcal{F}$ can be indexed either in the natural parameter space θ, or in a dual coordinate system η called the expectation parameter. Those dual coordinate systems are related by the Legendre transformation [6] $F^*(\eta) = \sup_{\theta\in\Theta} \theta^\top\eta - F(\theta)$. It follows that $\eta = \nabla F(\theta) = E_\theta[t(X)]$ (hence the name *expectation* parameter) and $\theta = \nabla F^*(\eta)$. The Kullback-Leibler divergence $\mathrm{KL}(p_{\theta_1} : p_{\theta_2}) = \int p_{\theta_1}(x)\log\frac{p_{\theta_1}(x)}{p_{\theta_2}(x)}\mathrm{d}\mu(x)$ (relative entropy) between two distributions of the same exponential family is equivalent to a Bregman divergence calculated on the swapped natural parameters: $\mathrm{KL}(p_{\theta_1} : p_{\theta_2}) = B(\theta_2 : \theta_1)$, where the Bregman divergence defined for the cumulant function F of the family is defined as $B(\theta : \theta') = F(\theta) - F(\theta') - (\theta - \theta')^\top\nabla F(\theta')$. Using the cumulant convex conjugate pair F and F^*, the Bregman divergence can be rewritten using the *canonical divergence* of dually flat spaces [5]: $B(\theta_2 : \theta_1) = A(\theta_2 : \eta_1) = F(\theta_2) + F^*(\eta_1) - \theta_2^\top\eta_1$. Observe that the canonical divergence computation relies on the *mixed coordinate system* θ/η. Thus we have the following equivalent expressions of the Kullback-Leibler divergence at our disposal:

$$\mathrm{KL}(p_{\theta_1} : p_{\theta_2}) = B(\theta_2 : \theta_1) = A(\theta_2 : \eta_1) = A^*(\eta_1 : \theta_2) = B^*(\eta_1 : \eta_2), \quad (4)$$

using the Legendre involution $(F^*)^* = F$ for the strictly convex and differentiable generator. Dually flat manifolds enjoy dual affine connections [5] that we denote by ∇^e (geodesics straight in θ) and ∇^m (geodesics straight in η).

For hypothesis distributions $P_1, ..., P_n$ belonging to the same exponential family, we write the log density of conditional distributions $p_1(x), ..., p_n(x)$ as equivalent Bregman divergences using convex conjugation [6]:

$$\log p_i(x) = -B^*(t(x) : \eta_i) + F^*(t(x)) + k(x), \quad (5)$$

with $\eta_i = \nabla F(\theta_i) = \eta(P_{\theta_i})$. Thus it follows that the optimal MAP decision rule

$$\begin{aligned}\mathrm{MAP}(x) &= \mathrm{argmax}_{i\in\{1,...,n\}} w_i p_i(x) = \mathrm{argmax}_{i\in\{1,...,n\}} - B^*(t(x) : \eta_i) + \log w_i, \\ &= \mathrm{argmin}_{i\in\{1,...,n\}} B^*(t(x) : \eta_i) - \log w_i \quad (6)\end{aligned}$$

is a *nearest neighbor classifier* in the expectation parameter space for a dual Bregman divergence B^* with additive weights. This characterization of the MAP rule extends the preliminary observation made in the probability simplex for discrete distributions [7]. Each Bregman Voronoi cell defines a *decision region*. Given an observation x, the MAP rule amounts to compute a nearest neighbor query on the weighted expectation parameter points. This can be solved using computational geometry in several ways, as follows:

- Build a left-sided additive weighted Bregman Voronoi diagrams [8] on weighted point set $\{(\eta_1, -\log w_1), ..., (\eta_n, -\log w_n)\}$. Since the bisectors $\mathrm{Bi}_{i,j} : B^*(t(x) : n_j) - B^*(t(x) : \eta_i) + \log \frac{w_i}{w_j} = 0$ are hyperplanes once re-parameterized with $y = t(x)$, we end up with an *affine diagram*[4] of complexity $O(n^{\lceil \frac{D}{2} \rceil})$ [8], where D is the order of the family. Observe that the weights only shift the bisector without changing its orientation. Note that when the order of the family D is greater than the dimension of the support d, we only need to compute the diagram on the restricted d-dimensional hyper-surface $\{(t_1(x), ..., t_D(x)) \quad | \quad x \in \mathbb{R}^d\}$. We then answer nearest neighbor queries by performing proximity location in the Voronoi cell arrangement [9]. This approach is limited to small dimensions. (Recently, additively-weighted Bregman Voronoi diagrams have also been used to learn mixtures of exponential families [10].)
- Use non-metric tree search structures like Bregman ball trees [11] or Bregman vantage point trees [12] that can be straightforwardly extended by taking into account a weight on each point.
- Perform brute-force searching using GPU [13].

