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

Markus Nussbaum-Thom[1], *Eugen Beck*[1], *Tamer Alkhouli*[1], *Ralf Schlüter*[1], *Hermann Ney*[1,2]

[1]Computer Science Dept. 6, RWTH Aachen University, Aachen, Germany
[2] Spoken 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 branching 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, benefiting 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 $pr(x, c)$. The *Bayes* $c_{pr}(x)$ and model $c_q(x)$ decision rules corresponding to the true and model distributions are defined as:

$$c_{pr}(x) := \underset{c\in\mathcal{C}}{\operatorname{argmax}}\{pr(x,c)\}, \quad c_q(x) := \underset{c\in\mathcal{C}}{\operatorname{argmax}}\{q(x,c)\}.$$

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

$$\Delta := \int pr(x, c_{pr}(x)) - pr(x, c_q(x)))\,\mathrm{d}x.$$

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

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

which is closely connected to the perplexity difference

$$PP(pr||q) = \exp(D_{\mathrm{KL}}(pr||q)).$$

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:

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

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

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

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

$$\log\left(\frac{2+V}{2-V}\right) - \frac{2V}{2-V} \leq D_{\mathrm{KL}}(pr||q),$$
$$\sum_{i=1}^{\infty} k_i^{2i} V^{2i} \leq D_{\mathrm{KL}}(pr||q) \qquad (2)$$

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 $D_{\mathrm{KL}}(q|pr)$ ("reversed" since pr and q are switched). The bound has the form:

$$\Delta^2 \le 1-\exp(-2D_{\mathrm{KL}}(q||pr)). \tag{3}$$

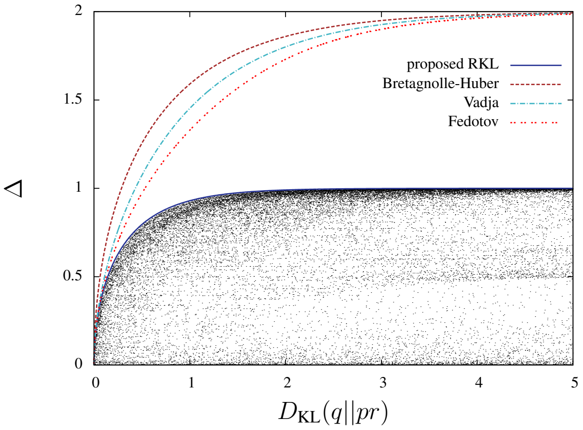

Figure 1: *The proposed RKL bound and simulations in comparison to Bretagnolle-Huber's, Vajda's and Fedotov's bounds.*

Figure 1 shows plots of the proposed RKL bound compared to existing bounds discussed in Section 2. These bounds are parametrized through $D_{\mathrm{KL}}(pr||q)$ due to the symmetry of the variational distance. The simulations correspond to simulating the distributions $pr(x,c)$ and $q(x,c)$ using 3 classes and 2 observations. The simulations clearly suggest that the proposed RKL bound is tight, in contrast to previous bounds. Additional simulations will be shown in Section 5. The conditions for the tightness of the bound are derived in the proof provided in Section 4. The RKL bound can be generalized using the *f-Divergence*.

Definition 1 *If $f:\mathbb{R}^+\to\mathbb{R}$ is a convex function and $f(1)=0$ then*

$$D_f(pr||q) := \int \sum_{c\in\mathcal{C}} q(x,c) f\left(\frac{pr(x,c)}{q(x,c)}\right) \mathrm{d}x$$

is defined as the f-Divergence[8, 9, 11].

Theorem 1 *The f-Divergence is lower-bounded by a function of the mismatch, which implicitly represents an upper bound to the mismatch as a function of the f-Divergence :*

$$D_f(pr||q) \ge \frac{1}{2}(f(1+\Delta)+f(1-\Delta)).$$

Equality is obtained with shared class-conditional probabilities,

$$\forall x\in\mathcal{X}, c\in\mathcal{C}: q(x|c)=pr(x|c)$$

and a special choice of priors for any pair of classes $c_1,c_2\in\mathcal{C}$ with $c_1\ne c_2$, and $\lambda\in[0.5;1.0]$, such that:

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

and

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

or, trivially, with the model distribution set equal to the true distribution.

