Relative Error Bounds for Statistical Classifiers based on the $f$-Divergence

Markus Nussbaum-Thom1, Eugen Beck1, Tamer Alkhouli1, Ralf Schlüter1, Hermann Ney1,2

1Computer Science Dept. 6, RWTH Aachen University, Aachen, Germany
2Spoken Language Processing Group, LIMSI CNRS, Paris, France
{nussbaum, beck, alkhouli, schlueter, ney}@i6.informatik.rwth-aachen.de

Abstract
In language classification, measures like perplexity and Kullback-Leibler divergence are used to compare language models. While this bears the advantage of isolating the effect of the language model in speech and language processing problems, the measures have no clear relation to the corresponding classification error. In practice, an improvement in terms of perplexity does not necessarily correspond to an improvement in the error rate.

It is well-known that Bayes decision rule is optimal if the true distribution is used for classification. Since the true distribution is unknown in practice, a model distribution is used instead, introducing suboptimality. We focus on the degradation introduced by a model distribution, and provide an upper bound on the error difference between Bayes decision and a model-based decision rule in terms of the $f$-Divergence between the true and model distributions. Simulations are first presented to reveal a special case of the bound, followed by an analytic proof of the generalized bound and its tightness. In addition, the conditions that result in the boundary cases will be discussed. Several instances of the bound will be verified using simulations, and the bound will be used to study the effect of the language model on the classification error.

Index Terms: generalization bounds, language modeling, perplexity, confidence measures, $f$-Divergence, error mismatch.

1. Introduction
Perplexity was introduced in the early days of automatic speech recognition (ASR) [1, p.137] as a model complexity or branch- ing factor measure for the recognition task from an information-theoretic point of view, by considering the recognition problem as a source-channel decoding problem. Although the term model complexity gives a notion of the difficulty of the considered classification task, the actual relation to a relevant evaluation measure is left open. In language modeling across different techniques it is often observed that a decrease in perplexity does not necessarily yield an improvement in classification error [2, 3, 4]. In this work, the relation between information-theoretic measures among others and the classification error will be investigated in more detail. To this end, we introduce a tight class of bounds on the error mismatch between a Bayes (true) and a model classifier, benefitting from simulations used to reveal the existence of the reversed Kullback-Leibler (RKL) bound and verify its tightness. The RKL bound is generalized to a class of bounds based on the $f$-Divergence and an analytic proof of the generalized bound is presented. The $f$-Divergence bounds are further supported by simulations.

The next sections are organized as follows: Section 2 revisits related error bounds, and Section 3 introduces the RKL bound and its generalization, followed by the proof presented in Section 4. Several examples of the introduced bound are given in Section 5. Section 6 extends the bound to string classes, and finally Section 7 concludes the paper.

2. Existing Error Bounds
Assume a statistical classification problem, where a model distribution $q(x,c)$ of the continuous observations $x \in \mathcal{X}$ and classes $c \in \mathcal{C}$ is used to classify samples of the unknown true distribution $p(x,c)$. The Bayes decision $q_{pr}(x)$ and model $c_{q}(x)$ decision rules corresponding to the true and model distributions are defined as:

$c_{pr}(x) := \arg\max_{c \in \mathcal{C}} p(x,c)\} \text{ and } c_{q}(x) := \arg\max_{c \in \mathcal{C}} q(x,c)\}.

The quality of the model can be measured by the error mismatch associated with the decision rules:

$\Delta := \int |p_{pr}(x)| - p_{q}(x,c_{q}(x))|\,dx.

Another quality measure is the Kullback-Leibler (KL) divergence or relative entropy:

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

which is closely connected to the perplexity difference $PP(p||q) = \exp(D_{KL}(p||q))$.

In the last two decades, efforts were made to clarify the relation between the Kullback-Leibler divergence and the mismatch, either directly or as a by-product of other results. In [5], the Kullback-Leibler divergence was shown to be an upper bound on the mismatch, with the intention to derive empirical training criteria. In addition, the more general relation between the total variational distance $V$ and the mismatch was introduced:

$V := \int \sum_{c \in \mathcal{C}} |p_{pr}(x,c) - q(x,c)|\,dx \geq \Delta. (1)$

In machine learning [6, p.30], the Bretagnolle-Huber bound is known in the context of density estimation:

$\Delta^2 \leq \nu^2 \leq 4(1 - \exp(-D_{KL}(p||q)))

Aiming for a refinement of the Pinsker inequality, Vajda and Fedotov et al. [7] respectively introduced the following relations:

$\log \left( \frac{2 + V}{2 - V} \right) - \frac{2V}{2 - V} \leq D_{KL}(p||q),

\sum_{i=1}^{\infty} k_{i}^{2}V^{2i} \leq D_{KL}(p||q) (2)

where $k_{i}, i = 1, 2, \ldots$ are the constants of the Taylor expansion. Unfortunately, these two bounds have no explicit representation in $V$. 

