A FAMILY OF DISCRIMINATIVE TRAINING CRITERIA BASED ON THE F-DIVERGENCE FOR DEEP NEURAL NETWORKS

Markus Nussbaum-Thom[1,2]*, Xiaodong Cui*[1]*, Ralf Schlüter*[2]*, Vaibhava Goel*[1]*, Hermann Ney*[2,3]

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY, 10598
[2] Computer Science Dept. 6, RWTH Aachen University, Aachen, Germany
[3] Spoken Language Processing Group, LIMSI CNRS, Paris, France
{nussbaum, schlueter, ney}@i6.informatik.rwth-aachen.de, {xcui, vgoel}@us.ibm.com

ABSTRACT

We present novel bounds on the classification error which are based on the *f-Divergence* and, at the same time, can be used as practical training criteria. There exist virtually no studies which investigate the link between the *f-Divergence*, the classification error and practical training criteria. So far only the *Kullback-Leibler f-Divergence* has been examined in this context to formulate a bound on the classification error and to derive the cross-entropy criterion. We extend this concept to a larger class of *f-Divergences*. We also successfully investigate if the novel training criteria based on the *f-Divergence* are suited for frame-wise training of deep neural networks on the Babel Vietnamese and Bengali speech recognition tasks.

Index Terms— discriminative training, classification error bound, f-Divergence, deep neural network

1. INTRODUCTION

The *f-Divergence* is a well known quantity [1, 2, 3] in mathematics for which the relation to the variational distance has often been studied in [4, 5] or [6, p.30]. Apart from our work reported in [7, 8, 9] there exist virtually no studies on the relation between the *f-Divergence*, the classification error or practical training criteria. Only in [8] the *Kullback-Leibler f-Divergence* has been shown to be an upper bound on the classification error difference between the *Bayes* and model decision rule. At the same time in [8], the cross-entropy criterion has been be derived from the *Kullback-Leibler* divergence in a theoretical sound way. We extend this concept to a larger class of *f-Divergences*: Establish bounds on the classification error difference and derive practical training criteria from these bounds. This continues our latest study from [9] on bounds on the classification error difference based on the *f-Divergence* to the derivation of practical training criteria based on the *f-Divergence*.

We also successfully examine if these novel training criteria are suited for frame-wise training of deep neural networks (DNNs) the Babel Vietnamese and Bengali automatic speech recognition (ASR) tasks in comparison to the cross-entropy criterion [10, 11, 12] which is the state-of-the-art criterion for frame-wise training of DNNs. On sequence-level, the minimum phone error, state-level minimum *Bayes* risk (sMBR) and boosted maximum mutual information have been shown to successfully improve the neural network [11, 13, 14].

The next sections are organized as follows: Section 2 relates to prior work, and gives a brief outlook on the results of this paper. In Section 3 a proof of global error bounds is presented from which in Section 4 empirical training criteria are derived. Section 5 shows experimental results for frame-wise DNN training using the novel criteria and Section 6 concludes the paper.

2. PRIOR WORK AND OUTLOOK

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 posteriors $pr(c|x)$ and $q(c|x)$ are defined as:

$$c_{pr}(x) := \operatorname*{argmax}_{c\in\mathcal{C}}\{pr(c|x)\} \quad \text{and} \quad c_q(x) := \operatorname*{argmax}_{c\in\mathcal{C}}\{q(c|x)\}$$

The quality of the model can be measured by the local and global classification error difference associated with the decision rules:

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

In [8] the standard cross entropy training criterion has been derived from the *Kullback-Leibler f-Divergence* using the following concept:

- Establish a local posterior bound,
- extend this bound to global level,
- and derive practical training criteria.

There, the *Kullback-Leibler f-Divergence* has also been shown to limit the squared classification error difference and has been used to derived the cross-entropy criterion for the labeled samples $(x_n, c_n)_{n=1}^N$:

$$2\int pr(x)\sum_{c\in\mathcal{C}} pr(c|x)\log\left(\frac{pr(c|x)}{q(c|x)}\right)\mathrm{d}x \Rightarrow \frac{1}{N}\sum_{n=1}^{N}\log\frac{1}{q(c_n|x_n)}$$

We generalize this concept to a larger class of *f-Divergences*.

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

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

is defined as the f-Divergence [1, 2, 3].