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 $c_1,c_2\in\mathcal{C}$ are two classes s.t. $c_1\ne c_2$ and assume that $\forall x\in\mathcal{X}: c_{pr}(x)\ne c_q(x)$ 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 $\mathcal{X}$ 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 $\overline{pr}(x,c) := pr(x,\pi_x(c))$ and $\overline{q}(x,c) := \overline{q}(x,\pi_x(c))$ be permuted versions of the true and model distributions s.t. $\pi_x(c_{pr}(x)) = c_1$, $\pi_x(c_1)=c_{pr}(x)$, $\pi_x(c_q(x))=c_2$, $\pi_x(c_2)=c_q(x)$, and $\pi_x(c)=c$ 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 $p_1,p_2,q_1,q_2\in\mathbf{R}^+$, the following inequality applies:

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

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

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

$$D_f(pr||q) \ge D_f(\overline{pr}||\overline{q}) \qquad \text{(Permutation)}$$

$$= \sum_c \overline{q}(c) \int \overline{q}(x|c) f\left(\frac{\overline{pr}(x|c)\overline{pr}(c)}{\overline{q}(x|c)\overline{q}(c)}\right) \mathrm{d}x$$

(equality, iff $\overline{q}(x|c) = \overline{pr}(x|c)$)

$$\ge \sum_c \overline{q}(c) f\left(\underbrace{\int \overline{q}(x|c)\frac{\overline{pr}(x|c)}{\overline{q}(x|c)}}_{=1} \frac{\overline{pr}(c)}{\overline{q}(c)}\right) \mathrm{d}x$$

$$\left(\begin{array}{c}\text{Jensen; eq. iff} \\ q(x|c)=pr(x|c)\end{array}\right)$$

$$= \sum_{c\in\{c_1,c_2\}} \overline{q}(c) f\left(\frac{\overline{pr}(c)}{\overline{q}(c)}\right) + \sum_{c\in\mathcal{C}\setminus\{c_1,c_2\}} \overline{q}(c) f\left(\frac{\overline{pr}(c)}{\overline{q}(c)}\right)$$

$$\ge \sum_{c\in\{c_1,c_2\}} \overline{q}(c) f\left(\frac{\overline{pr}(c)}{\overline{q}(c)}\right) \qquad \left(\begin{array}{c}\text{Aggr.; eq. for effective} \\ \text{2-class subspaces}\end{array}\right)$$

$$+ 2\frac{1-\overline{q}(c_1)-\overline{q}(c_2)}{2} f\left(\frac{\frac{1-\overline{pr}(c_1)-\overline{pr}(c_2)}{2}}{\frac{1-\overline{q}(c_1)-\overline{q}(c_2)}{2}}\right)$$

$$\ge \beta f\left(\frac{\lambda}{\beta}\right) + (1-\beta) f\left(\frac{1-\lambda}{1-\beta}\right) \qquad \left(\begin{array}{c}\text{Aggr.; eq. for effective} \\ \text{2-class subspaces}\end{array}\right)$$

with the definitions:

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

for which we obtain:

$$\begin{aligned} 2\lambda - 1 &= \lambda - (1-\lambda) = \overline{pr}(c_1) - \overline{pr}(c_2) = \Delta \qquad (4) \\ 2\beta - 1 &= \beta - (1-\beta) = \overline{q}(c_1) - \overline{q}(c_2). \end{aligned}$$

Using $\Delta \geq 0$ and $\overline{q}(c_2) - \overline{q}(c_1) \geq 0$, it follows that: $\beta \leq \frac{1}{2} \leq \lambda$. Now assume the following definitions:

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

Also, note that using the aggregation property yields the following inequality:

