Novel Tight Classification Error Bounds under Mismatch Conditions based on f-Divergence

Ralf Schlüter[1], Markus Nussbaum-Thom[1], Eugen Beck[1], Tamer Alkhouli[1], and Hermann Ney[1,2]
[1]Lehrstuhl für Informatik 6, Computer Science Dept., RWTH Aachen University, Aachen, Germany
[2] Spoken Language Processing Group, LIMSI CNRS, Paris, France
{schlueter, nussbaum, ney}@cs.rwth-aachen.de

***Abstract*—By default, statistical classification/multiple hypothesis testing is faced with the model mismatch introduced by replacing the true distributions in *Bayes* decision rule by model distributions estimated on training samples. Although a large number of statistical measures exist w.r.t. to the mismatch introduced, these works rarely relate to the mismatch in accuracy, i.e. the difference between model error and *Bayes* error. In this work, the accuracy mismatch between the ideal *Bayes* decision rule/*Bayes* test and a mismatched decision rule in statistical classification/multiple hypothesis testing is investigated explicitly. A proof of a novel generalized tight statistical bound on the accuracy mismatch is presented. This result is compared to existing statistical bounds related to the total variational distance that can be extended to bounds of the accuracy mismatch. The analytic results are supported by distribution simulations.**

I. Introduction

In statistical classification (also cf. multiple hypothesis testing), the most important performance criterion is the classification error. The classification error is minimized by *Bayes*' decision rule (*Bayes*' test), which implies knowledge of the true probability distributions underlying the classification problem. Nevertheless, in practice usually model distributions estimated from training samples are substituted for the true distributions in *Bayes*' decision rule. The implications of this approach on the classification error are not well understood, and corresponding statistical bounds on the accuracy mismatch, i.e. the difference between *Bayes*' error and the error of the mismatched decision rule are rarely discussed to the knowledge of the authors. In [5], one of the authors introduced novel bounds on the accuracy mismatch with applications to model estimation/training. In [8], also a training criterion representing a model-based smoothed classification error was shown to provide an upper bound on the *Bayes* error.

In this work, statistical bounds on the accuracy mismatch between the *Bayes* decision rule involving the true distributions and mismatched decision rules are analyzed. Existing bounds on the total variational distance are related to the accuracy mismatch in Sec. II. A proof of a generalized tight upper bound in terms of *f-Divergence*s [7] is provided in Sec. III. Special cases of this generalized bound will be discussed supported by simulations and compared to existing bounds, supported by distribution simulations in Sec. IV.

II. Existing Classification Error Bounds

Assume a statistical classification problem with a continuous observation space $\mathcal{X}$ and a set of classes $\mathcal{C}$. $pr(x,c)$ denotes the true joint probability density, with observations $x \in \mathcal{X}$, and classes $c \in \mathcal{C}$. Assuming uniform cost for classification errors, *Bayes* decision rule is defined by

$$c_{pr}(x) := \arg\max_c pr(c|x) = \arg\max_c pr(x,c).$$

The corresponding *Bayes* accuracy, i.e. optimal average probability of success using the bayes rule then can be written as:

$$A_{pr} = \int_{\mathcal{X}} pr(x, c_{pr}(x))\ dx.$$

In statistical classification the true distributions usually are unknown and need to be estimated from samples. The standard approach here is to replace the true distributions in *Bayes* decision rule by estimated model probability distributions $q(x,c)$, leading to the mismatched decision rule:

$$c_q(x) := \arg\max_c q(c|x) = \arg\max_c q(x,c).$$

The accuracy of the mismatched decision rule, i.e. the average probability of success using the mismatched rule c_q then is

$$A_q = \int_{\mathcal{X}} pr(x, c_q(x))\ dx.$$

The effect of the mismatch between model and true distributions can be represented by the corresponding accuracy mismatch between *Bayes* and mismatched decision rule:

$$\Delta := A_{pr} - A_q = \int pr(x, c_{pr}(x)) - pr(x, c_q(x)))\ dx.$$

Instead of the accuracy mismatch, in the literature, often the total variational distance V [1, p. 369] is discussed:

$$V := \int \sum_{c\in\mathcal{C}} |pr(x,c) - q(x,c)|\ dx,$$

for which a number of bounds based on the *Kullback-Leibler* distance D_{KL} or relative entropy [1, p. 19] have been derived:

$$D_{\mathrm{KL}}(pr||q) := \int \sum_{c\in\mathcal{C}} pr(x,c) \log\left(\frac{pr(x,c)}{q(x,c)}\right) dx.$$