We derive local posterior bounds by reformulating our recent study on global classification error bounds based on the *f-Divergence* [9] to posteriors and local error bounds:

$$2D_f^x(pr||q) \geq f(1 + \Delta(x)) + f(1 - \Delta(x)). \tag{1}$$

Eq. 1 can be proven by replicating the steps in [9] for posterior probabilities and the local classification error difference. We derive for f where f''' is monotonically increasing a bound on the squared global classification error difference. This bound can be resolved, for conjugate convex functions $f(u) = ug(1/u)$ and g monotonically decreasing, to practical training criteria:

$$2\int pr(x)D_f^x(pr||q)\,\mathrm{d}x \Rightarrow F_f(q) = \frac{1}{N}\sum_{n=1}^{N} g(q(c_n|x_n))$$

3. GLOBAL CLASSIFICATION ERROR BOUNDS

In this Section, the local classification error bound based on the *f-Divergence* from Eq. 1 is extended to global level, followed by a discussion on the tightness property. The global classification error bound is proven using the *Taylor's* theorem [15].

Taylor's theorem Let $k \in \mathbb{N}$ and the function $f : \mathbb{R} \to \mathbb{R}$ be k times differentiable at point $y_0 \in \mathbb{R}$. Then there exists a constant $\mu_y \in [y_0, y]$ with $R_k(y) = \frac{f^{(k+1)}(\mu_y)}{k!}(y-y_0)^{k+1}$ such that $f(y) = \sum_{n=0}^{k-1} \frac{f^{(n)}(a)(y-y_0)^n}{n!} + R_k(y)$.

Theorem 1 *If $f : \mathbb{R}^+ \to \mathbb{R}$ is convex, $f(1) = 0$, f''' exists and is monotonically increasing, then the following bound on the global classification error difference is valid:*

$$f''(1)\Delta^2 \leq \int pr(x) D_f^x(pr||q)\ \mathrm{d}x$$

Proof: By integration the local classification error bound from Eq. 1

$$2D_f^x(pr||q) \geq f(1-\Delta(x)) + f(1+\Delta(x)),$$

is extendable using *Jensens* inequality [16]:

$$\begin{aligned}
&2\int pr(x) D_f^x(pr||q)\ \mathrm{d}x \\
&\geq \int pr(x)\left(f(1-\Delta(x)) + f(1+\Delta(x))\right)\ \mathrm{d}x \\
&\boxed{\text{Jensens inequality: } E(f(X)) \geq f(E(X)) \text{ for convex } f} \\
&\geq f\left(1 - \int pr(x)\Delta(x)\ \mathrm{d}x\right) + f\left(1 + \int pr(x)\Delta(x)\ \mathrm{d}x\right) \\
&\geq f(1-\Delta) + f(1+\Delta)
\end{aligned}$$

Locally in $y_0 = 1$, the application of *Taylor's* Theorem results in:

$$\begin{aligned}
&f(1+\Delta) + f(1-\Delta) \\
=&f(1) + f'(1)\Delta + \frac{f''(1)\Delta^2}{2!} + \frac{f'''(\mu_{1+\Delta})}{3!}\Delta^3 \\
&+ f(1) - f'(1)\Delta + \frac{f''(1)\Delta^2}{2!} - \frac{f'''(\mu_{1-\Delta})}{3!}\Delta^3 \\
=&\underbrace{2f(1)}_{=0} + 2f''(1)\Delta^2 + \frac{\Delta^3}{3!}(f'''(\mu_{1+\Delta}) - f'''(\mu_{1-\Delta})) \\
\geq&2f''(1)\Delta^2 + \frac{\Delta^3}{3!}\Big(\underbrace{\min_{a\in[1,1+\Delta]} f'''(a)}_{\geq f'''(1)} + \underbrace{\max_{b\in[1-\Delta,1]} -f'''(b)}_{\geq -f'''(1)}\Big) \\
\geq&2f''(1)\Delta^2 + \frac{\Delta^3}{3!}(f'''(1) - f'''(1)) \\
=&2\underbrace{f''(1)}_{\geq 0}\Delta^2
\end{aligned}$$

This leads to the bound on the squared global classification error difference:

$$\Delta^2 \leq \frac{1}{f''(1)} \int pr(x) D_f^x(pr||q)\ \mathrm{d}x \qquad \blacksquare$$

Tightness: The derived bound is tightly bounded with the squared global error difference iff the model $q(c|x)$ approaches $pr(c|x)$, which is also fulfilled for the *Kullback-Leibler f-Divergence* from which the cross-entropy has been derived in [8]. Due to $D_f^x(pr||pr) = 0$ this property is intuitive.

