1. Introduction: hypothesis testing, divergences and affinities

1.1. Hypothesis testing

Consider the following fundamental binary hypothesis testing problem: Let $X_1, \ldots, X_n$ be $n$ identically and independently distributed (IID) random variables following distribution $Q$ with support $\mathcal{X}$. We consider two (simple) hypotheses:

$$H_1 : Q \sim P_1 (\text{null hypothesis}),$$

$$H_2 : Q \sim P_2 (\text{alternative hypothesis}),$$

and we design a test $g(X_1, \ldots, X_n) : \mathcal{X}^n \rightarrow \{1, 2\}$ to decide which hypothesis to select. The decision region $R_1 \subseteq \mathcal{X}^n$ corresponds to

the set of sequences $X^n = (X_1, \ldots, X_n)$ mapped to $H_1$, and the decision region $R_2 = \mathcal{X}^n \setminus R_1$ is the complementary region.

To illustrate this setting, consider for example the task of distinguishing a texture [21], modeled by edgelets $\mathcal{E}$ centered at 2D image lattice positions, from two textures $T_1$ and $T_2$, given by their respective edgelet probability distributions $P_1$ and $P_2$ (assuming the IID hypothesis). In practice, we observe a texture sample, that is a data set $x^n = (x_1, \ldots, x_n)$ sampled from $\mathcal{X}^n$.

There are two kinds of error [5] associated with any test:

- Type I error (misclassification when the true hypothesis is $H_1$): $\epsilon_1(n) = \Pr(g(X_1, \ldots, X_n) = 1|H_1)$, and
- Type II error (misclassification when the true hypothesis is $H_2$): $\epsilon_2(n) = \Pr(g(X_1, \ldots, X_n) = 1|H_2)$.

In target/noise detection theory [5], a test is called a detector, and those type I and type II errors are respectively called probability of false alarm and probability of miss.

There are two main approaches for hypothesis testing that have been developed in the literature [5,7]: The first Neyman–Pearson
approach seeks to minimize the probability of miss given the probability of false alarm, without any prior information for the hypothesis. The second Bayesian approach makes use of prior information on class-a priori and class-conditional probabilities for the hypothesis. We concisely review the links between hypothesis testing and statistical distances between distributions for the first non-Bayesian approach in Section 1.2, and the links between hypothesis testing and statistical similarities between distributions for the second Bayesian approach in Section 1.3.

1.2. Statistical divergences in hypothesis testing

The first approach asks to minimize the probability of miss $\epsilon_2$ (false negative) given the probability of false alarm $\epsilon_1$ (false positive):

$$\min_{\epsilon_1} \epsilon_1 \leq \epsilon,$$

where $\epsilon$ is a prescribed error threshold. In the literature, the significance level (or size) of a test is $\epsilon_1$ and the power of a test is $1 - \epsilon_2$. Thus we seek to maximize the power of a test given a prescribed significance level. A key result is the Neyman–Pearson lemma [5] which states the optimality of the Likelihood Ratio Test (LRT) (or equivalently its log-likelihood ratio):

$$\Lambda(X_1, \ldots, X_n) = \log \frac{P_1(X_1, \ldots, X_n)}{P_2(X_1, \ldots, X_n)} = \sum_{i=1}^{n} \log \frac{P_1(X_i)}{P_2(X_i)} \leq \lambda,$$  

(4)

to reject $H_1$ in favor of $H_2$ (with $Pr(\Lambda(X_1, \ldots, X_n) \leq \lambda|H_1) = \epsilon_1$ and $\lambda = \lambda(\epsilon_1)$). Note that the larger the log-likelihood ratio, the more probable the sequence $X^n$ comes from $P_1$ (and the more likely hypothesis $H_1$). From the IDD assumption, we better analyze sequences via the method of types [21,6]. The type $h(x^n)$ is the empirical probability distribution of elements of $x$ (say, a discrete alphabet with $d$ letters) met in $x^n$. That is, for alphabet $X = \{x_1, \ldots, x_d\}$, the type $h(x^n)$ of a sample sequence is the frequency empirical histogram of elements: $h(x^n) = (h_1(x^n), \ldots, h_d(x^n))$, where $h_i(x^n) = \frac{\#(x^n|\text{\textquoteleft}x_i\text{\textquoteleft})}{n}$. Observe that although there are $d^n$ distinct sequences of length $n$ (that is, exponential in $n$), there is only a polynomial number of types (bounded by $(n+1)^d$) since each of the $d$ elements $E_i$ of $X$ has counting number $n_i = nh_i(x^n)$ an integer between 0 and $n$.

The log-likelihood ratio of Eq. (4) rewrites as:

$$\Lambda(X_1, \ldots, X_n) = \sum_{i=1}^{n} \log \frac{P_1(X_i)}{P_2(X_i)} = \frac{d}{n} \log \frac{P_1(E_j)}{P_2(E_j)}.$$  

(5)

Observe that the log-likelihood ratio can be conveniently written as an inner product between two histograms $h(x^n)$ and $h_2(x^n)\Lambda_2 = \left(\log \frac{P_1(E_1)}{P_2(E_1)}, \ldots, \log \frac{P_1(E_d)}{P_2(E_d)}\right) : \Lambda(X_1, \ldots, X_n) = n(h(x^n)|\Lambda_2)$. In [21], the inner product $(h(x^n)|\Lambda_2)$ is called the reward of sequence sample $x^n$.

The expected average log-likelihood ratios with respect to $P_1$ and $P_2$ are:

$$\frac{1}{n} E_{P_1}[\Lambda(X_1, \ldots, X_n)] = \sum_{j=1}^{d} P_1(E_j) \log \frac{P_1(E_j)}{P_2(E_j)} = KL(P_1 : P_2),$$  

(6)

and

$$\frac{1}{n} E_{P_2}[\Lambda(X_1, \ldots, X_n)] = \sum_{j=1}^{d} P_2(E_j) \log \frac{P_2(E_j)}{P_1(E_j)} = -KL(P_2 : P_1),$$  

(7)

where $KL(P_1 : P_2) = \sum_{j=1}^{d} P_1(E_j) \log \frac{P_1(E_j)}{P_2(E_j)}$ denotes the Kullback–Leibler divergence between distributions $P_1$ and $P_2$. The difference between the two expected average log-likelihood ratio is the Jeffreys divergence $J(P_1 : P_2) = KL(P_1 : P_2) + KL(P_2 : P_1)$, that symmetrizes the Kullback–Leibler divergence. It follows that those Kullback–Leibler and Jeffreys measures can be interpreted as measures of separability, that is distances between distributions. Note that the KL distance is asymmetric: $KL(P_1 : P_2) \neq KL(P_2 : P_1)$. The KL and $d$ distances are not metric because they violate the triangular inequality [5].