From machine learning, the *Bretagnolle-Huber* bound has been known for some time in the context of density estimation [10, p.30]:

$$V^2 \leq 4(1 - \exp(-D_{\mathrm{KL}}(pr|q))). \tag{1}$$

Also, the *Pinsker* inequality connects total variational distance and *Kullback-Leibler* distance [1, p. 370] in the following way:

$$V^2 \leq 2D_{\mathrm{KL}}(pr||q). \tag{2}$$

An improvement of *Pinsker*'s inequality was presented by *Vajda* [2]:

$$D_{\mathrm{KL}}(pr||q) \geq \log\left(\frac{2+V}{2-V}\right) - \frac{2V}{2+V}. \tag{3}$$

Vajda also proposed a tight bound, for which *Fedotov* presented a parametrized form in [2]. Unfortunately, both bounds do not have explicit representations in V.

Unfortunately, the bounds presented do not relate directly to the mismatch in classification accuracy. In [5], a number of bounds were presented which relate the accuracy mismatch to statistical measures. Specifically, a bound of the local accuracy mismatch (i.e. conditioned on the observation x) on the (local) variational distance is presented. The corresponding derivation presented in [5] can be generalized to the global case:

$$\Delta \leq V \tag{4}$$

We will not provide a proof for this global case here, since it will be covered as a special case of the proposed *f-Divergence* based bound presented in this work, cf. IV-E.

Due to monotonicity, Ineq. (4) can be substituted into Ineqs. (1-3) and *Fedotov*'s inequality, i.e. the total variational distance V can be replaced by the accuracy mismatch Δ. This leads to a number of inequalities, bounding the accuracy mismatch in the same way as the total variational distance by functions of the *Kullback-Leibler* distance.

In [5], also a global bound of the accuracy mismatch on the *Kullback-Leibler* distance was derived directly:

$$\frac{\Delta^2}{2} \leq D_{\mathrm{KL}}(pr||q).$$

Nevertheless, this bound is identical to the case of substituting Ineq. (4) into the *Pinsker* Ineq. (2).

Fig. 1 shows plots of the derived bounds discussed above.

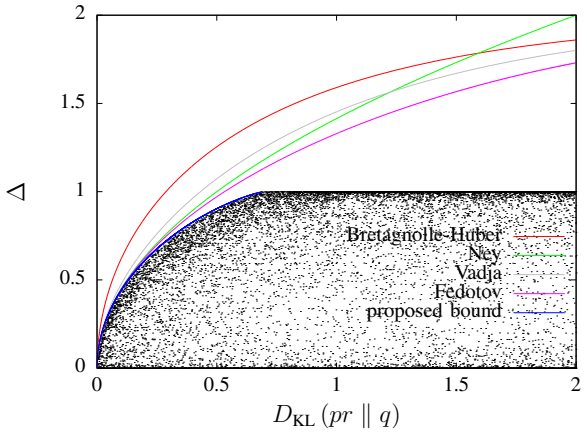

Fig. 1. *Bretagnolle-Huber, Pinsker, Vajda, and Fedotov bound on the accuracy mismatch parametrized by the Kullback-Leibler distance, compared to simulations of pairs of distributions. The tight Kullback-Leibler bound is a special case of the f-Divergence bound derived in this work. The black dots refer to results from generating distributions via simulations.*

To further analyze the relation between accuracy mismatch and *Kullback-Leibler* distance, Fig. 1 further shows simulations of pairs of probability distributions, with each black dot representing the result of a single simulation. As could be observed, the simulations suggest that the bounds obtained by substituting Ineq. (4) into *Kullback-Leibler* bounds on the total variational distance are not tight. Instead, the simulations suggest a bound of the form:

$$D_{\mathrm{KL}}(pr||q) \geq \frac{1}{2}\left[(1+\Delta)\log(1+\Delta) + (1-\Delta)\log(1-\Delta)\right],$$

which also is plotted in Fig, 1. Later, it will be shown that this bound is covered by the proposed *f-Divergence* based bound presented in this work, cf. Sec. IV-A.