The next section derives empirical training criteria based on *f-Divergences* from this bound.

4. EMPIRICAL TRAINING CRITERIA

In this Section, empirical training criteria are derived from the global classification error bounds developed in Section 3. The derivation is restricted to a special category of *f-Divergence*. First, the global classification error bound is extended. Then, practical training criteria are derived from this extension using the empirical distribution.

4.1. From Error Bounds to Empirical Training Criteria

Theorem 2 *Consider model $q(c|x)$ and an f-Divergence with a function $f(u) = ug(1/u)$ such that:*

- *g is monotonically decreasing, convex and , $g(1) = 0$,*
- *f''' is monotonically increasing (in compliance with Theorem 1).*

Then, the global classification error bound from Theorem 1 based on the f-Divergence can be extended to:

$$\int pr(x) D_f^x(pr||q)\ \mathrm{d}x \leq \int \sum_{c\in\mathcal{C}} pr(x,c) g(q(c|x))\ \mathrm{d}x$$

Proof: The global classification error bound evaluates to:

$$\begin{aligned}
&\int pr(x) D_f^x(pr||q)\ \mathrm{d}x && \left(f(u) = ug\left(\tfrac{1}{u}\right)\right) \\
&= \int pr(x) \sum_{c\in\mathcal{C}} q(c|x) \frac{pr(c|x)}{q(c|x)} g\left(\frac{q(c|x)}{pr(c|x)}\right)\ \mathrm{d}x \\
&= \int \sum_{c\in\mathcal{C}} pr(x,c) g\left(\frac{q(c|x)}{pr(c|x)}\right)\ \mathrm{d}x \\
&\leq \int \sum_{c\in\mathcal{C}} pr(x,c) g\left(\frac{q(c|x)}{1}\right)\ \mathrm{d}x && (g \text{ monotonically decreasing}) \\
&= \int \sum_{c\in\mathcal{C}} pr(x,c) g(q(c|x))\ \mathrm{d}x && \blacksquare
\end{aligned}$$

According to the scheme of [8, p.642] practical training criteria are derived from this extended bound using the empirical distribution of the training data. The empirical distribution of the labeled samples $(x_n, c_n)_{n=1}^N$ is defined by:

$$pr(x,c) = \frac{1}{N}\sum_{n=1}^{N} \delta(x-x_n)\delta(c,c_n)$$

Here, $\delta(x)$ denotes the continuous *Dirac* delta and $\delta(c,\tilde{c})$ (1 iff $c = \tilde{c}$, 0 otherwise) is the discrete *Kronecker* delta. The *Dirac* delta has the useful sifting property [17] for integrable functions $\int h(x)\delta(x-x_0)\ \mathrm{d}x = h(x_0)$ which is used in the next proof. The training criterion to determine the optimum parameter set of the model can be written as:

$$\begin{aligned}
F_f(q) &= \int \sum_{c\in\mathcal{C}} pr(x,c) g(q(c|x))\ \mathrm{d}x && (\text{emp. } pr(x,c)) \\
&= \int \sum_{c\in\mathcal{C}} \frac{1}{N}\sum_{n=1}^{N} \delta(x-x_n)\delta(c,c_n) g(q(c|x))\ \mathrm{d}x \\
&= \frac{1}{N}\sum_{n=1}^{N} g(q(c_n|x_n)) && (\text{sifting property})
\end{aligned}$$

We would like to emphasize that these training criteria are consistent with the true distribution in case of infinite training data in that sense

that the optimum model results in a monotonously increasing transformation of the true posterior which retains the *Bayes* decision rule. A complete proof of this property is beyond the scope of this paper and will be discussed in a further publication. In the next Section examples for practical training criteria are shown.

4.2. Examples

We introduce examples for the training criteria from the family of criteria of Section 4.1. Table 1 shows the function g from $f(u) = ug(1/u)$ corresponding to the *Kullback-Leibler* (CE), *Lin* and conjugate power approximation (α-CPA, $\alpha \in [0, 1]$) *f-Divergences* with additional characteristic properties. Also the third derivative is presented which easily can be shown to be monotonically increasing in the positive domain as demanded by Theorem 1.