3. Novel RKL and f-divergence Bounds

In this Section, we introduce a tight error bound on the mismatch based on the reversed KL divergence $D_{KL}(q||pr)$ ("reversed" since $pr$ and $q$ are switched). The bound has the form:

$$\Delta^2 \leq 1 - \exp(-2D_{KL}(q||pr)). \quad (3)$$

As Theorem 1 indicates, the bound is tight for any function $f$ fulfilling the properties of the $f$-divergence. Notice, that the special case of $f$-Divergence bound in Ineq. (3) can also be derived using another proof based on the results presented in [12]. This proof and a second kind of tight bound will be covered in another publication.

4. Proof of $f$-divergence Bound

In this section, a proof of the proposed bound is provided and the boundary conditions will be discussed. In the following, assume $c_1, c_2 \in C$ are two classes s.t. $c_1 \neq c_2$ and assume that $\forall x \in X : c_{pr}(x) \neq c_{q}(x)$ without loss of generality. The extension of the proof including the case of Bayes and model decision rules leading to equal results in subspaces of $X$ is straightforward, but needs to be presented in a further publication due to the lack of space. The proof uses the following properties of the $f$-Divergence:

**Permutation** Let the distributions $\hat{pr}(x, c) := pr(x, \pi_x(c))$ and $\hat{q}(x, c) := q(x, \pi_x(c))$ be permuted versions of the true and model distributions s.t. $\pi_x(c_{pr}(x)) \neq c_1, \pi_x(c_{q}(x)) \neq c_2, \pi_x(c_{pr}(x)) = c_2, \pi_x(c_{q}(x)) = c_1$, and $\pi_x(c_{pr}(x)) = \pi_x(c_{q}(x)) = c$ otherwise. The permutation neither changes the $f$-Divergence value nor the mismatch for the permuted distributions.

**Aggregation** This is also known as the lumping property [8, p.32] of the $f$-Divergence or the log-sum inequality. Consider the aggregation of two summands of the $f$-Divergence, with $p_1, p_2, q_1, q_2 \in R^+$, the following inequality applies:

$$q_1 f \left( \frac{p_1}{q_1} \right) + q_2 f \left( \frac{p_2}{q_2} \right) \geq (q_1 + q_2) \cdot f \left( \frac{p_1 + p_2}{q_1 + q_2} \right).$$

**Jensen** According to Jensen’s inequality [10] for a convex function $f$, the expectation $E$ fulfills $f(E(x)) \leq E(f(x)).$

**Proof of Theorem 1** Consider the $f$-Divergence of the original pair of distributions $q$ and $pr$:

$$D_f(pr||q) \geq D_f(\hat{pr}||\hat{q}) \quad \text{(Permutation)}$$

$$= \sum_c \hat{q}(c) \int \frac{\hat{pr}(x|c)\hat{pr}(c)}{\hat{q}(x|c)\hat{q}(c)} dx \quad \text{ (equality, iff } \hat{q}(x|c) = \hat{pr}(x|c))$$

$$\geq \sum_c \hat{q}(c) f \left( \frac{\hat{pr}(c)}{\hat{q}(c)} \right) \quad \text{ (Jensen; eq. iff } \frac{\hat{pr}(c)}{\hat{q}(c)} \text{ is convex})$$

$$= \sum_{c \in \{c_1, c_2\}} \hat{q}(c) f \left( \frac{\hat{pr}(c)}{\hat{q}(c)} \right) + \sum_{c \notin \{c_1, c_2\}} \hat{q}(c) f \left( \frac{\hat{pr}(c)}{\hat{q}(c)} \right)$$

$$\geq \sum_{c \in \{c_1, c_2\}} \hat{q}(c) f \left( \frac{\hat{pr}(c)}{\hat{q}(c)} \right) \quad \text{(Aggr.: eq. for effective 2-class subspaces)}$$

$$+ 2 \frac{1 - \hat{q}(c_1) - \hat{q}(c_2)}{2} \left( \frac{\hat{pr}(c_1) - \hat{pr}(c_2)}{\hat{pr}(c_1) - \hat{pr}(c_2)} \right) \quad \text{(Aggr.: eq. for effective 2-class subspaces)}$$

with the definitions:

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

for which we obtain:

$$2\lambda - 1 = \lambda - (1 - \lambda) = \hat{pr}(c_1) - \hat{pr}(c_2) = \Delta \quad (4)$$

$$2\beta - 1 = \beta - (1 - \beta) = \hat{q}(c_1) - \hat{q}(c_2).$$

Figure 1: The proposed RKL bound and simulations in comparison to Bretagnolle-Haber’s, Vajda’s and Fedotov’s bounds.
Using $\Delta \geq 0$ and $\eta(c_2) - \eta(c_1) \geq 0$, it follows that: $\beta \leq \frac{1}{2} \leq \lambda$. 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 that using the aggregation property yields the following inequality:

$$\frac{1}{2} (f(2\lambda) + f(2(1 - \lambda))) \geq f(1) = 0. \tag{5}$$

Then, the following simplification can be carried out:

$$D_f(pr||q) \geq \beta f \left( \frac{\lambda}{\beta} \right) + (1 - \beta) f \left( \frac{1 - \lambda}{1 - \beta} \right) = \beta f \left( \frac{\lambda}{\beta} \right) + a f \left( \frac{q}{a} \right) + (1 - \beta) f \left( \frac{1 - \lambda}{1 - \beta} \right) + b f \left( \frac{b}{b} \right) \geq (\beta + a) f \left( \frac{\lambda + a}{\beta + a} \right) + (1 - \beta + b) f \left( \frac{1 - \lambda + b}{1 - \beta + b} \right) \geq \frac{1}{2} \left( f(2\lambda) + f(2(1 - \lambda)) \right) \geq 0, \text{ cf. ineq. } (5)$$

$$= \frac{2\lambda - 2\beta}{2\lambda - 1} \cdot \frac{1}{2} \left( f(2\lambda) + f(2(1 - \lambda)) \right) \geq \frac{1}{2} \left( f(1 + \Delta) + f(1 - \Delta) \right) \quad \text{(cf. (4))}$$

### 5. Simulation of f-Divergence Bounds

In this Section, some f-Divergences and their corresponding bounds are presented. More possible f-Divergences can be found in [11]. In the following, let $u = pr(x, c)/q(x, c)$.

**Kullback-Leibler** The Kullback-Leibler divergence is obtained by setting $f(u) = -\log u$. The associated bound becomes then:

$$2D_{KL}(pr||q) \geq (1 + \Delta) \log(1 + \Delta) + (1 - \Delta) \log(1 - \Delta)$$

**Reversed Kullback-Leibler** The reversed Kullback-Leibler divergence is obtained by setting $f(u) = -\log u$. The associated bound becomes then:

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

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

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

**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:

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

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

$$\Delta^2 \leq D_f^2 (pr||q)$$
Unfortunately, the bounds derived from the Kullback-Leibler and the Hellinger distance do not result in a closed-form expression in terms of $\Delta$, in which case numerical approximations can be used. The bound derived from the RKL divergence takes values in $[0,1]$ and might give a reasonable upper bound on the mismatch. On the other hand, the bounds derived by the Chi-Squared and Vajda divergences are not limited to the mismatch domain, which makes them useless for those cases where the trivial bound $\Delta \leq 1$ is tighter.

The simulations in Figure 2 through 5 were performed by generating a shared class-conditional and two (true and model) class prior distributions, where 3 classes and 2 observations were assumed. The same tendency was confirmed in simulations using more classes and observations. In general, several million distributions were generated, followed by filtering to achieve better visualization.

6. Language Modeling Example

The pursuit to bound the mismatch in this work can be applied to quantify the language model effect in the string case, where sequences of discrete random variables $x_i^N$ and $c_i^N$ of length $N$ are used as observations and classes, respectively. A possible real-world example on such a string case decision problem is Part-of-speech (POS) tagging, where $x_i^N$ is the word sequence, and $c_i^N$ is the POS tag sequence. Assume the true and model distributions to have respectively higher- and lower-order $n$-gram language models. The lower-order language model is considered a model of the true higher-order language model, and linking the two is done by deriving the former from the latter. For instance, if we seek to relate two recognition systems having bigram and unigram language models, we define the position-dependent derived unigram as the marginalization of the language model distribution modeled according to the bigram assumption as follows:

$$q_n(c) = \sum_{c_i^N} \prod_{i=1}^{N} pr(c_i | c_{i-1})$$

Studying the mismatch in this case would yield insights as to the extent of improvement possible when replacing a unigram with its corresponding bigram language model. Assuming the two systems have a shared class-conditional model, so as to isolate the effect of the language model, the errors corresponding to the bigram and unigram systems are given respectively as:

$$E = 1 - \sum_{x_i^N} \max_{c_i^N} \left\{ \prod_{i=1}^{N} pr(c_i | c_{i-1}) pr_i(x_i | c_i) \right\}$$

$$\bar{E} = 1 - \sum_{x_i^N} \max_{c_i^N} \{ q_i(c_i) pr_i(x_i | c_i) \}$$

and the mismatch in this case would be $\Delta = E - \bar{E}$. Figure 6 compares simulation results against the RKL bound for strings, where the RKL is given by:

$$D_{KL}(q(c_i^N)||pr(c_i^N)) = \sum_{c_i^N} \prod_{i=1}^{N} q_i(c_i) \log \left( \prod_{n=1}^{N} \frac{q_n(c_n)}{q(c_n | c_{n-1})} \right)$$

The simulations confirm that the bound holds, while the gap between simulation points and the bound arises due to the introduced dependence between the two language models, which is an additional constraint not used in the previous sections.

7. Conclusion

A novel category of bounds based on $f$-Divergences was introduced, seeking to relate the error mismatch between Bayes and model-based statistical recognition systems. An analytic proof of the bounds was presented and supported by simulations indicating their validity and tightness. Furthermore, the RKL bound was used to study the effect of the language model on the classification error in string systems.

8. Acknowledgements

This work was partly realized under the QUAOER Programme, funded by OSEO, French State agency for innovation and EU-BRIDGE European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n 287658. H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.
9. References