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

Then, the following simplification can be carried out:

$$\begin{aligned}
D_f(pr||q) &\geq \beta f\left(\frac{\lambda}{\beta}\right) + (1-\beta) f\left(\frac{1-\lambda}{1-\beta}\right)\\
&= \beta f\left(\frac{\lambda}{\beta}\right) + a\,\underbrace{f(\frac{a}{a})}_{=0} + (1-\beta) f\left(\frac{1-\lambda}{1-\beta}\right) + b\,\underbrace{f(\frac{b}{b})}_{=0}\\
&\geq (\beta + a) f\left(\frac{\lambda + a}{\beta + a}\right) + (1-\beta+b) f\left(\frac{1-\lambda+b}{1-\beta+b}\right)\\
&\hfill \text{(Aggregation)}\\
&= \frac{2\lambda - \overbrace{2\beta}^{\leq 1}}{2\lambda - 1} \underbrace{\frac{1}{2}\Big(f(2\lambda) + f(2(1-\lambda))\Big)}_{\geq 0, \text{cf. ineq. (5)}}\\
&\geq \frac{1}{2}\Big(f(2\lambda) + f(2(1-\lambda))\Big)\\
&= \frac{1}{2}\Big(f\big(1+\Delta\big) + f\big(1-\Delta\big)\Big) \qquad \text{(cf. (4))}
\end{aligned}$$

5. Simulation of f-Divergence Bounds

In this Section, some *f-Divergence*s and their corresponding bounds are presented. More possible *f-Divergence*s 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 $f(u) = u \log u$. The associated bound becomes then:

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

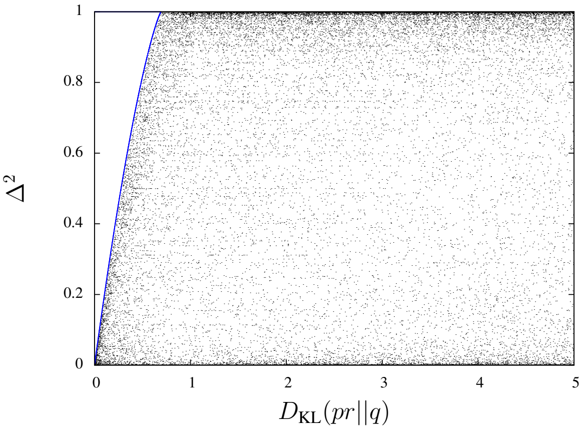

Figure 2: *Kullback-Leibler Bound.*

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

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

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

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

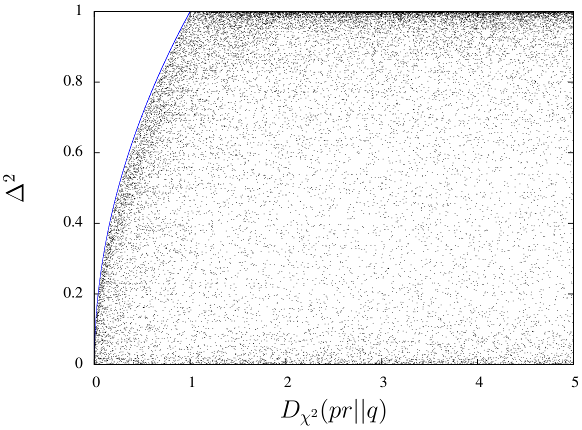

Figure 3: *Chi-Squared Bound.*

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:

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

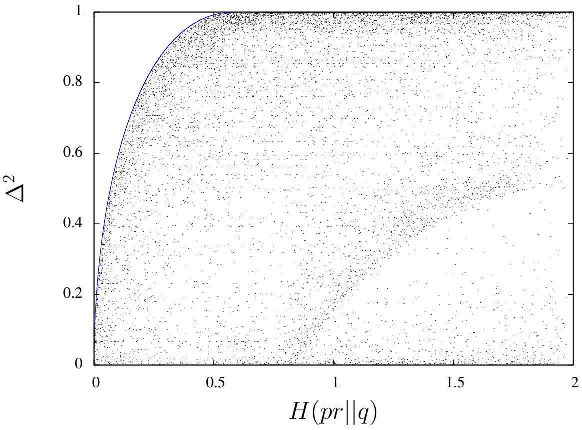

Figure 4: *Hellinger Bound.*

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

$$\Delta^2 \leq D_f^{\frac{2}{\alpha}}(pr||q)$$

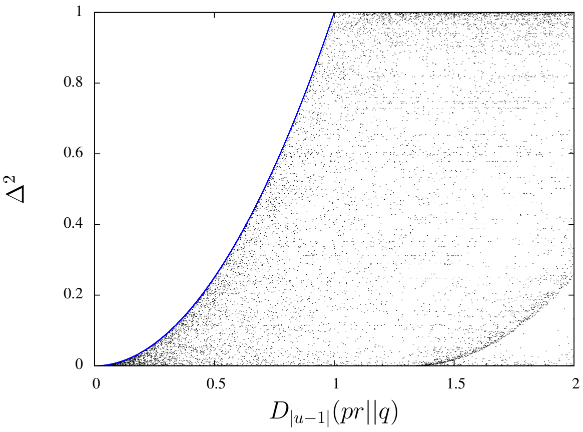

Figure 5: *Vajda Bound for* $f(u) = |u-1|$.