The probability that a sequence sample $x^n$ from $P_1$ has lower reward than a sequence sample $x^n$ from $P_2$ is bounded by (see [21, Theorem 2]):

$$(n + 1)^{-\beta} \leq \Pr(\Lambda(X_1, \ldots, X_n) \leq \Lambda_2),$$  

(8)

where $B(P_1, P_2)$ denotes the Bhattacharyya divergence:

$$B(P_1, P_2) = -\log \sum_{j=1}^{d} \sqrt{P_1(E_j)P_2(E_j)}.$$  

(9)

Although the Neyman–Pearson lemma [5] characterizes the optimal decision test, it does not specify the threshold $\lambda$. The false alarm error $\epsilon_1$, and miss error $\epsilon_2$ probabilities decay exponentially as the sample size $n$ increases (see [21, Theorem 1]). Thus in the asymptotic regime, we are rather interested in characterizing the error exponents defined as the rate of that exponential decay:

$$\alpha = \lim_{n \to \infty} -\frac{1}{n} \log \epsilon_1(n), \quad \beta = \lim_{n \to \infty} -\frac{1}{n} \log \epsilon_2(n),$$  

(10)

where $\epsilon_1(n)$ is the probability of false alarm $\epsilon_1$ and $\epsilon_2(n)$, the probability of error, the optimal threshold is $\lambda = 0$, and the error rate is the Chernoff information [5,15]:

$$C(P_1, P_2) = \min_{\alpha \in [0,1]} B_\alpha(P_1 : P_2),$$  

(11)

where $B_\alpha$ denotes the skewed Bhattacharyya divergence:

$$B_\alpha(P_1 : P_2) = -\sum_{j=1}^{d} \left( P_1(E_j) \right)^\alpha P_2(E_j)^{1-\alpha},$$  

(12)

generalizing the Bhattacharyya divergence: $B(P_1, P_2) = B_1(P_1 : P_2)$.

Those notions of statistical divergences can be extended to continuous distributions by replacing the discrete sum by an integral (and interpreted $x$ as a continous alphabet).

We now consider the Bayesian paradigm in hypothesis testing, and show how to bound the probability of misclassification error using statistical similarity measures.

1.3. Statistical similarities in hypothesis testing

The Bayesian framework of hypothesis testing assumes that we are given prior beliefs over the probabilities of the two hypothesis, and we seek to minimize the expected probability of error (also called error probability): $Pe = \epsilon_1 Pr(H_1) + \epsilon_2 Pr(H_2)$. In this setting, both the class a priori ($w_i > 0$) and the class conditional probabilities ($p_i$) are known beforehand (or estimated from a training labeled sample [7]). Let $q_1$ and $q_2$ be the a posteriori probabilities derived from Bayes theorem:

$$q_i(x) = \frac{wp(x)}{p(x)},$$  

(13)

where $p(x)$ is the mixture density $p(x) = w_1 p_1(x) + w_2 p_2(x)$ (and $w_1 + w_2 = 1$). Let $C = [c_{ij}]$ be the $2 \times 2$ design matrix, with $c_{ij}$ denoting the cost of deciding $x \in C_i$ when $x \in C_j$, with $1 \leq i, j \leq 2$. Furthermore, denote by

$$r_1(x) = c_{11}q_1(x) + c_{12}q_2(x)$$  

(14)

and

$$r_2(x) = c_{21}q_1(x) + c_{22}q_2(x).$$  

(15)
the respective conditional costs of deciding \( x \in C_i \), for \( i \in \{1, 2\} \). To classify \( x \), consider the decision rule:

\[
 r_i(x) = \frac{c_i}{c_j} \geq r_j(x). 
\]  

The conditional cost of this decision rule is:

\[
 r(x) = \min(r_1(x), r_2(x)). 
\]  

Bayes error \( B_e \) (see [7], p. 57) is defined as the expected cost of this decision rule:

\[
 B_e = E(r(x)),
\]  

\[
 = \int p(x) \min(r_1(x), r_2(x)) \, dx,
\]  

\[
 = \int_{R_1} (c_{11} w_1 p_1(x) + c_{12} w_2 p_2(x)) \, dx + \int_{R_2} (c_{21} w_1 p_1(x) + c_{22} w_2 p_2(x)) \, dx,
\]  

where \( R_1 = \{ x \mid r_1(x) \leq r_2(x) \} \) and \( R_2 = \{ x \mid r_2(x) \leq r_1(x) \} \) are the decision regions induced by the decision rule of Eq. (16). Thus Bayes test for minimum cost writes as:

\[
 (c_{12} - c_{22}) w_2 p_2(x) \geq (c_{21} - c_{11}) w_1 p_1(x),
\]  

or equivalently:

\[
 \frac{p_1(x)}{p_2(x)} \geq \frac{w_2 (c_{12} - c_{22})}{w_1 (c_{21} - c_{11})}.
\]  

The term \( l(x) = \frac{p \mathcal{A}}{p \mathcal{B}} \) is called the likelihood ratio. It is equivalent and often mathematically simpler to consider the test using the log-likelihood ratio (e.g., think of the multivariate Gaussian class-conditional probabilities):

\[
 \log p_1(x) - \log p_2(x) \geq \log \frac{w_2 (c_{12} - c_{22})}{w_1 (c_{21} - c_{11})}. 
\]  

The function \( h(x) = \log p_1(x) - \log p_2(x) - \frac{w_2 (c_{12} - c_{22})}{w_1 (c_{21} - c_{11})} \) is called the discriminant function.

For symmetrical cost function \( c_{21} = c_{11} = c_{12} = c_{22} \), the expected cost is called the probability of error \( P_e \). For the probability of error, we do not incur a cost for correctly classifying and pay a unit cost for misclassification. The design matrix for the probability of error is therefore:

\[
 C = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. 
\]  

As stated earlier, the probability of error can be decomposed as the sum of two misclassification costs:

\[
 P_e = w_1 \epsilon_1 + w_2 \epsilon_2,
\]  

with

\[
 \epsilon_1 = \int_{R_1} p_1(x) \, dx, \quad \epsilon_2 = \int_{R_2} p_2(x) \, dx,
\]  

where \( R_1 = \{ w_2 p_2(x) \leq w_1 p_1(x) \} \) and \( R_2 = \{ w_1 p_1(x) \leq w_2 p_2(x) \} \).

In practice, Bayes error and the probability of error are quite tricky to calculate as we need to compute integrals on decision regions \( R_1 \) and \( R_2 \). Even if those domains can be expressed simply, say for Gaussians, it is often intractable to compute analytically those integrals (e.g., for multivariate class-conditional probabilities). See Appendix A for a review of formulas when class-conditional distributions are Gaussians. Therefore, we need to set good lower and upper bounds to characterize Bayes error \( B_e \) (or the probability of error \( P_e \)).

A first upper bound on the probability of error \( P_e \) is the Bhattacharyya bound [10]:

\[
 P_e \leq \sqrt{w_1 w_2} \rho(p_1, p_2),
\]  

with \( \rho(p_1, p_2) = \int \sqrt{p_1(x) p_2(x)} \, dx \) denoting the Bhattacharyya coefficient. This first upper bound was tightened by Chernoff [4] as follows: We have

\[
 P_e \leq w_1^{-x} w_2^{-x} \rho(p_1, p_2),
\]  

with \( \rho_x(p_1, p_2) = \int (p_1(x)^x p_2(x))^{-x} \, dx \), the \( x \)-Chernoff coefficient defined for \( x \in (0, 1) \). Therefore the tightest upper bound is:

\[
 P_e \leq w_1^{x} w_2^{x} \rho_x(p_1, p_2),
\]  

with \( \rho_x \), the Chernoff coefficient obtained from the following minimization problem:

\[
 \rho_x(p_1, p_2) = \min_{x \in [0,1]} \rho_x(p_1, p_2),
\]  

where \( x^* \in [0,1] \) denotes the optimal value that minimizes \( \rho_x(p_1, p_2) \). Note that the Bhattacharyya coefficient is a particular case of the Chernoff \( x \)-coefficient (obtained for \( x = 1/2 \)). \( \rho = \rho_2 \).