3.2 Geometry of the best error exponent

Case of binary hypothesis The best error exponent of a binary hypothesis testing amounts to compute the Chernoff distance $C(P_1, P_2) = \max_{\alpha \in (0,1)} -\log \int p_1^\alpha(x) p_2^{1-\alpha}(x) \mathrm{d}\mu(x)$ for distributions P_1 and P_2 belonging to the same exponential family. We summarize the results reported in [14] when handling exponential families:

- First, it is shown that the α-Chernoff coefficient $c_\alpha(P_1 : P_2)$ amounts to compute another divergence [15] in the natural parameter space:

$$c_\alpha(P_{\theta_1} : P_{\theta_2}) = \int p_{\theta_1}^\alpha(x) p_{\theta_2}^{1-\alpha}(x) \mathrm{d}\mu(x) = \exp(-J_F^{(\alpha)}(\theta_1 : \theta_2)), \qquad (7)$$

  where $J_F^{(\alpha)}(\theta_1 : \theta_2)$ is a skew Jensen divergence defined for F on the natural parameter space as:

$$J_F^{(\alpha)}(\theta_1 : \theta_2) = \alpha F(\theta_1) + (1 - \alpha) F(\theta_2) - F(\theta_{12}^{(\alpha)}), \qquad (8)$$

[4] Equivalent to a power diagram. See [8].

where $\theta_{12}^{(\alpha)} = \alpha\theta_1 + (1-\alpha)\theta_2 = \theta_2 - \alpha\Delta\theta$, with $\Delta\theta = \theta_2 - \theta_1$. It follows that maximizing the α-Chernoff divergence $C_\alpha(P_{\theta_1} : P_{\theta_2}) = -\log c_\alpha(P_{\theta_1} : P_{\theta_2})$ amounts equivalently to maximizing the skew Jensen divergence with respect to α.

– Second, for the optimal value α^* of α, the Chernoff distance amounts to calculate a Bregman divergence: $C(P_{\theta_1} : P_{\theta_2}) = B(\theta_1 : \theta_{12}^{(\alpha^*)}) = B(\theta_2 : \theta_{12}^{(\alpha^*)})$, where α^* is the *unique* value satisfying $\nabla F(\theta_{12}^{(\alpha)})^\top(\theta_1 - \theta_2) = F(\theta_1) - F(\theta_2)$. An alternative definition of Chernoff information for exponential family distributions is $C(P_{\theta_1} : P_{\theta_2}) = \min_{\theta\in\Theta}\{\mathrm{KL}(p_\theta : p_{\theta_1}), \mathrm{KL}(p_\theta : p_{\theta_2}\} = \min_{\theta\in\Theta}\{B(\theta_1 : \theta), B(\theta_2 : \theta)\}$.

It follows a geometric characterization of the Chernoff distribution $P^* = P_{\theta_{12}^*}$ of two distributions P_1 and P_2 belonging to the same exponential family: It is the unique point on the exponential family manifold that belongs to both the e-geodesic and the m-bisector: $P^* = P_{\theta_{12}^*} = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)$, with

$$G_e(P_1, P_2) = \{E_{12}^{(\lambda)} \mid \theta(E_{12}^{(\lambda)}) = (1-\lambda)\theta_1 + \lambda\theta_2, \lambda \in [0,1]\}, \tag{9}$$