Unfortunately, the bounds derived from the *Kullback-Leibler* and the Hellinger distance do not result in a closed-form expression in terms of Δ, 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_1^N and c_1^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_1^N is the word sequence, and c_1^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:

$$q_n(c) = \sum_{c_1^N : c_n = c} \prod_{i=1}^{N} pr(c_i|c_{i-1})$$

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:

$$E = 1 - \sum_{x_1^N} \max_{c_1^N} \left\{ \prod_{i=1}^{N} pr(c_i|c_{i-1}) pr_i(x_i|c_i) \right\}$$

$$\bar{E} = 1 - \sum_{x_1^N} \prod_{i=1}^{N} \max_{c_i} \left\{ q_i(c_i) pr_i(x_i|c_i) \right\}$$

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:

$$D_{\text{KL}}(q(c_1^N)||pr(c_1^N)) = \sum_{c_1^N} \prod_{i=1}^{N} q_i(c_i) \log \left(\prod_{n=1}^{N} \frac{q_n(c_n)}{q(c_n|c_{n-1})} \right)$$

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.

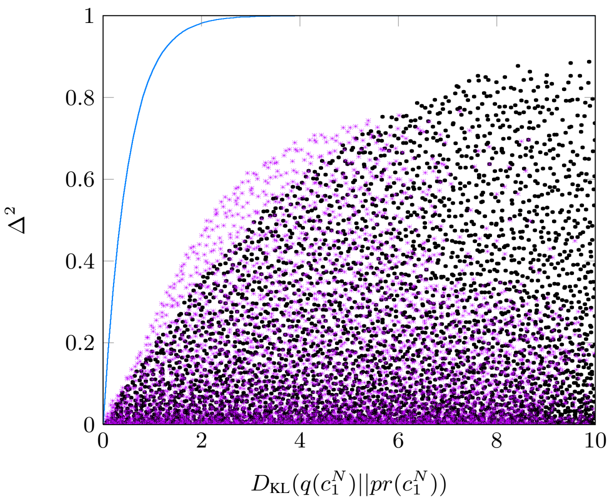

Figure 6: Bigram vs. derived unigram simulations for strings of length 4 (star-shapes) and 6 (dots) for 2 classes and 2 observations, along with the RKL bound shown as the blue curve.

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 QUAERO 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

[1] Jelinek, F., “Statistical Methods for Speech Recognition”, MIT Press, Cambridge Massachusetts, 137, 1997.

[2] Rosenfeld, R., “Two decades of statistical language modeling: Where do we go from here ?”, Proceedings of the IEEE, 1270-1278, 2000.

[3] Klakow, D. and Peters, J., “Testing the correlation of word error rate and perplexity”, Elsevier, Speech Communication, 38(1): 19–28, September, 2002.

[4] Schwenk, H., “Continuous space language models”, Computer Speech & Language, 21(3), 492-518, 2007.

[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, 636-645, June, 2003.

[6] Vapnik V., “Statistical learning theory”, Wiley, pp. 30-32, 1998.

[7] Fedotov, A. A. and Harremoës P. and Topsøe F., “Refinements of Pinsker’s inequality”, IEEE Transactions on Information Theory, 49(6), 1491-1498, 2003.

[8] Csiszár, I. and Shields, P. C., “Information Theory and Statistics: A Tutorial”, Foundations and Trends in Communications and Information Theory, 1(4), 2004.

[9] Liese F. and Vajda I., ”On Divergences and Informations in Statistics and Information Theory”, IEEE Transactions on Information Theory, 52(10), 2006, 4394-4412.

[10] 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 - No. 460, 201 - 205, 1973.

[11] Öesterreicher, 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.

[12] Guntuboyina, A., “Lower bounds for the minimax risk using f-divergences and applications”, IEEE Transactions on Information Theory, 2011.