The Bhattacharyya, \( x \)-Chernoff and Chernoff coefficients \( \rho, \rho_x \) and \( \rho = \rho_2 \) can be interpreted a similarity measures between distributions defined by measuring the degree of overlap of their densities. Those coefficients are also called affinities.

**Remark 1 (Affinity coefficients and divergences).** The Chernoff \( x \)-coefficient \( \rho_x \), the Chernoff coefficient \( \rho \), and the Bhattacharyya coefficient \( \rho \left( \rho = \rho_2 \right) \) provide upper bounds on \( P_e \): \( 0 < P_e < 1/2 \), \( 1/4 < \rho < 1/2 \) (for \( w_1 = w_2 = 1 \), see Eqs. (29) and (27)). We can transform any affinity coefficient \( 0 < A \leq 1 \) into a corresponding divergence by applying a monotonously increasing function \( f \) to \( 1/2 \) (with \( 1/2 \in [1, \infty) \)). By choosing \( f(x) = \log(x) \), we end-up with the traditional divergences between statistical distributions: \( x \)-Chernoff divergence, Chernoff divergence (also called Chernoff information), and Bhattacharyya divergence.

Recall that those upper bounds are useful if they can be computed easily from closed-form formula (which is not the case of \( B_e \) or \( P_e \)).

### 1.4. Closed-form Bhattacharyya/Chernoff coefficients for exponential families

Many usual distributions like Gaussians, Poisson, Dirichlet or Gamma/Beta, etc. distributions are exponential families in disguise [11,16] for which the skewed affinity coefficient \( \rho_s \) (i.e., similarity distance within \([0,1]\)) can be computed in closed-form. An exponential family is a family \( \mathcal{F} \) of distributions:

\[
 \mathcal{F} = \{ p(x; \theta) = \exp(x^T \theta - F(\theta)) \mid \theta \in \Theta \},
\]  

indexed by a parameter \( \theta \in \Theta \). Space \( \Theta \) is the parameter domain, called the natural parameter space [16], \( F \) is a strictly convex and differentiable convex function called the log-normalizer. For example, the multivariate normal (MVN) distributions of mean \( \mu \) and covariance matrix \( \Sigma \) are exponential families for parameter \( \theta = (\mu, \Sigma^{-1}) \in \mathbb{R}^d \times S_{d+} \) (where \( S_{d+} \) denotes the space of symmetric positive definite \( d \times d \) matrices). The function \( F \) (called log-normalizer) expressed in the natural coordinate system is [15]:

\[
 F_{\text{MVN}}(\theta_1, \theta_2) = \frac{1}{2} \theta_1^T \theta_2^{-1} \theta_1 - \log |\theta_2|,
\]  

where \( | \cdot | \) denotes the determinant for a matrix operand.

Wlog., let the class-conditional probabilities \( p_1 \) and \( p_2 \) belong to the same exponential family, then we have [11,16]:
where \( f_p^{(0)}(\theta_1, \theta_2) \) denote the Jensen skewed divergence [16]:

\[
\rho_{2}(p_1, p_2) = e^{f_p^{(0)}(\theta_1, \theta_2)},
\]

(33)

denotes the total variation distance extended to positive distributions (i.e., not necessarily normalized probability distributions). In particular, for the probability of error, we have \( a_1 = w_1 \) and \( a_2 = w_2 \) (with \( a_1 + a_2 = 1 \)) and get:

\[
P_e = \frac{1}{2} \cdot \text{TV}(w_1p_1, w_2p_2).
\]

(43)

Note that \( \text{TV}(w_1p_1, w_2p_2) = w_1\text{TV}(p_1, \frac{w_2}{w_1}p_2) = w_2\text{TV}(\frac{w_1}{w_2}p_1, p_2) \). Therefore in the special case \( w_1 = w_2 = \frac{1}{2} \), we get the probability of error related to the total variation distance (a metric) on probability distributions by:

\[
P_e = \frac{1}{2} (1 - \text{TV}(p_1, p_2)).
\]

(44)

For sanity check, notice that when \( p_1 = p_2 \) (undistinguishable distributions), we have \( \text{TV}(p_1, p_2) = 0 \) and \( P_e = \frac{1}{2} \). Clearly, \( 0 \leq \text{TV}(p_1, p_2) \leq 1 \) and \( 0 \leq P_e \leq \frac{1}{2} \). We summarize the result in the following theorem:

**Theorem 1.** The Bayes error \( B_e \) for the cost design matrix \( C = [c_{ij}] \) is related to the total variation metric distance \( \text{TV}(p, q) = \frac{1}{2} \int |p(x) - q(x)| \, dx \) by

\[
B_e = \frac{2}{\text{TV}(p_1, p_2)}.
\]

(45)

3. Upper bounds with generalized means

Without loss of generality, consider the probability of error \( P_e \):

\[
P_e = \int \min(w_1p_1(x), w_2p_2(x)) \, dx = S(w_1p_1, w_2p_2).
\]

(46)

The probability of error can be interpreted as a similarity measure \( S(w_1p_1, w_2p_2) \), extending the definition of histogram intersection [20] to continuous domains.

Chernoff [4] made use of the following mathematical trick:

\[
\min(a, b) \leq a b^{1-z}, \quad \forall a, b > 0,
\]

(47)

to define the Chernoff information upper bounding \( P_e \):

\[
P_e \leq w_1^z w_2^{1-z} \rho_z(p_1, p_2).
\]

(48)

with

\[
\rho_z(p_1, p_2) = \min_{z \in [0, 1]} \rho_z(p_1, p_2), \quad \rho_z(p_1, p_2) = \int (p_1(x))^z (p_2(x))^{1-z} \, dx.
\]

(49)

We shall revisit this technique using the wider scope of generalized weighted means.

By definition, a mean \( M(a, b) \) is a smooth function such that \( \min(a, b) \leq M(a, b) \leq \max(a, b) \). Similarly, we can define a weighted mean as a smooth function \( M(a, b; x) \) that fulfills the interness property \( M(a, b; x) \in [\min(a, b), \max(a, b)] \forall x \in [0, 1] \). Let us consider the quasi-arithmetic means (also called Kolmogorov–Nagumo f-means [12,14]) for a strictly monotonous generator function \( f \):

**Lemma 1** [1]. The quasi-arithmetic weighted mean \( M_f(a, b; x) = f^{-1}(xf(a) + (1 - x)f(b)) \) of two real values \( a \) and \( b \) for a strictly monotonous function \( f \) satisfies the interness property:

\[
\min(a, b) \leq M_f(a, b; x) \leq \max(a, b).
\]

(50)
For all $a$, $b$, $x$, we have $f_a(a, b; x) = M_f(a, b; x)$. Thus, if $f(a) = f(b)$ and $f(a) = f(b)$, then $M_f(a, b; x) = M_f(b, a; x)$. If $b = a$, we similarly have $M_f(a, b; x) = a$. Therefore, minimum $M_f(a, b; x) = (a, b)$. The proof is identical when $f$ is strictly decreasing.

Interestingly, one important property of quasi-arithmetic means is their dominance relationship. That is, if $f(x) < g(x)$ then $M_f(a, b; x) = M_g(a, b; x)$ (with equality when $a = b$). This property generalizes the well known arithmetic–geometric–harmonic (AGH) inequality property of Pythagorean means:

$$M_f(a, b; x) \geq M_g(a, b; x) \geq M_h(a, b; x),$$

with $f_a(x) = x, g_a(x) = \log(x)$ and $h(x) = 1/x$ denoting the generators for the arithmetic, geometric, and harmonic means, respectively.