Due to the symmetry of the variational distance, note that the above Ineqs. relating variational distance and *Kullback-Leibler* distance $D_{\mathrm{KL}}(pr||q)$ also are fulfilled for the *reverse Kullback-Leibler* distance (or likelihood disparity [7]) $D_{\mathrm{KL}}(q||pr)$. For this case, Fig. 2 compares the different bounds derived for the accuracy mismatch, and corresponding simulations.

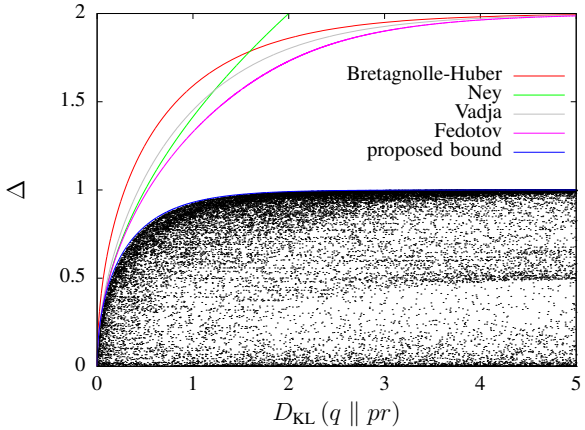

Fig. 2. *Bretagnolle-Huber, Pinsker, Vajda, and Fedotov bound on the accuracy mismatch parametrized by the reverse Kullback-Leibler distance, compared to simulations of pairs of distributions. The tight reversed Kullback-Leibler bound is a special case of the f-Divergence bound derived in this work. The black dots refer to results from generating distributions via simulations.*

Again, note that all bounds on the accuracy mismatch, which are derived via the total variational distance are not tight. As could be seen in Fig. 2, a lower, tight bound is suggested by the simulations, which is of the form:

$$\Delta \leq \sqrt{1-\exp(-2D_{\mathrm{KL}}(q||pr))}.$$

In the next section it will be shown that this bound is a special case of the *f-Divergence* bound presented in this work, cf. Sec. IV-B.

III. Error Bounds based on f-Divergence

Theorem 1: The accuracy mismatch implicitly is tightly bounded by a function of the *f-Divergence* of pr from q in the following way:

$$D_f(q||pr) \geq \frac{1}{2}\Big(f\big(1+\Delta\big) + f\big(1-\Delta\big)\Big) \tag{5}$$

with the *f-Divergence* [9]

$$D_f(q||pr) = \int_{\mathcal{X}} \sum_c q(x,c) f\Big(\frac{pr(x,c)}{q(x,c)}\Big)\, dx, \tag{6}$$

and f being a convex function $f : \mathrm{R}^+ \to \mathrm{R}$, with the specific property $f(1) = 0$.

The equality is obtained for specific true and model distributions pr' and q' with equal class conditionals

$$q'(x|c) = pr'(x|c) \qquad \forall\, x \in \mathcal{X}, c \in \mathcal{C}, \tag{7}$$

and with a specific choice of the class priors for any pair of classes $c_1, c_2 \in \mathcal{C}$ with $c_1 \neq c_2$, and $\lambda \in [0.5; 1.0]$:

$$pr'(c) = \begin{cases} \lambda & c = c_1 \\ 1-\lambda & c = c_2 \\ 0 & otherwise \end{cases}$$

and

$$q'(c) = \lim_{\epsilon \to 0^+} \begin{cases} \frac{1}{2} - \epsilon & c = c_1 \\ \frac{1}{2} + \epsilon & c = c_2 \\ 0 & otherwise, \end{cases}$$

Note: ϵ ensures $c_2 = \arg\max_c q'(c) \neq \arg\max_c pr'(c) = c_1$.