Table 1. *Values of $g(1/u)$, $g'(q(c_n|x_n))$, $f''(1)$ and $f'''(u)$ for the CE, LIN and α-CPA criteria with $u = pr(c|x)/q(c|x)$.*

	CE	LIN	α-CPA
$g\left(\frac{1}{u}\right)$	$-\log\left(\frac{1}{u}\right)$	$-\log\left(\frac{1}{2}\left(1+\frac{1}{u}\right)\right)$	$-\frac{\left(\frac{1}{u}\right)^{\alpha}-1}{\alpha}$
$g'(q(c\|x))$	$-\frac{1}{q(c\|x)}$	$-\frac{1}{1+q(c\|x)}$	$-\frac{1}{(q(c\|x))^{1-\alpha}}$
$f''(1)$	1	$1/4$	$(1-\alpha)$
$f'''(u)$	$-\frac{1}{u^3}$	$-\frac{1+3u}{u^2(1-u)^3}$	$-\frac{1-\alpha^2}{u^{\alpha+2}}$

The training criteria associated with these divergences become:

$$F_f(q) = -\frac{1}{N}\sum_{n=1}^{N}\log q(c_n|x_n) \qquad \text{(CE)}$$

$$F_f(q) = -\frac{1}{N}\sum_{n=1}^{N}\log\left(\frac{1}{2}(1+q(c_n|x_n))\right) \qquad \text{(LIN)}$$

$$F_f(q) = \frac{1}{N}\sum_{n=1}^{N}\frac{\left(1-(q(c_n|x_n))^{\alpha}\right)}{\alpha} \qquad (\alpha\text{-CPA})$$

Due to simplicity, the previous criteria are referred to as the CE, LIN and α-CPA or in general *f-Divergence* criteria.

Optimization: The neural networks using the novel training criteria are optimized using the SGD framework. Within this framework, only the gradient of the neural network with respect to the input layer of the softmax-layer has to be modified. The gradient optimization of other hidden layers remains unchanged as for the CE criterion. For the proposed criteria with $f(u) = ug(1/u)$ the gradient has a canonical form:

$$\nabla F_f(q) = \sum_{n=1}^{N} g'(q(c_n|x_n))\nabla q(c_n|x_n)$$

Table 1 shows values of $g'(q(c_n|x_n))$ and $f''(1)$ for the CE, *Lin*, α-CPA and 0.5-CPA criterion.

We would like to emphasize that the factor $2/f''(1)$ from Theorem 1 theoretically describes the correct weighting between different *f-Divergence* criteria because only using these factors the corresponding bounds limit the classification error difference without bias.

Linear Combination: Another property is that the derived *f-Divergence* bounds and criteria are closed under linear combination, since a linear combination of *f-Divergences* is still a *f-Divergence*. Therefore, a linear combination of *f-Divergence* bounds and criteria remains within the same family. Furthermore, a linear combination of criteria $F_{f_1}(q), \ldots, F_{f_I}(q)$ with weights $\lambda_1, \ldots, \lambda_I \in \mathbb{R}$ comes with no additional computational cost since the optimization of the hidden layers except the last layer remains unchanged. Additional costs only occur for the frame-wise weight $\sum_{i=1}^{I}\lambda_i g_i'(q(c_n|x_n))$ on top of the standard CE derivative $\nabla q(c_n|x_n)$ which can be integrated efficiently using a matrix multiplication on GPUs. The question has to be asked: Why should a combination of criteria result in an improved model? As outlined previous, the novel *f-Divergence* criteria discussed in Section 4 are consistent with a monotonous increasing transformation of the true posterior. In theory during optimization all *f-Divergence* criteria as well as their combination should head into a direction towards a posterior which fulfills the *Bayes* decision rule. Therefore, the optimum model learned with a combination of *f-Divergence* criteria should perform at least as good as a single criterion. In the next Section, we try to confirm this assumption experimentally.

5. APPLICATION TO ASR

In this section, the theoretically derived training criteria are tested on the Babel ASR limited language pack tasks.

5.1. Experimental Setup

The novel training criteria are tested on Babel Vietnamese and Bengali ASR tasks. Both training and development sets are composed similar to the Babel tasks described in [18]. The Babel training and development sets for each language consist of about 20 h conversational and scripted telephony speech each. In addition for Vietnamese, 14 h of the Babel 2013 evaluation set are available, the remaining data is still under non-disclosure. Overall, the data poses a good challenge to acoustic modeling in terms of spontaneous speaking style, dialect, channel, and speaker diversity.

The acoustic front-end for Vietnamese and Bengali is comprised of Perceptive Linear Predictive (PLP) features. Nine consecutive frames are concatenated and a Linear Discriminant Analysis (LDA) is used to reduce the dimension to 40. The Vietnamese system is speaker independent while the Bengali uses Vocal tract length normalization and speaker adaptive training based on feature-space maximum likelihood linear regression (FMLLR) to reduce speaker variability.