That is, we have:

$$x_a + (1 - x)b \geq ab \geq \frac{ab}{x_a + (1 - x)b}.$$ (50)

Similarly to Chernoff [4], we define the following generalized affinity coefficient $\rho_f$ using generalized weighted means as follows:

**Definition 1.** The Chernoff-type similarity coefficient (affinity) for a strictly monotonic function $f$ is defined by:

$$\rho_f(p_1, p_2) = \min_{x \in [0, 1]} \int M_f(p_1(x), p_2(x); x) \, dx \leq \int p_1(x) \, dx = 1$$

and define the generalized Chernoff information as:

$$C_f(p_1, p_2) = -\log \rho_f(p_1, p_2) = \max_{x \in [0, 1]} \int M_f(p_1(x), p_2(x); x) \, dx \geq 0. $$ (52)

**Corollary 1.** The traditional Chernoff similarity, information, and upper bound are obtained by choosing the weighted geometric mean, by setting the generator $f_{\text{Chernoff}}(x) = \log(x)$ (with $f_{-1}(x) = \exp(x)$). We get $M_{\text{Chernoff}}(p_1(x), p_2(x); x) = p_1(x)^{p_2(x)}$. When we do not optimize over the parameter $x$, but assume it fixed to $1/2$, we extend the Bhattacharyya coefficient and divergence as follows:

**Definition 3.** The generalized skew Bhattacharyya-type similarity coefficient (affinity) for a strictly monotonic function $f$ is defined by:

$$\rho_f'(p_1, p_2) = \int M_f(p_1(x), p_2(x); x) \, dx \leq \int p_1(x) \, dx = 1$$

and the generalized skew Bhattacharyya-type divergence is defined as $B_f' = -\log \rho_f'(p_1, p_2)$. The generalized Bhattacharyya coefficient $\rho_f(p_1, p_2) = \int M_f(p_1(x), p_2(x); 1/2) \, dx$ and divergence $B_f(p_1, p_2) = -\log \rho_f(p_1, p_2)$.\n
**Theorem 2.** Using quasi-arithmetic means, we can bound the probability of error as follows:

$$P_e = \int \min(w_1 p_1(x), w_2 p_2(x)) \, dx \leq \int M_f(w_1 p_1(x), w_2 p_2(x); x) \, dx.$$ (54)

The upper bound proves useful for well-chosen $f$ yielding closed-form expression of the rhs.

In particular, by choosing the power means $M_{\alpha}$ obtained for $f_\alpha(x) = x^\alpha$, we get a tight bound in the limit case since $M_{\alpha}(p, q) \to \min(p, q)$ when $\alpha \to -\infty$. However, in order for the generalized affinity, distance and upper bound to be useful, we need to be able to compute them in closed-form for some statistical distribution families. We illustrate how to derive closed form formula for $\rho_f$ and closed form upper bounds on $P_e$ for several statistical distribution families.

The generalized Bhattacharyya upper bound $\rho_f^2$ is obtained by setting $x = \frac{1}{2}$:

$$P_e = \int \min(w_1 p_1(x), w_2 p_2(x)) \, dx \leq B_f(w_1 p_1(x), w_2 p_2(x)), $$ (55)

$$B_f(w_1 p_1(x), w_2 p_2(x)) = \int M_f\left(w_1 p_1(x), w_2 p_2(x); 1/2 \right) \, dx.$$ (56)

In order for the Chernoff information to improve over the Bhattacharyya bound, we need the quasi-arithmetic $x$-weighted mean to be a convex function with respect to parameter $x$. For the geometric mean, we check that:

$$M_f(a, b; x) = x^\alpha \log^{1/b}$$ (57)

is strictly convex with respect to $x$ (since $M_f(a, b; x) = (\log(b))^{-1} M_f(a, b; x) > 0$ for $a \neq b$). Similarly, the weighted harmonic mean is convex with respect to $x$ since $\frac{dx}{M_f(a, b; x)} = 2ab(a-b)^{3}(x(a-b) + b)^{3} > 0$.

**Remark 2.** Note that not all quasi-arithmetic means yield convex weighted means. Indeed, let $M_f(a, b; x) = f(\alpha f^{-1}(a) - f^{-1}(b)) + f^{-1}(b)$ then $\frac{dx}{M_f(a, b; x)} = (f^{-1}(a) - f^{-1}(b))^{2} f(\alpha f^{-1}(a) - f^{-1}(b)) + f^{-1}(b)$. Functions $f$ and $f^{-1}$ are strictly monotonic but can be convex, concave or arbitrary in general.

**Remark 3.** Sometimes, we prefer to parameterize the weighted mean as the smooth interpolant from $a$ to $b$, when $x$ varies from 0 to 1 (kind of geodesic parameterization). In that case, we may prefer the parameterization $M_f(a, b; x') = f^{-1}(1-x')f(a) + x'f(b)$. This is not important for Bhattacharyya-type symmetric bounds nor for Chernoff-type bounds that optimize over the $x'$ range (or equivalently over the $x$ range).

Let us now examine how the “quasi-arithmetic bounding techniques” apply for several families of statistical distributions.

### 4. Some illustrating examples

**4.1. Geometric means and the Chernoff bound for exponential families**

First, we recall the well-known formula [11,16] for the case of exponential families.

In order to compute a closed form for the right-hand side of Eq. (58):

$$P_e \leq \int M_f(w_1 p_1(x), w_2 p_2(x); x) \, dx,$$

$$\leq \int f^{-1}(\alpha f(w_1 p_1(x)) + (1 - \alpha) f(w_2 p_2(x))) \, dx,$$

we consider the geometric mean obtained for $f(x) = \log(x)$. Since $p_1(x) = \exp(x \cdot h_1 - F(h_1))$ and $p_2(x) = \exp(x \cdot h_2 - F(h_2))$ belong to the exponential families, we get:

$$M_f(w_1 p_1(x), w_2 p_2(x); x) = e^{\alpha \log w_1 p_1(x) + (1 - \alpha) \log w_2 p_2(x)},$$

$$= w_1^{\alpha} w_2^{1-\alpha} \rho_f^2.$$ (61)

It follows that:
\[ P_x \leq w_1^3 w_2^{-2} \int f^{-1}(xf(p_1(x))) + (1 - x)f(p_2(x))) \, dx. \]  

**Remark 4.** In fact, the geometric mean is a limit case of a family of linear-scale free means defined for \( f_2(x) = x^{-\alpha} \) with \( f_1(x) = \log x \). In general, in order to slide the \( a \) priori weights \( w_1 \) and \( w_2 \) out of the integral, we would like to use a homogeneous function \( f \) (with \( f(\lambda x) = \lambda f(x) \)).