(linear interpolation on the natural parameter), and

$$\mathrm{Bi}_m(P_1, P_2) : \{P \mid F(\theta_1) - F(\theta_2) + \eta(P)^\top\Delta\theta = 0\}, \tag{10}$$

a hyperplane equation in the η-coordinate system. This intersection point can be found by bisecting the exponential geodesic, as described in [14]. Furthermore, at the intersection point, we have the *orthogonality* property of the primal e-geodesic with the dual m-bisector proved in [8] (see Figure 1): We say that triangle $\triangle PQR$ is orthogonal at Q if and only if $B(\theta(P) : \theta(Q)) + B(\theta(Q) : \theta(R)) = B(\theta(P) : \theta(R))$. This amounts to check equivalently that $(\theta(P) - \theta(Q))^\top(\eta(R) - \eta(Q)) = 0$, see [8]. Here, we have $G_e(P_1, P_2) \perp \mathrm{Bi}_m(P_1, P_2)$. Observe that the Chernoff point $P^* = G_e(P_1, P_2) \cap \mathrm{Bi}_m(P_1, P_2)$ can also be interpreted as the left-sided Kullback-Leibler projections of source distributions to their m-bisector (or equivalently, the right-sided Bregman projections):

$$\theta^* = \theta_{12}^{(\alpha^*)} = \mathrm{argmin}_{\theta\in\Theta} B(\theta_1 : \theta) = \mathrm{argmin}_{\theta\in\Theta} B(\theta_2 : \theta). \tag{11}$$

Case of multiple hypothesis The best error exponent of a n-ary MHT [1] is determined by the *minimum pairwise Chernoff distance*: $C(P_1, ..., P_n) = \min_{i,j\neq i} C(P_i, P_j)$. It follows that for the observation sequence X^m, the probability of error $P_e^m \leq e^{-mC(P_{i^*}, P_{j^*})}$ where $(i^*, j^*) = \mathrm{argmin}_{i,j\neq i} C(P_i, P_j)$. In the natural parameter space, this amounts to find the *closest pair* of parameters among a set of n points in D dimension (family order) with respect to the Chernoff distance. Although symmetric, Chernoff distance fails the triangular inequality and is therefore not a metric. When the (non-additive) Bregman Voronoi diagram[5] is already available, we may compute the closest Chernoff

[5] Bregman Voronoi diagrams can either be built from equivalent power diagrams or as vertical projections of a $(d+1)$-dimensional polytope. See demo at `http://www.sonycsl.co.jp/person/nielsen/BVDapplet/`

pair as follows (see Figure 1): We compute for each pair of natural neighbors P_{θ_i} and P_{θ_j}, the Chernoff distance $C(P_{\theta_i}, P_{\theta_j})$ that amounts to find the distance $\mathrm{KL}(P_{\theta_i} : P_{\theta^*_{ij}}) = \mathrm{KL}(P_{\theta_j} : P_{\theta^*_{ij}})$ between the Voronoi sites and the information projection $P_{\theta^*_{ij}}$ on the border bisector. The geometry of the Bayesian error exponent is independent of the prior probabilities and therefore relies only on the non-additive Bregman Voronoi diagram. Note that it is only necessary[6] to consider natural neighbor sites in the Voronoi diagram: That is, to inspect pairs of sites whose bisector contribute to the Voronoi diagram (yielding linear time algorithm when $D = 2$ or worst-case quadratic time otherwise). This generalizes to exponential family manifolds the geometric interpretation [7] formerly studied for the probability simplex case.