The baseline acoustic model is a DNN trained using the CE criterion. The DNN consists of 5 hidden layers with 1024 nodes and 1000 output nodes and has an input dimension of 360. A layer-wise pre-training is used to initialize the DNN. The alignments used for the training of the neural network were obtained by alternating optimization of DNN training and realignment. The deep neural network is optimized using the mini-batch based stochastic gradient descent (SGD) algorithm [10, 11] which can efficiently be performed using GPUs. On top of the best frame-wise trained DNN models, a Hessian-free (HF) sMBR optimization [11] is applied to improve the final performance of the DNNs. Due to missing language model data, from the acoustic training data a modified *Kneser-Ney* [19] trigram for Vietnamese and bigram model for Bengali was estimated for recognition. All improvements are measured in token error rate (TER) which is the WER for language dependent tokens instead of words.

5.2. Experimental Results

In initial experiments the DNNs were trained using single criteria solely. It turned out, only the criteria resulting from the *Lin* and α-CPA divergence (for small values of α) are able to produce comparable but slightly worse results than the CE criterion.

In the next series of experiments, the linear combination using weights 1, 2 and 3 of *Lin* and 0.5-CPA with CE was tested. The TER of the baseline CE and the most successful combinations are shown in Table 2. In all cases the linear combination outperforms the CE baseline slightly but consistent. Most of the improvement using the combination also remains after sequence-based training.

Table 2. *TER[%] results for frame-wise training using the CE criterion in combination with the 0.5-CPA and Lin (+LIN) criteria evaluated on the Vietnamese (V) and Bengali (B) test sets, followed by HF sMBR training.*

task	TER[%] criteria					
	CE	+HF	+2·0.5-CPA	+HF	+2·LIN	+HF
V dev	75.6	73.8	74.7	73.0	74.9	73.1
V eval	75.5	74.3	74.9	73.7	75.0	73.9
B dev	70.1	67.1	69.7	67.1	69.4	66.8

In a second series of experiments, the TER progress across different combination weights and initial learning rates of the last layer was investigated for the 0.5-CPA criteria to figure out a good weight across varying learning rates. Fig. 1 shows the TER progress for different learning rates for combinations of weight 1, 2 and 3 between the 0.5-CPA and CE criterion for the Bengali development set. The same observation was made for the Vietnamese development set as well. Interestingly, combination weight 2 is the ratio suggested by the factor $1/f''(1)$ of Theorem 1 for the 0.5-CPA criterion. Although the sample of this measurement is not large enough to make a real statement about this assumption this is an interesting direction for future work.

In the third series of experiments, the TER progress was measured for an equal combination of the α-CPA and CE criteria for different values of $\alpha = 1, 0.9, \ldots, 0.2$ and initial learning rates of the last layer. The corresponding results are shown in Fig. 2. The combination of both criteria results in improved results for most of the α values. Furthermore, a stable interval for α is observed from 0.4 through 0.7. The next section concludes the paper.

6. CONCLUSION

A novel family of discriminative training criteria based on the *f-Divergence* was derived from bounds on the classification error difference between the *Bayes* and model decision rule. An analytical proof of the bounds and criteria was presented which extends the classical derivation of the cross-entropy criterion from the Kullback-Leibler divergence to general *f-Divergences*. The novel criteria are closed under linear combination. Therefore, the linear combination of the cross-entropy with the novel criteria was successfully investigated in practice for frame-wise training of deep neural networks on the Babel tasks. The linear combination performed slightly but consistently better than the baseline cross-entropy criterion, most of the improvement also remained after sequence-based discriminative training. This gives reason to assume that a combination of training criteria performs better than a single criterion. Furthermore, this combination of the criteria comes with almost no additional computational cost compared to the cross-entropy criterion.

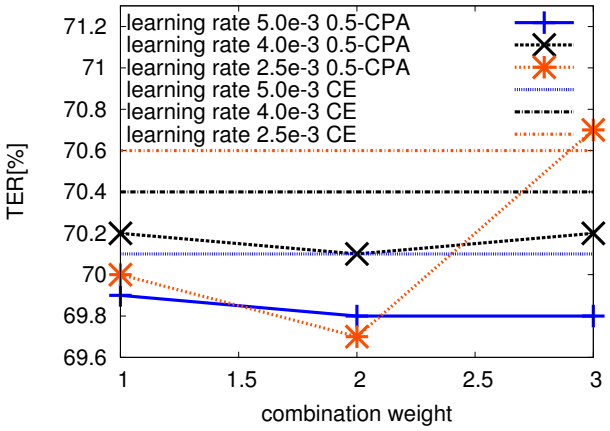

Fig. 1. TER[%] progress vs. the linear combination of the *Lin* and 1.5-CPA criterion with CE on the Bengali Babel test set.