Furthermore, since \( m_2(x; \theta_1, \theta_2) = \log(p_1(x)) + (1 - \alpha) \log(p_2(x)) = x'((\theta_1 + (1 - \alpha)\theta_2) = -\alpha \theta_1 - (1 - \alpha) \theta_2) \), we would like to get \( f^{-1}(m_2(x; \theta_1, \theta_2)) = c_{\theta_1, \theta_2} p(x; (\theta_1^2(\theta_2))) \) so that we can slide the integral operand inside the expression, and use the fact that we recognize a member \( \theta_1^2(\theta_2) \) of the exponential family so that its integration over the support is 1. For members of the same exponential family, we have:

\[
\begin{align*}
&f^{-1}(m_2(x; \theta_1, \theta_2)) = e^{x'h_1(\theta_1, \theta_2)} + p(x; \theta_1)\theta_2(1 - \alpha) b_2(x) \\
&= e^{x'h_1(\theta_1, \theta_2)} p(x; \theta_1 + (1 - \alpha) \theta_2) \\
&= e^{x'h_1(\theta_1, \theta_2)} b_2(x) \\
&= e^{-x'h_1(\theta_1, \theta_2)} b_2(x)
\end{align*}
\]

Thus

\[ P_x \leq w_1^3 w_2^{-2} e^{-x'h_1(\theta_1, \theta_2)} \int p(x; \theta_1 + (1 - \alpha) \theta_2) \, dx. \]  

Since the natural parameter space \( \Theta \) is convex for exponential families, we have \( \theta_1^2(\theta_2) = \theta_2 + (1 - \alpha) \theta_2 \in \Theta \) and therefore \( \int p(x; \theta_1 + (1 - \alpha) \theta_2) \, dx = 1 \). We end up with:

\[ P_x \leq \min_{\Theta} w_1^3 w_2^{-2} e^{-x'h_1(\theta_1, \theta_2)}. \]  

**Definition 4.** The \( \alpha \)-Chernoff distance is an asymmetric statistical distance defined by \( \rho_\alpha(p_1, p_2) = -\log p_1^\alpha(x)p_2^{1-\alpha}(x) \, dx \geq 0 \). The Bhattacharyya symmetric distance \( \rho_2(p_1, p_2) \) is a particular member of the family of the \( \alpha \)-Chernoff distances.

Thus we always have \( \rho_2(p_1, p_2) \geq \rho_\alpha(p_1, p_2) = \min_{\alpha \in (0, 1)} \rho_\alpha(p_1, p_2) \).

The optimal Chernoff bound is obtained for the optimized weight \( x' \) that has been characterized geometrically on the statistical manifolds [15].

To derive other Chernoff-type upper bounds, we shall therefore consider non-exponential families of distributions. The most prominent family, to start with, is the Cauchy family (a member of the Student \( t \)-distribution families) and the multivariate Pearson type VII elliptical distributions [19] (with its scaled multivariate \( t \)-distributions).

### 4.2. Harmonic means and the Chernoff-type bound for Cauchy distributions

Consider the family of Cauchy distributions with density:

\[ p(x; s) = \frac{1}{\pi} \frac{s}{x^2 + s^2}, \]  

defined over the support \( \mathbb{R} \). This is a scale family with heavy tails indexed by a scale parameter \( s \):

\[ p(x; s) = \frac{1}{\pi} p_0(x), \quad p_0(x) = \frac{1}{\pi} \frac{1}{x^2 + 1}, \]  

where \( p_0(x) \) denotes the standard Cauchy distribution \( C_0 \). Cauchy distributions do not belong to the exponential families. (Indeed, the mean is undefined.) Let us take the harmonic mean \( M_H = M_1 \) defined for the strictly monotonous generator \( f(x) = f^{-1}(x) = \frac{1}{x} \).

We have:

\[ P_x = \int \min(w_1 p_1(x), w_2 p_2(x)) \, dx \leq M_f(w_1 p_1(x), w_2 p_2(x); x) \, dx. \]  

Wlog., to simplify calculations exhibiting the method, consider \( w_1 = w_2 = \frac{1}{2} \).

\[ P_x \leq \int \frac{1}{2} p_1(x) p_2(x) \, dx \leq \int \frac{1}{2} (1 - \alpha)p_1(x) + \alpha p_2(x) \, dx \]

\[ \leq \int \frac{1}{2} (1 - \alpha) \frac{1}{\pi(\theta_1^2 + \theta_2^2)} + \alpha \frac{1}{\pi(\theta_2^2 + \theta_3^2)} \, dx. \]

\[ \leq \frac{1}{2} \int \frac{1}{\pi(1 - \alpha)s_1 \left( s_1^2 + s_2^2 + 2s_3(x^2 + s_1^2) \right)} \, dx. \]

\[ \leq \frac{1}{2} \int \frac{1}{\pi(1 - \alpha)s_1 + 2s_3 \left( s_1^2 + s_3 s_1^2 + 2s_3^2 \right)} \, dx. \]

\[ \leq \frac{1}{2} \left( \frac{s_1 s_2}{s_1} - \frac{s_3^2 s_1}{2s_3} + \frac{s_1 s_2 s_3}{s_1 s_2} \right) \]

since \( s_3 > 0 \) belongs to the parameter space \( \Theta \), with:

\[ s_3 = \sqrt{\frac{(1 - \alpha)s_1 s_2 + 2s_3^2}{(1 - \alpha)s_1 + 2s_3^2}}. \]

Thus the weighted harmonic mean provides a Chernoff-type upper bound:

\[ P_x \leq \frac{\frac{1}{2} s_1 s_2}{\left( (1 - \alpha)s_1 + 2s_3^2 \right)} \]  

A sanity check \( s_1 = s_3 = s \) shows that \( P_x = \frac{1}{2} \) as expected (class-conditional distributions are not distinguishable).

The Bhattacharyya-type bound obtained for \( x = \frac{1}{2} \) yields:

\[ s_3 = \frac{s_1 s_2^2 + s_3^2 s_1^2}{s_1 s_2}, \]  

and the upper bound:

\[ P_x \leq \frac{s_1 s_2}{(s_1 + s_2)(s_1 s_2^2 + s_3^2 s_1^2)} \]  

The harmonic mean \( M_H(a, b; x) \) is linear-scale free:

\[ M_H(a, b; x) = \lambda M_H(a, b; x). \]  

Then we write the probability of error as:

\[ P_x \leq \frac{1}{2} \left( \frac{\frac{1}{2} s_1 s_2}{\left( (1 - \alpha)s_1 + 2s_3^2 \right)} \right)^{\frac{1}{2}} \]

In particular, we upper bound the probability of error by the following Bhattacharyya bound:

\[ P_x \leq \left( \frac{x'}{2} \right) \]

The Chernoff-type bound proceeds by minimizing Eq. (80) with respect to \( x \). We have \( P_x^{(0)} = P_x^{(1)} = \frac{1}{2} \). To find the minimum value over the \( \alpha \)-range \([0, 1]\), we study function \( P_x^{(0)} \). Using a computer-algebra system, we find that it is a convex function that always admits a minimum at \( x = \frac{1}{2} \). Indeed, the derivative of \( P_x \) with respect to \( x \) is:

\[ P_x^{(0)}(x) = \sqrt{x} \]

\[ P_x^{(1)}(x) = \frac{1}{x + 1}. \]  

\[
\frac{d}{dz} p(z; \lambda) = \frac{1}{2}(\lambda - 1)^2 z^2 / (2z - 1) / (\lambda z + 1) / (2z + 1),
\]
that is zero if and only if \( z = \frac{1}{2} \). This is a remarkable example that shows that the Chernoff bound amounts to the Bhattacharyya bound.