Proof: The proof of Theorem 1 is structured in the following way. A series of permutations of class labels, and transformations are performed on an arbitrary pair of true and model distributions pr and q respectively that localize the maximizing classes of true and model distribution respectively into specific constant class labels independent of the observation x, while preserving both accuracy mismatch and *f-Divergence*. Subsequently, the convexity property is used to derive a tight lower bound on the *f-Divergence* as a function of the accuracy mismatch.

A preliminary version of the proof of Theorem 1 is given in [6] under the assumption that the decisions based on model and true distribution always differ, i.e. $c_{pr}(x) \neq c_q(x)\ \forall\, x \in \mathcal{X}$. Here, the full proof is presented.

A. Permutation of Maximizing Class Labels

For each observation $x \in \mathcal{X}$, a joint permutation of the class labels in both the joint true and model distributions conserves both the local and global accuracies, as well as the *f-Divergence* between both. Accuracies are not affected, i.e. changing the identity of the maximizing class label does not have any effect, since the values of the maxima of the class posterior distributions are conserved. Also, the *f-Divergence* is conserved, since the same (joint) permutation is done on the true and the model distributions, which simply leads to a permutation of summands in the *f-Divergence* only. Therefore, without loss of generality, in the following we assume that for each observation $x \in \mathcal{X}$ the class resulting from *Bayes* decision rule is $c_{pr}(x) = c_1$, and, for those cases where *Bayes* and mismatched decision rule differ, i.e., $c_q(x) \neq c_{pr}(x)$, the maximizing class of the mismatched decision rule is $c_q(x) = c_2$, with an arbitrary, but fixed pair of classes $c_1, c_2 \in \mathcal{C}$ with $c_1 \neq c_2$.

B. Aggregation/Equalization

Consider the aggregation of the probabilities of two summands of an *f-Divergence*, with $p_1, p_2, q_1, q_2 \in \mathbf{R}^+$. Then, using *Jensen*'s inequality, this part of the *f-Divergence* can be bounded from above:

$$q_1 f\big(\frac{p_1}{q_1}\big) + q_2 f\big(\frac{p_2}{q_2}\big) \geq (q_1+q_2)\cdot f\big(\frac{p_1+p_2}{q_1+q_2}\big) \tag{8}$$

$$= 2\frac{q_1+q_2}{2} f\left(\frac{\frac{p_1+p_2}{2}}{\frac{q_1+q_2}{2}}\right), \tag{9}$$

i.e., both aggregation and equalization of the probability distributions for a pair of classes leads to a decrease of the corresponding contribution to the *f-Divergence*.

C. Elimination of Equal Decisions

In the next step, we modify the permuted distributions to better represent the case of equal decisions. For this, the following modification is applied to the permuted true and model distributions:

$$\widehat{pr}(x,c) = \begin{cases} \dfrac{pr(x,c_1)+pr(x,c_2)}{2} & \forall\, c \in \{c_1, c_2\},\ x \in \mathcal{X}_= \\ pr(x,c) & \text{otherwise} \end{cases}$$

$$\widehat{q}(x,c) = \begin{cases} \dfrac{q(x,c_1)+q(x,c_2)}{2} & \forall\, c \in \{c_1, c_2\},\ x \in \mathcal{X}_= \\ q(x,c) & \text{otherwise} \end{cases}$$

The accuracy mismatch of the original distributions[1] can then be simplified to:

$$\begin{aligned} \Delta = A_{pr} - A_q &= \int_{\mathcal{X}} (pr(x, c_{pr}(x)) - pr(x, c_q(x)))\, dx \\ &= \int_{\mathcal{X}_\neq} (pr(x,c_1) - pr(x,c_2))\, dx \\ &= \int_{\mathcal{X}} (\widehat{pr}(x,c_1) - \widehat{pr}(x,c_2))\, dx = \widehat{pr}(c_1) - \widehat{pr}(c_2), \end{aligned} \tag{10}$$

with

$$\widehat{pr}(c) = \int_{\mathcal{X}} \widehat{pr}(x,c)\, dx.$$

The *f-Divergence* of the modified distributions then becomes a lower bound for the *f-Divergence* of the original distributions:

$$\begin{aligned} &D_f(q||pr) - D_f(\widehat{q}||\widehat{pr}) \\ &= \int_{\mathcal{X}_=} \Bigg[q(x,c_1) f\big(\frac{pr(x,c_1)}{q(x,c_1)}\big) + q(x,c_2) f\big(\frac{pr(x,c_2)}{q(x,c_2)}\big) \\ &\qquad - \big(2\frac{q(x,c_1)+q(x,c_2)}{2} f\left(\frac{\frac{pr(x,c_1)+pr(x,c_2)}{2}}{\frac{q(x,c_1)+q(x,c_2)}{2}}\right)\Bigg]\, dx \\ &\geq 0, \quad \text{cf. Eq. (8)} \end{aligned} \tag{11}$$

i.e. we have $D_f(q||pr) \geq D_f(\widehat{q}||\widehat{pr})$.

Similarly, for the modifications of the model distribution we obtain

$$0 \leq \int_{\mathcal{X}} (q(x, c_q(x)) - q(x, c_{pr(x)}))\, dx \tag{12}$$

$$\begin{aligned} &= \int_{\mathcal{X}_\neq} (q(x,c_2) - q(x,c_1))\, dx \\ &= \int_{\mathcal{X}} (\widehat{q}(x,c_2) - \widehat{q}(x,c_1))\, dx = \widehat{q}(c_2) - \widehat{q}(c_1), \end{aligned} \tag{13}$$

[1] Note that the accuracy mismatch of the original distributions (pr, q) and the modified distributions $(\widehat{pr}, \widehat{q})$ is not necessarily identical.

with

$$\widehat{q}(c) = \int_{\mathcal{X}} \widehat{q}(x,c)\, dx.$$

D. Lower Bound for f-Divergence

Consider the *f-Divergence* of the original pair of distributions q and pr:

$$\begin{aligned}
&D_f(q||pr) \geq D_f(\widehat{q}||\widehat{pr}) \\
&\quad \text{(cf. Ineq. (11); equality for } \mathcal{X}_{\neq} = X, \text{ or } \mathcal{X}_{\neq} = \emptyset) \\
&= \int_{\mathcal{X}} \sum_c \widehat{q}(x,c) f\big(\frac{\widehat{pr}(x,c)}{\widehat{q}(x,c)}\big)\, dx \\
&= \sum_c \widehat{q}(c) \int_{\mathcal{X}} \widehat{q}(x|c) f\big(\frac{\widehat{pr}(x|c)\widehat{pr}(c)}{\widehat{q}(x|c)\widehat{q}(c)}\big)\, dx \\
&\geq \sum_c \widehat{q}(c) f\Bigg(\underbrace{\Big[\int_{\mathcal{X}} \widehat{q}(x|c)\frac{\widehat{pr}(x|c)}{\widehat{q}(x|c)}\, dx\Big]}_{=1} \frac{\widehat{pr}(c)}{\widehat{q}(c)}\Bigg) \\
&\quad \text{(convexity of } f\text{: } \textit{Jensen}\text{'s inequality; equality for} \\
&\quad\ q(x|c) = pr(x|c)\ \forall\, c \in \{c' \in \mathcal{C}|\widehat{q}(c) > 0\},\ \forall\, x \in \mathcal{X}) \\
&= \sum_c \widehat{q}(c) f\big(\frac{\widehat{pr}(c)}{\widehat{q}(c)}\big) \\
&\geq \widehat{q}(c_1) f\big(\frac{\widehat{pr}(c_1)}{\widehat{q}(c_1)}\big) + \widehat{q}(c_2) f\big(\frac{\widehat{pr}(c_2)}{\widehat{q}(c_2)}\big) \\
&\quad + \Big(\sum_{c\in\mathcal{C}\setminus\{c_1,c_2\}} \widehat{q}(c)\Big) f\big(\frac{\sum_{c\in\mathcal{C}\setminus\{c_1,c_2\}} \widehat{pr}(c)}{\sum_{c\in\mathcal{C}\setminus\{c_1,c_2\}} \widehat{q}(c)}\big) \\
&\quad \text{(applying Ineq. (8) iteratively; equality} \\
&\quad\ \text{for } \widehat{pr}(c) = \widehat{q}(c) = 0 \quad \forall\, c \in \mathcal{C}\setminus\{c_1,c_2\}, \\
&\quad\ \text{also cf. stronger condition below)} \\
&= \widehat{q}(c_1) f\big(\frac{\widehat{pr}(c_1)}{\widehat{q}(c_1)}\big) + \widehat{q}(c_2) f\big(\frac{\widehat{pr}(c_2)}{\widehat{q}(c_2)}\big) \\
&\quad + 2\frac{1-\widehat{q}(c_1)-\widehat{q}(c_2)}{2} f\big(\frac{\frac{1-\widehat{pr}(c_1)-\widehat{pr}(c_2)}{2}}{\frac{1-\widehat{q}(c_1)-\widehat{q}(c_2)}{2}}\big) \\
&\geq \frac{1+\widehat{q}(c_1)-\widehat{q}(c_2)}{2} f\big(\frac{\frac{1+\widehat{pr}(c_1)-\widehat{pr}(c_2)}{2}}{\frac{1+\widehat{q}(c_1)-\widehat{q}(c_2)}{2}}\big) \\
&\quad + \frac{1-\widehat{q}(c_1)+\widehat{q}(c_2)}{2} f\big(\frac{\frac{1-\widehat{pr}(c_1)+\widehat{pr}(c_2)}{2}}{\frac{1-\widehat{q}(c_1)+\widehat{q}(c_2)}{2}}\big) \\
&\quad \text{(cf. Ineq. (8); equality for} \\
&\quad\ \widehat{pr}(c) = \widehat{q}(c) = 0 \quad \forall\, c \in \mathcal{C}\setminus\{c_1,c_2\}) \\
&= \beta f\big(\frac{\lambda}{\beta}\big) + (1-\beta) f\big(\frac{1-\lambda}{1-\beta}\big) \qquad (14)
\end{aligned}$$

