

Investigations on Neural Networks, Discriminative Training Criteria and Error Bounds

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der
RWTH Aachen University zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Diplom-Informatiker
Markus Nußbaum-Thom
aus Mönchengladbach

Berichter: Universitätsprofessor Dr.-Ing. Hermann Ney
Universitätsprofessor Dr.-Ing. Reinhold Häb-Umbach

Tag der mündlichen Prüfung: 13.11.2020

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

ACKNOWLEDGMENTS

First, I would like to thank Prof. Dr.-Ing. Hermann Ney for giving me the opportunity to work and to research at the Lehrstuhl für Informatik 6. I learnt many things which still help me in scientific work today. I also would like to thank Prof. Dr.-Ing. Reinhold Häb-Umbach agreeing to review this thesis. Special thanks to Dr. Ralf Schlüter for his interest in my work, agreeing to review this thesis, the collaboration throughout the years, and his helpful feedback. Furthermore, I would like to thank the colleagues from the I6 institute Kazuki Irie, David Nolden, Simon Wiesler, Martin Sundermeyer, Pavel Golik, Jens Forster, Albert Zeyer, Saab Mansour, Markus Freitag, Stephan Peitz, Jörn Wübker as well as the colleagues from the IBM Watson Research Center Xiaodong Cui, Jia Cui, Bhuvana Rhamabhadran, Brian Kingsbury and Michael Picheny. Thanks to Stefan Koltermann and Kai Frantzen for maintaining the infrastructure where the experiments for this thesis were conducted. And thanks to the secretaries, in particular Steffi Jansen and Andrea Kierdorf, for interesting conversations and their support in formal matters. Last but not least, very special thanks go to my family. I have really appreciated their encouraging support during all the years. And my wife, I can not thank you enough for your endless support, your encouraging words, your empathy and your love. Talking with you in hard times was very important to get this done. Thank you.

ABSTRACT

The task of an automatic speech recognition system is to convert speech signals into written text by choosing the recognition result according to a statistical decision rule. The discriminative training of the underlying statistical model is an essential part to improve the word error rate performance of the system. In automatic speech recognition a mismatch exists between the loss used in the word error rate performance measure, the loss of the decision rule and the loss of the discriminative training criterion. In the course of this thesis the analysis of this mismatch leads to the development of novel error bounds and training criteria. The novel training criteria are evaluated in practical speech recognition experiments. In summary, we come to the conclusion the statistical model is able to compensate for this mismatch if the discriminative training criterion involves the loss of the performance measure.

Automatic speech recognition is based on *Bayes* decision rule. This rule chooses the most probable sentence as the recognition result for a given speech signal. The word error rate measures the performance of the recognition result. This measure is based on the *Levenshtein* loss and calculates the minimum number of insertions, deletions, and substitutions to transform the spoken into the recognized sentence. However, this choice of performance measure bears a fundamental mismatch to the one targeted in the maximum probability decision rule, as by definition, *Bayes* decision rule minimizes the sentence error rate, which does not guarantee to optimize the performance measure of automatic speech recognition — the word error rate. The straightforward approach to overcome this problem incorporates the *Levenshtein* loss into *Bayes* decision rule by choosing the recognition result according to the sentence minimizing the posterior-expected *Levenshtein* loss. Nevertheless, the evaluation of this decision rule is too time and memory consuming. It only is performed as a post-processing step after the search of the maximum probability decision rule.

In practice, we have to make a model assumption to *Bayes* decision theory. The theory assumes the true distribution, which is the empirical prior of all speech signals and spoken sentences. This distribution is unknown in practice. To stay as close to the principle of *Bayes* decision rule, a model distribution with free parameters substitutes the true distribution. The corresponding maximum probability decision rule using the model is called the model-based decision rule. The free parameters of the model are learned from training data, e.g., with generative training. Subsequently, discriminative training finetunes the model. For automatic speech recognition, the type of discriminative training criterion plays a crucial role. For example, the Minimum Phone Error (MPE) criterion, which involves the *Levenshtein* loss, performs better than other discriminative criteria like cross-entropy or maximum-mutual-information.

Apart from its superior practical performance, the MPE criterion has a lack of theoretical justification. In contrast to this criterion, the cross-entropy criterion can be derived based on a formal derivation scheme from the *Kullback-Leibler* divergence comparing the true and model distribution. In this scheme, the *Kullback-Leibler* divergence is an upper bound to the error difference between the model-based and *Bayes* decision rule. The error difference measures the

performance difference between both decision rules. For the MPE criterion, different from the cross-entropy criterion, no such derivation scheme exists relating the training criterion to an upper bound on the error difference. In this thesis, we close this gap and give a theoretical justification for the MPE criterion.