Let us compute the total variation distance between two scaled Cauchy distributions \( \alpha_1 p(x; s_1) \) and \( \alpha_2 p(x; s_2) \) defined over the real-line support \( \mathbb{R} \) with:
\[
p(x; s) = \frac{s}{\pi x^2 + s^2}
\]
For \( s_1 \neq s_2 \), the two distinct positive densities intersect in exactly two values of the support:
\[
x_1 = \frac{s_1 s_2 (a_2 s_1 - a_1 s_2)}{\sqrt{a_1 s_1 - a_2 s_2}},
\]
\[
x_2 = \frac{s_1 s_2 (a_2 s_1 - a_1 s_2)}{\sqrt{a_1 s_1 - a_2 s_2}}.
\]
In particular, when \( a_1 = a_2 = 1 \), we have \( x_1 = -\sqrt{s_1 s_2} \) and \( x_2 = \sqrt{s_1 s_2} \). By abuse of notations, we let \( x_0 = -\infty \) and \( x_3 = \infty \), and apply the generic 1D total variation formula of Eq. (114) with \( k = 2 \):
\[
TV(p_1, p_2) = \frac{1}{2} \sum_{i=1}^{2} \left| P_i (x_i) + P_i (x_{i+1}) - P_i (x_{i-1}) - P_i (x_i) \right|.
\]
Since the Cauchy scaled cumulative distribution is \( P_i (x) = a_i (\frac{1}{2} \arctan \left( \frac{x}{b} \right) + \frac{1}{2}) \) with \( P_i (x_0) = 0 \) and \( P_i (x_3) = a_i \).

In particular, the probability of error when \( w_1 = w_2 = \frac{1}{2} \) is:
\[
P_e = \frac{1}{2} (1 - TV(p_1, p_2)),
\]
\[
TV(p_1, p_2) = |P_1 (x_1) - P_1 (x_2) - P_2 (x_1) + P_2 (x_2)|.
\]
That is, we find an exact analytic expression for the total variation (and hence for Bayes error \( B_0 \) and the probability of error \( P_e \)):
\[
TV(p_1, p_2) = \frac{1}{2} \left( \arctan \left( \frac{x_2}{s_1} \right) - \arctan \left( -\frac{x_2}{s_1} \right) \right)
\]
\[
+ \arctan \left( -\frac{s_2}{s_1} \right) - \arctan \left( \frac{s_2}{s_1} \right).
\]
Using the identity \( \arctan(x) + \arctan(1/x) = \frac{\pi}{2} \) and the fact that \( \arctan(-x) = -\arctan(x) \), we get a more compact formula:
\[
TV(p_1, p_2) = \frac{2}{\pi} \left( \arctan \left( \frac{s_2}{s_1} \right) - \arctan \left( \frac{s_2}{s_1} \right) \right).
\]

**Remark 5.** Note that for Cauchy distributions with scale parameters \( s_1 \) and \( s_2 \), we have \( TV(s_1, s_2) = TV(s_1, s_2), \forall s > 0 \) (because \( \sqrt{s_1 s_2} = \sqrt{s_1 s_2} \)). Therefore, we may renormalize by considering \( s_1 = 1 \) and \( s_2 = s_1 \).

It follows that the probability of error is:
\[
P_e = \frac{1}{2} - \frac{1}{\pi} \left( \arctan \left( \frac{\sqrt{2}}{2} \right) - \arctan \left( \frac{1}{\sqrt{2}} \right) \right).
\]

Consider the following numerical example: \( s_1 = 10 \) and \( s_2 = 50 \) (\( w_1 = w_2 = \frac{1}{2} \)). Then, we get \( P_e \approx 0.2677 \), the Bhattacharyya-type bound \( B \approx 0.3726 \) and the Chernoff-type bound \( C = B \).

**Remark 6.** Let us study the tightness of the upper bound. We can express analytically the gap as \( \Delta = C - P_e = C - \frac{1}{2} + TV(p_1, p_2) \). That is, we get:
\[
\Delta(\lambda) = \frac{\sqrt{2}}{1 + \lambda} - 1 + \frac{2}{\pi} \arctan \left( \frac{\sqrt{2}}{\lambda} \right) > 0.
\]

Note that \( \lambda = 1 \), we have \( \Delta(1) = 0 \): That is, the gap is tight when distributions coincide. Using a computer-algebra system, we find that the gap is maximized at \( \lambda = \frac{2 + \sqrt{2}}{2} \approx 4.5 \) and \( \Delta_{\text{max}} \approx 0.1 \).

This Cauchy case study illustrates well that the geometric mean may not always be the most appropriate weighted mean to use to upper bound \( P_e \) Note that for Cauchy distributions, we also obtained an analytic form of \( P_e \) and \( B_0 \) using the total variation distance expressed using the cumulative distribution.

We showed that the harmonic mean is tailored to derive closed-form bound for the Cauchy distributions. However, we may apply the harmonic mean to other distributions. This has in fact be done in the literature detailed in the following remark:

**Remark 7.** The harmonic mean \( M_H(a, b) = \frac{ab}{a+b} \) has been used for defining the nearest neighbor error bound [7], always better than the Bhattacharyya bound (obtained for \( x = \frac{1}{2} \)).

(See for exponential families, the nearest neighbor error bound is not available in closed-form.)

### 4.3. Pearson type VII distributions

Consider the \( d \)-dimensional elliptically symmetric Pearson type VII distribution \([19,22]\) with density:
\[
p(x; \mu, \Sigma, \lambda) = \pi^{-\frac{d}{2}} \left| \frac{1}{\Gamma(\frac{\lambda}{2})} \right| \Sigma^{-\frac{1}{2}} \left( 1 + (x - \mu) \Sigma^{-1} (x - \mu) \right)^{-\frac{\lambda}{2}}.
\]
where \( \Gamma(\gamma) = \int_0^\infty t^{\gamma-1} e^{-t} \, dt \) denotes the Gamma function extending the factorial function (i.e., \( \Gamma(1) = 1, \Gamma(n+1) = n! \)) and \( \Gamma(n) = (n-1)! \) for \( n \in \mathbb{N} \). Parameter \( \lambda > \frac{1}{d} \) represents the degree of freedom [19].

For sake of simplicity, wlog, let us consider the zero-centered sub-family with \( \mu = 0, v = 1 \) and \( \lambda > \frac{1}{d} \) fixed. This sub-family is defined on the cone of symmetric positive definite matrices \( \Theta = \{ \Sigma | \Sigma > 0 \} \) with density:
\[
p(x; \Sigma) = c_d(\lambda) \left| \Sigma \right|^{-\frac{d}{2}} \left( 1 + x^T \Sigma^{-1} x \right)^{-\frac{\lambda}{2}},
\]
where \( c_d(\lambda) = \pi^{-\frac{d}{2}} \left( \frac{\Gamma(1/2)}{\Gamma(d/2)} \right) \) is the normalizing constant [13]. This multivariate family does not belong to the exponential families. Consider the \( x \)-weighted \( f \)-mean with \( f(x) = x \), for prescribed \( \lambda > \frac{1}{d} \) and \( f^{-1}(x) = x^2 \).