with the definitions:

$$\lambda := \frac{1}{2} + \frac{\widehat{pr}(c_1)}{2} - \frac{\widehat{pr}(c_2)}{2}, \quad \beta := \frac{1}{2} + \frac{\widehat{q}(c_1)}{2} - \frac{\widehat{q}(c_2)}{2},$$

where especially we have

$$\begin{aligned}
&2\lambda - 1 = \widehat{pr}(c_1) - \widehat{pr}(c_2) = A_{pr} - A_q = \Delta \quad \text{(cf. Eq. (10))} \\
&2\beta - 1 = \widehat{q}(c_1) - \widehat{q}(c_2).
\end{aligned}$$

Also, from $A_{pr} - A_q \geq 0$ and from Ineq. (13) follows:

$$\lambda \geq 1/2, \quad \beta \leq 1/2. \qquad (15)$$

Now assume the following definitions:

$$a := \lambda \cdot \frac{1-2\beta}{2\lambda-1} \geq 0, \quad b := (1-\lambda)\cdot\frac{1-2\beta}{2\lambda-1} \geq 0$$

Also, note the following inequality:

$$\begin{aligned}
\frac{1}{2}\big(f(2\lambda) + f(2[1-\lambda])\big) &= \frac{1}{2}\Big(f(\frac{\lambda}{1/2}) + f(\frac{1-\lambda}{1/2})\Big) \\
&\geq \big(\frac{1}{2}+\frac{1}{2}\big)\cdot f\Big(\frac{\lambda+1-\lambda}{\frac{1}{2}+\frac{1}{2}}\Big) = f(1) = 0 \qquad (16)
\end{aligned}$$