In the first part of this thesis, we develop a scheme to derive discriminative training criteria from bounds on the error difference between the model-based and *Bayes* decision rule. The *f-Divergence* is the basis for the examined error bounds. This divergence family is a generalization of the *Kullback-Leibler* divergence and is used to compare two distributions. We start by formulating proofs to derive upper *f-Divergence* bounds on the classification error difference. These proofs are then extended to error bounds based on a more general loss. These also include error bounds based on the *Levenshtein* loss, which are relevant to the mismatch between performance measure and model-based decision rule in automatic speech recognition. We ultimately find a type of explicit bound which is suitable to derive discriminative training criteria. Before this thesis, no derivation scheme for more general losses like the *Levenshtein* loss existed relating the training criterion to an upper bound on the error difference. Practical automatic speech recognition experiments evaluate our novel training criteria. These experiments include frame-wise training of neural network training as well as sequence training of log-linear mixture models. We show that our novel *f-Divergence* training criteria achieve a competitive or better performance than the conventional cross-entropy and minimum phone error criteria.

The second part of this thesis summarizes our successful participation in the QUAERO project evaluation campaign. We contributed the automatic speech recognition system for German in all project periods achieving the best or competitive results.

KURZFASSUNG

Die Aufgabe der Spracherkennung ist es mittels einer statistischen Entscheidungsregel Sprachsignale in geschriebene Sprache zu transformieren. Das diskriminative Training des entsprechenden statistischen Modells ist eine wichtige Komponente, um die Wortfehlerrate des Systems zu verbessern. Allerdings existiert in der Spracherkennung eine fundamentale Diskrepanz zwischen der Verlustfunktion der Wortfehlerrate, der Verlustfunktion der Entscheidungsregel und der Verlustfunktion des diskriminativen Trainingskriteriums. Im Laufe dieser Arbeit führt die Analyse dieser Diskrepanz zu neuen Fehlerschranken und Trainingskriterien, die in praktischen Spracherkennungsexperimenten ausgewertet werden. Zusammenfassend kann man sagen, dass das statistische Modell dazu fähig ist, die Diskrepanz zwischen der Verlustfunktion der Wortfehlerrate und der Entscheidungsregel zu überwinden, wenn das diskriminative Trainingskriterium die Verlustfunktion der Wortfehlerrate berücksichtigt.

Spracherkennung beruht auf der *Bayes* Entscheidungsregel. Diese wählt für ein gegebenes Sprachsignal den wahrscheinlichsten Satz als Erkennungsergebnis aus. Die Qualität der Erkennung wird anhand der Wortfehlerrate gemessen. Diese basiert auf der *Levenshtein* Verlustfunktion und gibt die minimale Anzahl von Einfüge-, Lösch- und Ersetzungsoperationen an, um den gesprochenen in den erkannten Satz zu transformieren. Doch dies birgt eine grundlegende Diskrepanz zwischen dem Qualitätsmaß und der Entscheidungsregel. Denn die *Bayes* Entscheidungsregel minimiert per Definition die Satzfehlerrate, was nicht zwingend zu einer Minimierung des Qualitätsmaßes – der Wortfehlerrate – führt. Um dieses Problem zu beseitigen, wäre es naheliegend, die *Levenshtein* Verlustfunktion in die *Bayes* Entscheidungsregel miteinzubeziehen. Hiermit würde die *Bayes* Entscheidungsregel die erwartete *Levenshtein* Verlustfunktion minimieren. Aber dieser Ansatz beansprucht in der Praxis zu viele Ressourcen (wie Zeit und Speicher) und kann daher nur als anschließender Verarbeitungsschritt zu einer Suche nach dem wahrscheinlichsten Satz durchgeführt werden.

In der Praxis müssen Modellannahmen an die *Bayes* Entscheidungsregel gemacht werden, da die *Bayes* Entscheidungstheorie auf der wahren Verteilung beruht, die durch die relativen Häufigkeiten des Sprachsignals und der gesprochenen Sätze definiert ist, und diese in der Praxis nicht bekannt sind. Um weiterhin dem Prinzip der *Bayes* Entscheidungsregel zu folgen, ersetzt man die wahre Verteilung durch eine Modellverteilung. Die entsprechende Entscheidungsregel heißt modellbasierte Entscheidungsregel. Die freien Parameter des Modells werden durch Trainingsdaten gelernt, zum Beispiel durch Training eines generativen Modells. Anschließend wird das Modell in mehreren Schritten durch diskriminatives Training verbessert.