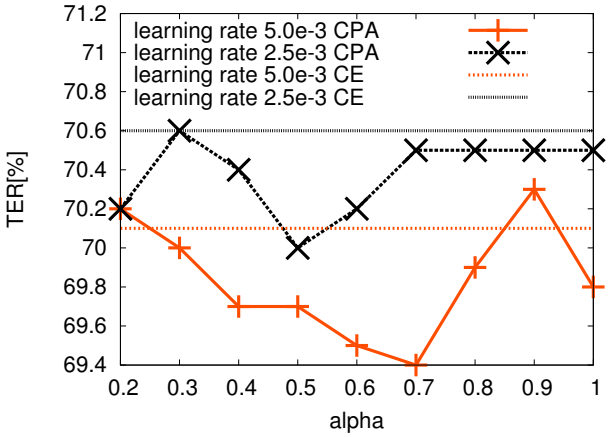

Fig. 2. TER[%] progress vs. factor α of the α-CPA criterion measured on the Bengali development set.

Acknowledgment: The first author would like to thank Xiaodong Cui and Vaibhava Goel for the supervision and helpful discussion during the internship. The Research leading to these results has received funding from the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense U.S. Army Research Labotory (DoD/ARL) contract no. W911NF-12-C-0012. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOD/ARL, or the U.S. Government. H. Ney was partially supported by a senior chair award from DIGITEO, a French research cluster in Ile-de-France.

7. REFERENCES

[1] I. Csiszár and P. C. Shields, "Information Theory and Statistics: A Tutorial," vol. 1, no. 4, 2004.

[2] F. Liese and I. Vajda, "On Divergences and Informations in Statistics and Information Theory," vol. 52, no. 10, pp. 4394–4412, 2006.

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

[4] A. A. Fedotov, P. Harremoës, and F. Topsøe, "Refinements of Pinsker's Inequality," vol. 49, no. 6, pp. 1491–1498, 2003.

[5] A. Guntuboyina, "Lower Bounds for the Minimax Risk using f-Divergences and Applications," in *IEEE Transactions on Information Theory*, 2011.

[6] V. Vapnik, Ed., *Statistical Learning Theory*, Wiley, 1998.

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

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

[9] M. Nußbaum-Thom, E. Beck, T. Alkhouli, R. Schlüter, and H. Ney, "Relative Error Bounds for Statistical Classifiers Based on the f-Divergence," in *Interspeech*, Lyon, France, Aug. 2013.

[10] F. Seide, G. Li, and D. Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks," in *INTERSPEECH*, 2011, pp. 437–440.

[11] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization," in *INTERSPEECH*, 2012.

[12] P. Golik, P. Doetsch, and H. Ney, "Cross-Entropy vs. Squared Error Training: a Theoretical and Experimental Comparison," in *Interspeech*, Lyon, France, Aug. 2013, pp. 1756–1760.

[13] K. Vesely, A. Ghoshal, L. Burget, and D. Povey, "Sequence-Discriminative Training of Deep Neural Networks," in *Interspeech*, 2013.

[14] H. Su, G. Li, D. Yu, and F. Seide, "Error Back Propagation for Sequence Training of Context-Dependent Deep Networks for Conversational Speech Transcription," in *ICASSP*, 2013.

[15] M. Hazewinkel, Ed., *Enyclopedia of Mathematics*, Springer, 2001.

[16] P. Lah and M. Ribarič, "Converse of Jensens Inequality for Convex Functions," in *Publications de la faculte d'Electrotechnique de Universite a Belgrade, ser, Mathematics et Physique, No. 412 - No. 460*, 1973, pp. 201–205.

[17] R. Bracewell, *The Sifting Property*, McGraw-Hill, 3rd ed. new york edition, 1999.

[18] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T. N. Sainath, and A. Sethy, "Developing Speech Recognition Systems for Corpus Indexing under the IARPA Babel Program," in *ICASSP*, 2013.

[19] R. Kneser and H. Ney, "Improved Backing-Off for m-Gram Language Modeling," in *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Detroit, Michigan, May 1995, vol. I, pp. 181–184.