Then, continuing from Ineq. (14), we obtain:

$$\begin{aligned}
&D_f(q||pr) \\
&\geq \beta f\big(\frac{\lambda}{\beta}\big) + (1-\beta) f\big(\frac{1-\lambda}{1-\beta}\big) \\
&= \beta f\big(\frac{\lambda}{\beta}\big) + a\underbrace{f\big(\frac{a}{a}\big)}_{=0} + (1-\beta) f\big(\frac{1-\lambda}{1-\beta}\big) + b\underbrace{f\big(\frac{b}{b}\big)}_{=0} \\
&\geq (\beta+a) f\big(\frac{\lambda+a}{\beta+a}\big) + (1-\beta+b) f\big(\frac{1-\lambda+b}{1-\beta+b}\big) \\
&\quad \leq 1 \text{, cf. Ineq. (15)} \\
&= \frac{2\lambda - \overbrace{2\beta}}{2\lambda-1} \underbrace{\frac{1}{2}\big(f(2\lambda) + f(2(1-\lambda))\big)}_{\geq 0 \text{, cf. Ineq. (16)}} \\
&\geq \frac{2\lambda-1}{2\lambda-1}\frac{1}{2}\big(f(2\lambda) + f(2(1-\lambda))\big) \\
&\quad \text{(equality for } \beta = \tfrac{1}{2}) \\
&= \frac{1}{2}\Big(f\big(1+[2\lambda-1]\big) + f\big(1-[2\lambda-1]\big)\Big) \\
&= \frac{1}{2}\Big(f\big(1+[A_{pr}-A_q]\big) + f\big(1-[A_{pr}-A_q]\big)\Big) \qquad (17)
\end{aligned}$$

In the course of the derivations (14) and (17), the conditions for equality/tightness are given, and show that the bound is reached in case of equality of the class conditionals and effective two class priors as described in Theorem 1. ■

IV. Exemplary f-Divergence Bounds

In this section, a number of explicit error bounds will be derived from Theorem 1 by substituting specific examples of *f-Divergence*s. Each choice of *f-Divergence* is defined by a convex function $f(u)$ with $f(1) = 0$. Details on the *f-Divergence*s used here can be found in [7].

A. Kullback-Leibler

The *Kullback-Leibler* distance is obtained from the *f-Divergence* by setting $f(u) = u\log u$. The associated bound following from Theorem 1 then becomes:

$$\begin{aligned}
D_f(q||pr) &= D_{\mathrm{KL}}(pr||q) \\
&\geq \frac{1}{2}\left[(1+\Delta)\log(1+\Delta) + (1-\Delta)\log(1-\Delta)\right]
\end{aligned}$$

A plot of the bound is shown in Fig. 1, together with simulations supporting the tightness of the bound.

B. Reversed Kullback-Leibler (Likelihood Disparity)

The reversed *Kullback-Leibler* distance, also termed *likelihood disparity* [7] is obtained by setting $f(u) = -\log u$. The associated bound then becomes:

$$D_f(q||pr) = D_{\mathrm{KL}}(q||pr) \geq \frac{1}{2}(\log(1+\Delta) + \log(1-\Delta))$$
$$\Leftrightarrow \quad \Delta \leq \sqrt{1 - \exp(-2D_{\mathrm{KL}}(q||pr))}$$

A plot of this bound is shown in Fig. 2, again together with simulations supporting the tightness of the bound. As pointed out in [6], this bound can also be derived from [3]. The corresponding non-trivial derivation will be presented in further work.

C. Chi-Squared

The χ^2-distance D_{χ^2} is obtained by setting $f(u) = u^2 - 1$. The associated bound then becomes:

$$\Delta^2 \leq D_f(q||pr) =: D_{\chi^2}(q||pr) = \int \sum_{c\in\mathcal{C}} \frac{pr^2(x,c)}{q(x,c)}\, dx - 1$$

D. Hellinger

The Hellinger distance is obtained using $f(u) = (\sqrt{u}-1)^2$ or $f(u) = 2(1-\sqrt{u})$. The associated bound becomes then:

$$D_f(q||pr) =: D_{\mathrm{H}}(q||pr) \geq 2 - \sqrt{1+\Delta} - \sqrt{1-\Delta}$$