We have:
\[
x f(p_1) = x c_d(\lambda) \left| \Sigma_1 \right|^{-\frac{d}{2}} (1 + x^T \Sigma_1^{-1} x),
\]
\[
(1 - x) f(p_2) = (1 - x) c_d(\lambda) \left| \Sigma_2 \right|^{-\frac{d}{2}} (1 + x^T \Sigma_2^{-1} x).
\]

Let \( c_1 = c_d(\lambda) \left| \Sigma_1 \right|^{-\frac{d}{2}} \) and \( c_2 = c_d(\lambda) \left| \Sigma_2 \right|^{-\frac{d}{2}} \). Denote by:
\[
c_x = x c_1 + (1 - x) c_2,
\]
\[
= c_d(\lambda) \left( x \left| \Sigma_1 \right|^{-\frac{d}{2}} + (1 - x) \left| \Sigma_2 \right|^{-\frac{d}{2}} \right).
\]
we get:
\[
x f(p_1) + (1 - x) f(p_2) = x c_1 (1 + x^T \Sigma_1^{-1} x) + (1 - x) c_2 (1 + x^T \Sigma_2^{-1} x),
\]
\[
= c_1 (1 + x^T \Sigma_1^{-1} x),
\]
with
The left hand side has been elucidated in Eq. (44) and the right hand side is the Chernoff-type/Bhattacharyya-type similarity coefficients which should be available in closed-form for fast calculation. We are interested in characterizing the gaps \( \Delta = \rho_2(p_1, p_2) - P_e \) and \( \Delta_2 = \rho_2(p_1, p_2) - P_e \) between the Bhattacharyya/Chebyshev upper bounds and \( P_e \) (with \( \Delta \leq \Delta_2 \)). We start with some simple cases of univariate distributions, where \( P_e \) can be expressed analytically, and then considered the multivariate distributions where \( P_e \) need to be stochastically estimated.

4.5.1. Simple cases: univariate distributions

For univariate densities, we may calculate the total variation by computing the roots of \( p_1(x) = p_2(x) \) and then using the cumulative distributions \( P(t) = \int p(x \leq t) \) to explicit a formula. Assume \( x_1, \ldots, x_k \) are the \( k \) roots, and by abuse of notations, let \( x_0 = x_{\min} \) and \( x_{k+1} = x_{\max} \) denote the extra endpoints of the distribution support \( x \in [x_{\min}, x_{\max}] \). When the support is the full real line \((\text{supp}(p_1) = \mathbb{R})\), we set \( x_{\min} = -\infty \) and \( x_{\max} = \infty \). We have:

\[
\text{TV}(p_1, p_2) = \frac{1}{2} \sum_{i=1}^{k+1} \left| \int_{x_{i-1}}^{x_i} (p_1(x) - p_2(x)) \, dx \right|.
\]

This scheme allows one to compute the total variation distance of many univariate distributions like the Gaussian (see Eq. (A.13)), Rayleigh, Cauchy, etc. distributions. Therefore, we get analytic expressions of Bayes error relying on the cumulative distribution function (CDF) for many univariate distributions.

Those the Bhattacharyya gaps for the Cauchy or 1D Gaussians can be written analytically. Section 4.2 already addressed the gap for the Cauchy distributions. Next, we consider the general multivariate case (that includes those univariate examples) and report on our numerical experiments.

4.5.2. The case of multivariate distributions

We consider multivariate t-distributions \([8, 9]\) (MVT) that includes the univariate normal (MVN) distributions in the limit case (when the number of degree of freedom tends to infinity). To perform experiments, we used the mvtnorm package\(^6\) on the R software platform.\(^5\)

Since the total variation (TV) does not admit a closed-form formula (nor the probability of error \( P_e \)), we estimate those quantities by performing stochastic integrations as follows:

\[
\widehat{P}_e(p_1, p_2) = \frac{1}{2} (1 - \text{TV}(p_1, p_2)).
\]

\[
\widehat{\text{TV}}(p_1, p_2) = \frac{1}{2n} \sum_{i=1}^{n} \frac{1}{p_i(x)} \left| p_i(x) - p_2(x) \right|.
\]

where \( x_1, \ldots, x_n \) are \( n \) identically and independently variates of \( p_1 \). Stochastic integration guarantees convergence to the true value in the limit: \( \lim_{n \to \infty} \widehat{P}_e(p_1, p_2) = P_e(p_1, p_2) \).

To give a numerical example, consider central bidimensional t-distributions with \( v = 6 \) and scale matrices \( \Sigma_1 = I \) and \( \Sigma_2 = 10 I \), where \( I \) denotes the identity matrix. Running the mvtnorm package\(^6\) code, we get the following

\[
\text{http://cran.r-project.org/web/packages/mvtnorm/index.html. To install the package, we used the command line: install.packages('mvtnorm_0.9-99996.zip', repos = NULL).}

\[ R \text{ can be freely downloaded at http://www.r-project.org.}
\]

\[ \text{The code is freely available at http://www.informationgeometry.org/TestMVR.r.} \]
estimates for $\overline{TV}$: 0.3210709 ($n = 100$), 0.3479519 ($n = 1000$), 0.3472926 ($n = 10000$), 0.347538 ($n = 100000$). We chose $n = 10000$ in the following experiments.

Consider central $t$-distributions ($\mu = 0$). We implemented the closed-form formula of Eq. (109) to calculate the $\varepsilon$-Chernoff coefficient $\rho_{\text{MVT}}(\Sigma, \Sigma)$. The optimal Chernoff coefficient (and the exponent $x^*$) is approximated by discretizing in 1000 steps the unit range for $\alpha$. We consider $\Sigma = I$ and $\Sigma = \lambda I$ (with $\lambda = 6$) for $d = 4$ and various values of the dimension. (Indeed, after an appropriate “whitening” transformation [7], we may assume wlog. that one parameter matrix is the identity while the other is diagonal.) We report the experimental results in Table 1.

**Table 1**

<table>
<thead>
<tr>
<th>Dimension</th>
<th>$\lambda$</th>
<th>$\overline{P}_I$</th>
<th>$\rho_{\text{MVT}}$</th>
<th>$\overline{\Delta}$</th>
<th>$\rho_{\text{MVT}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 2$</td>
<td>$\lambda = 3$</td>
<td>0.3302</td>
<td>0.4471817</td>
<td>0.455</td>
<td>0.4475742</td>
</tr>
<tr>
<td>$d = 3$</td>
<td>$\lambda = 4$</td>
<td>0.2578</td>
<td>0.3951277</td>
<td>0.462</td>
<td>0.3956298</td>
</tr>
<tr>
<td>$d = 5$</td>
<td>$\lambda = 6$</td>
<td>0.16215</td>
<td>0.2943599</td>
<td>0.487</td>
<td>0.2944589</td>
</tr>
<tr>
<td>$d = 10$</td>
<td>$\lambda = 11$</td>
<td>0.06045</td>
<td>0.1400655</td>
<td>0.548</td>
<td>0.141438</td>
</tr>
<tr>
<td>$d = 15$</td>
<td>$\lambda = 16$</td>
<td>0.02845</td>
<td>0.0744227</td>
<td>0.592</td>
<td>0.0784162</td>
</tr>
<tr>
<td>$d = 20$</td>
<td>$\lambda = 21$</td>
<td>0.0167</td>
<td>0.0439694</td>
<td>0.625</td>
<td>0.0494522</td>
</tr>
</tbody>
</table>

**5. Conclusion**

In this paper, we first reported a formula relating the Bayes error $B_\varepsilon$ (including the probability of error $P_I$) to the total variation metric $TV$ defined on scaled distributions (see Theorem 1). Second, we elucidated the Chernoff upper bound mechanism based on generalized weighted means: Chernoff [4] used the fact that $\min(a, b) \leq a^b - 1$ for $a, b > 0$ and $a \in [0, 1]$ to derive an upper bound that turned out to be well-suitied to the structure of exponential families [11]. We interpreted the right-hand side of this inequality as a weighted geometric mean, and considered extending the upper bound construction using generalized weighted mean $M(a, b; \lambda)$. A mean $M(a, b; \lambda)$ is indeed guaranteed to fall within its extrema by definition, thus yielding the bounds: $\min(a, b) \leq M(a, b; \lambda) \leq \max(a, b)$. We considered the family of quasi-arithmetic means [12,13,14] and showed how to derive new upper bounds by coupling the structure of the generalized mean with the structure of the probability distribution family at hand. We illustrated our method by considering three examples: The univariate Cauchy distributions, and the multivariate Pearson type VII and $t$-distributions (that includes the multivariate normal distributions in the limit case). For those families, we designed new affinity coefficients upper bounding $B_\varepsilon$. The best value $x^*$ of $\alpha$ yielding the tightest coefficient can be found by optimization on the statistical manifold [15]. We carried out numerical experiments that show that those novel upper bounds are helpful because not too distant to Bayes error (although not very tight), specially because they can be calculated in constant time using closed-form formula. Otherwise, for more precise approximations, the Bayes error can be estimated using computationally-intensive stochastic integrations.