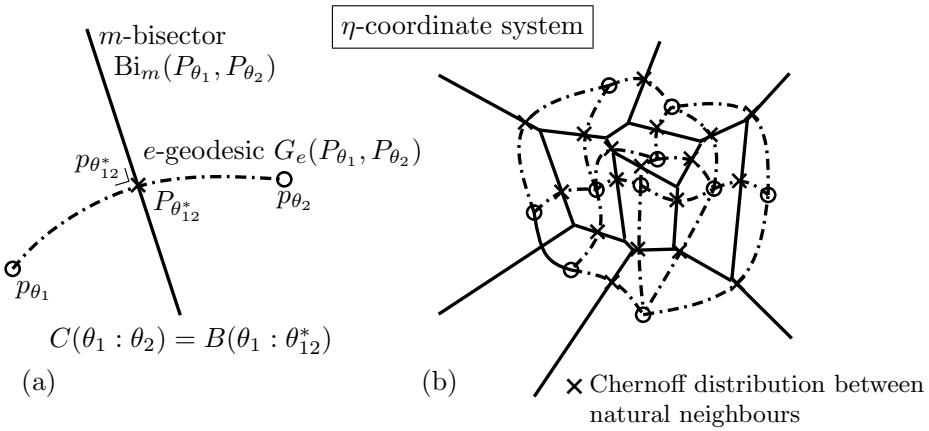

Fig. 1. Geometry of the best error exponent. Binary hypothesis (a): The Chernoff distance is equal to the Kullback-Leibler divergence from the midpoint distribution $P_{\theta^*_{12}}$ to the extremities (or vice-versa for the Bregman divergences), where the midpoint distribution $P_{\theta^*_{12}}$ ($\times$) is obtained as the left-sided KL projection of the sites to their bisector. (b) Multiple hypothesis testing: The Chernoff distance is the minimum of pairwise Chernoff distance that can be deduced from the non-additive Bregman Voronoi diagram by inspecting all Chernoff distributions ($\times$) lying on $(d-1)$-faces. Both drawings illustrated in the η-coordinate system where m-bisectors are hyperplanes. (In the dual θ-coordinate systems, e-geodesics are straight whilst m-geodesics are curved.)

4 Conclusion

In this paper, we focused on several geometric interpretations for Bayes detection theory on the exponential family manifolds, generalizing former work relying on

[6] Proof by contradiction using Bregman Pythagoras theorem for non-adjacent cells [8].

distributions on the probability simplex [7]. We described several computational information-geometric techniques: The MAP decision rule amounts to an additive dual Bregman nearest neighbor classifier that can be solved either using point location or via Bregman tree search data-structures (or GPU brute-force search). The best exponent error for binary hypothesis is interpreted geometrically as the unique (orthogonal) intersection point of an exponential geodesic with a mixture bisector. The best exponent error for multiple hypothesis reduces to a closest pair problem on the manifold that can be deduced from the non-additive Bregman Voronoi diagram or computed by any other appropriate techniques of computational geometry.

References

1. Leang, C.C., Johnson, D.H.: On the asymptotics of M-hypothesis Bayesian detection. IEEE Transactions on Information Theory **43**(1) (January 1997) 280–282
2. Hellman, M.E., Raviv, J.: Probability of error, equivocation and the Chernoff bound. IEEE Transactions on Information Theory **16** (1970) 368–372
3. Chernoff, H.: A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. Annals of Mathematical Statistics **23** (1952) 493–507
4. Brown, L.D.: Fundamentals of statistical exponential families: with applications in statistical decision theory. Institute of Mathematical Statistics, Hayworth, CA, USA (1986)
5. Amari, S., Nagaoka, H.: Methods of Information Geometry. Oxford University Press (2000)
6. Banerjee, A., Merugu, S., Dhillon, I.S., Ghosh, J.: Clustering with Bregman divergences. Journal of Machine Learning Research **6** (2005) 1705–1749
7. Westover, M.: Asymptotic geometry of multiple hypothesis testing. IEEE Transactions on Information Theory **54**(7) (July 2008) 3327–3329
8. Boissonnat, J.D., Nielsen, F., Nock, R.: Bregman Voronoi diagrams. Discrete & Computational Geometry **44**(2) (2010) 281–307
9. Boissonnat, J.D., Yvinec, M.: Algorithmic Geometry. Cambridge University Press, New York, NY, USA (1998)
10. Nielsen, F.: k-MLE: A fast algorithm for learning statistical mixture models. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE (2012) preliminary, technical report on arXiv.
11. Piro, P., Nielsen, F., Barlaud, M.: Tailored Bregman ball trees for effective nearest neighbors. In: European Workshop on Computational Geometry (EuroCG), LORIA, Nancy, France, IEEE (March 2009)
12. Nielsen, F., Piro, P., Barlaud, M.: Bregman vantage point trees for efficient nearest neighbor queries. In: Proceedings of the 2009 IEEE International Conference on Multimedia and Expo (ICME). (2009) 878–881
13. Garcia, V., Debreuve, E., Nielsen, F., Barlaud, M.: k-nearest neighbor search: Fast GPU-based implementations and application to high-dimensional feature matching. In: IEEE International Conference on Image Processing (ICIP). (2010) 3757–3760
14. Nielsen, F.: An information-geometric characterization of Chernoff information. IEEE Signal Processing Letters (SPL) **20**(3) (March 2013) 269–272
15. Nielsen, F., Boltz, S.: The Burbea-Rao and Bhattacharyya centroids. IEEE Transactions on Information Theory **57**(8) (August 2011) 5455–5466