E. Total Variational Distance

Assume the *f-Divergence* with $f(u) = |u-1|^\alpha$ and $\alpha \geq 1$. The associated bound then becomes:

$$D_f(q||pr) =: D_\alpha(q||pr) \geq \Delta^\alpha$$

Interestingly, the special case of $\alpha = 1$ shows that the accuracy

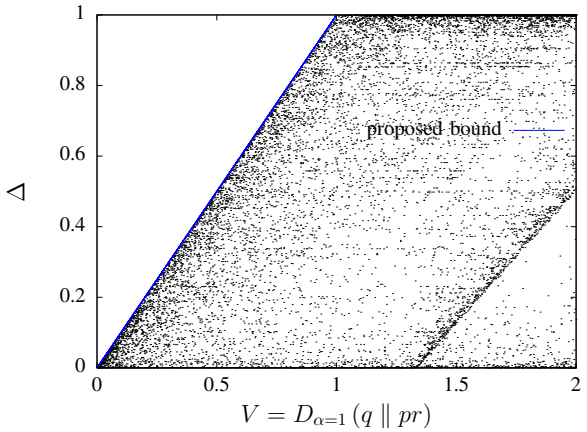

Fig. 3. *Total variational distance bound for $f(u) = |u-1|$. The black dots refer to simulations.*

mismatch is directly bounded by the total variational distance:

$$\Delta \leq V \tag{4}$$

Class prior simulations of the true and the model distribution are conducted for each bound induced by an *f-Divergence* using shared class conditional distributions. Exemplary, the relation of the bounds to the mismatch are shown in Figure 1 through 3 for a three class problem and two observations. The same tendency was confirmed in simulations using more classes and observations. In general, several million distributions along with the corresponding bounds value were generated, followed by a filtering to balance the plotting effort.

The bound derived from the reversed *Kullback-Leibler* distance is limited to the domain of the accuracy mismatch, and approaches the upper limit of the domain asymptotically, only. In contrast to this, On the other hand, the bounds derived from the other examples of *f-Divergence*s discussed here are not limited to the domain of the accuracy mismatch, which renders them useless for those cases where the trivial bound $\Delta \leq 1$ is more tight.

V. Conclusion

In this work, a class of tight bounds on the accuracy mismatch between the *Bayes*, and a mismatched decision rule is established. To the authors best knowledge, this presents the closest bounds on the accuracy mismatch known so far. A complete new category of bounds is introduced based on *f-Divergences*. Simulations were performed to support the analytical results.

Acknowledgments

This work was partly realized under the QUAERO Program, funded by OSEO, French State agency for innovation. H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

References

[1] Cover, T. M. and Thomas, J. A., *Elements of Information Theory,* 2nd edtion, John Wiley & Sons, New York, NY, 2006.

[2] Fedotov, A. A. and Harremoës P. and Topsøe F., "Refinements of Pinsker's inequality", IEEE Transactions on Information Theory, Vol. 49, No. 6, pp. 1491–1498, 2003.

[3] A. Guntuboyina, "Lower bounds for the minimax risk us- ing f-divergences and applications," *IEEE Transactions on Information Theory*, Vol. 57, pp. 2386–2399, 2011.

[4] Lah, P. and Ribarič, M., "Converse of Jensens inequality for convex functions", Publications de la faculte d'Electrotechnique de Universite a Belgrade, ser, Mathematics et Physique, No. 412–460, pp. 201–205, 1973.

[5] H. Ney, "On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition", Iberian Conference on Pattern Recognition and Image Analysis IbPRIA, Puerto de Andratx, Spain, pp. 636–645, June, 2003.

[6] M. Nussbaum-Thom, E. Beck, T. Alkhouli, R. Schlüter, H. Ney: "Relative Error Bounds for Statistical Classifiers based on the f-Divergence," in *Proc. Interspeech*, Lyon, France, August 2013.

[7] Österreicher, F., "Csizár's f-Divergences - Basic Properties", Talk presented at workshop of the Research Group in Mathematical Inequalities and Applications at the Victoria University, Melbourne, Australia, October, 2002.

[8] Schlüter, R. and Ney, H.: "Model-based MCE Bound to the True Bayes' Error," *IEEE Signal Processing Letters*, Vol. 8, No. 5, pp. 131–133, May 2001.

[9] Liese, F. and Vajda, I., "On Divergences and Informations in Statistics and Information Theory," *IEEE Transactions on Information Theory*, Vol. 52, No. 10, pp. 4394-4412, Oct. 2006.

[10] Vapnik V., "Statistical learning theory", Wiley, pp. 30–32, 1998.