Last but not least, this paper revealed novel interactions between the Bayes error and statistical diversifications: We show how to design $P_I(p_1, p_2) \leq P_I^{\mu_1}(p_1, p_2)$ upper bounds where $\rho_\mu$ is an affinity coefficient derived from a weighted quasi-arithmetic mean $M_\mu$ (skewed for Chernoff type and symmetric for a Bhattacharyya type). Since we can transform any affinity coefficient $\rho_\mu$ into a corresponding divergence by defining $D_\mu = -\log \rho_\mu$, we deduce that those novel statistical diversities $D_\mu$ can be used to upper bound the probability of error: $P_I(p_1, p_2) \leq e^{-D_\mu(p_1, p_2)}$.

**Acknowledgments**

The author is grateful for the valuable comments of the Reviewers that led to this revised work.

**Appendix A. Bayes error for class-conditional Gaussians**

We summarize the formula for the Bayes error $B_\varepsilon$ and the probability of error $P_I$ when dealing with univariate and multivariate normal class-conditional distributions:

$$B_\varepsilon = \frac{a_1 + a_2}{2} - TV(a_1 p_1, a_2 p_2),$$  
(A.1)

$$P_I = \frac{1}{2} - TV(w_1 p_1, w_2 p_2),$$  
(A.2)

where $a_1 = w_1 (c_1 + c_2)$ and $a_2 = w_2 (c_{12} + c_{22})$ are the positive weights derived from the cost design matrix and the a priori class weights.

First, consider the family of univariate normal distributions $\mathcal{F} = \{N(\mu, \sigma) | \mu \in R, \sigma \in (0, \infty)\}$. Consider two distinct normal densities $p_1$ and $p_2$. The two weighted densities $a_1 p_1(x) = a_2 p_2(x)$ intersect at exactly two positions when $\sigma_1 \neq \sigma_2$ or in exactly one position, otherwise:

$$\frac{a_1}{\sigma_1 \sqrt{2\pi}} e^{-\frac{(x - \mu_1)^2}{2\sigma_1^2}} = \frac{a_2}{\sigma_2 \sqrt{2\pi}} e^{-\frac{(x - \mu_2)^2}{2\sigma_2^2}}.$$  
(A.3)

Finding the roots amounts to solve the quadratic equation:

$$\left(\frac{X - \mu_1}{\sigma_1}\right)^2 - \left(\frac{X - \mu_2}{\sigma_2}\right)^2 - 2 \log \frac{a_1 \sigma_2}{a_2 \sigma_1} = 0.$$  
(A.4)

When $\sigma_1 = \sigma_2 = \sigma$ (with $\mu_1 \neq \mu_2$), we have one root:

$$x_1 = \frac{\mu_1^2 - \mu_2^2 - 2 \log \frac{a_1 \sigma_2}{a_2 \sigma_1}}{2 (\mu_1 - \mu_2)}.$$  
(A.5)

We find:

$$TV(a_1 p_1, a_2 p_2) = \frac{1}{2} \left[a_1 \Phi \left(x_1; \mu_2, \sigma_2\right) - a_2 \Phi \left(x_1; \mu_1, \sigma_1\right)\right],$$  
(A.6)

where $\Phi(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} (x - \mu)^2}$ is the cumulative distribution, and $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2} dt$ denotes the error function. That is,

$$B_\varepsilon = \frac{a_1 + a_2}{2} - \frac{1}{2} \left[a_2 \text{erf} \left(\frac{x_1 - \mu_2}{\sigma_2 \sqrt{2}}\right) - a_1 \text{erf} \left(\frac{x_1 - \mu_1}{\sigma_1 \sqrt{2}}\right)\right].$$  
(A.7)

$$x_1 = \frac{\mu_1^2 - \mu_2^2 - 2 \log \frac{a_1 \sigma_2}{a_2 \sigma_1}}{2 (\mu_1 - \mu_2)}.$$  
(A.8)

When $\sigma_1 \neq \sigma_2$, the quadratic equation expands as $ax^2 + bx + c = 0$, and we have two distinct roots $x_1$ and $x_2$:

$$a = 1, \quad b = \frac{\mu_2}{\sigma_2} - \frac{\mu_1}{\sigma_1}, \quad c = 0.$$  
(A.9)

$$b = 2 \frac{\mu_2}{\sigma_2} \frac{\mu_1}{\sigma_1},$$  
(A.10)

$$c = \frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} - 2 \log \frac{a_1 \sigma_2}{a_2 \sigma_1}.$$  
(A.11)

$$x_1 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}, \quad x_2 = \frac{-b + \sqrt{b^2 - 4ac}}{2a},$$  
(A.12)

with $\Delta = b^2 - 4ac \geq 0$ and the total variation writes as follows:
TV(a_1,p_1, a_2,p_2) = \frac{1}{2} \left( \text{erf} \left( \frac{x_1 - \mu_1}{\sigma_1 \sqrt{2}} \right) - \text{erf} \left( \frac{x_1 - \mu_2}{\sigma_2 \sqrt{2}} \right) \right) \\
+ \left| \text{erf} \left( \frac{x_2 - \mu_1}{\sigma_1 \sqrt{2}} \right) - \text{erf} \left( \frac{x_2 - \mu_2}{\sigma_2 \sqrt{2}} \right) \right| \\
(A.13)

Those formula generalize the probability of error reported in [3, p. 1375] to the most general case.

Second, consider the family of multivariate normals distributions \(\{N(\mu, \Sigma) \mid \mu \in \mathbb{R}^d, \Sigma = 0\}\). For densities \(p_1\) and \(p_2\) having the same covariance matrix \(\Sigma\), the probability of error is reported in [17], even for degenerate covariance matrices \(\Sigma\) by taking the pseudo-inverse matrix \(\Sigma^+\):

\[
P_e = 1 - \frac{1}{2} \text{erf} \left( \frac{1}{2 \sqrt{2}} \left\| (\Sigma^+)^{-\frac{1}{2}} (\mu_2 - \mu_1) \right\| \right). \\
(A.14)

When covariance matrices are distinct \((\Sigma_1 \neq \Sigma_2)\) but linear classifiers are considered, we also get a closed-form formula [18] for the probability of error. Otherwise, for the general case of quadratic classifiers of multivariate normals with distinct covariance matrices, no analytical formula is known. The best way to compute the probability of error is then by performing 1D integration of the conditional density of the discriminant function [7]. Note that since two matrices can always be simultaneously diagonalized [7], it is enough to consider the case of two Gaussians with the first covariance being set to the identity matrix \(I\) and the second covariance matrix set to a diagonal matrix \(\Lambda\).

References