Die mathematische Form des Trainingskriteriums spielt für die automatische Spracherkennung eine wichtige Rolle. Zum Beispiel erreicht das Kriterium, welches die erwartete *Levenshtein* Verlustfunktion der gesprochenen Phonemesequenz minimieren soll, in der Praxis die besten Ergebnisse. Theoretisch ist dieses Kriterium aber schlecht fundiert. Das entsprechende Kriterium wird auch das Minimum Phone Error Kriterium genannt. Im Gegensatz dazu ist das Kreuzentropiekri-

terium theoretisch gut fundiert und kann anhand eines formalen Schemas aus der *Kullback-Leibler* Divergenz zwischen der wahren und der Modellverteilung hergeleitet werden. In diesem Schema ist die *Kullback-Leibler* Divergenz eine obere Schranke an die Fehlerdifferenz zwischen der *Bayes* und modellbasierten Entscheidungsregel. Ein solches Schema existiert für das MPE Kriterium nicht. In dieser Arbeit schließen wir diese Lücke und leiten eine theoretische Motivation für diese Art von Kriterium her.

Im ersten Teil dieser Arbeit entwickeln wir ein Schema, um diskriminative Trainingskriterien aus oberen Schranken an die Fehlerdifferenz zwischen der *Bayes* und modellbasierten Entscheidungsregel herzuleiten. Alle untersuchten Fehlerschranken basieren auf der *f-Divergenz*, die eine Verallgemeinerung der *Kullback-Leibler* Divergenz ist. Beide Arten von Divergenzen werden verwendet, um jeweils zwei Verteilungen miteinander zu vergleichen. In dieser Arbeit formulieren wir Beweise von oberen *f-Divergenz*-Schranken an die Klassifikationsfehlerdifferenz. Diese Beweise werden dann erweitert zu Fehlerschranken für allgemeinere Verlustfunktionen. Sie umfassen auch den Fall der *Levenshtein* Verlustfunktion, und sind relevant für den Fall, in dem eine Diskrepanz zwischen dem Qualitätsmaß und der modellbasierten Entscheidungsregel vorliegt, wie zum Beispiel im Fall der Spracherkennung. Es zeigt sich, dass eine bestimmte Art von expliziten Schranken geeignet ist, um diskriminative Trainingskriterien herzuleiten. Es ist hervorzuheben, dass es vor dieser Arbeit kein Schema gab, welches das Trainingskriterium, basierend auf einer allgemeinen Verlustfunktion, wie der *Levenshtein* Verlustfunktion, aus einer oberen Schranke auf der Fehlerdifferenz herleiten kann. Die neuen Trainingskriterien werden in Experimenten auf praktischen Spracherkennungsdaten ausgewertet. In diesen Experimenten werden akustische Modelle, wie neuronale Netzwerke und log-lineare Mischverteilungen, diskriminativ trainiert. Wir verwenden für die Verfeinerung des akustischen Modells sowohl positionsweise als auch sequentielle Trainingskriterien. Wir zeigen, dass unsere neuen Trainingskriterien, die auf der *f-Divergenz* basieren, eine konkurrenzfähige Leistung im Vergleich zu konventionellen diskriminativen Kriterien erreichen.

Der zweite Teil dieser Arbeit fasst unsere erfolgreiche Teilnahme an der QUAERO Projekt-evaluierung zusammen, an der wir mit konkurrenzfähigen Spracherkennungssystemen in Deutsch teilgenommen haben.

CONTENTS

1	Introduction	1
1.1	Signal Analysis/Feature Extraction	3
1.2	Acoustic Modeling	3
1.3	Language Modeling	6
1.4	Search	6
2	State-of-the-Art in Training Criteria and Error Bounds	9
2.1	Discriminative Training Criteria	9
2.1.1	Frame-Wise Discriminative Training	9
2.1.2	Sequence Discriminative Training	10
2.1.3	Optimization	11
2.2	Bayes Decision Theory	11
2.2.1	Bayes Decision Theory for Decision Problems Using the 0-1 Loss	11
2.2.2	Classification Error	12
2.2.3	Model-Based Decision Rule	12
2.2.4	The Classification Error Difference as a Quality Measure	12
2.3	State-of-the-Art in Error Bounds	13
2.3.1	Loss Function Mismatch in Automatic Speech Recognition	15
2.4	Outline of this Thesis	15
2.5	Previously Published	16
3	Scientific Goals	19
4	Error Bounds	21
4.1	Derivation of Classification Error Bounds	21
4.1.1	Implicit Classification Error Bounds	26
4.1.2	Implicit Classification Error Bounds and Guntuboyinas Theorem	27
4.1.3	Implicit Classification Error Bounds with Closed Form Solution	27
4.1.4	Explicit Classification Error Bounds	29
4.2	Derivation of Error Bounds for General Loss Functions	31
4.2.1	Bayes Decision Theory for General Loss Functions	31
4.2.2	From the Local Error to Class Posteriors	32
4.2.3	Implicit Local Error Bounds	34
4.2.4	Implicit Error Bounds with Closed Form Solutions	35
4.2.5	Explicit Global Error Bounds	36
4.3	Joint Work	37

5	Error-Bound-Based Training Criteria	39
5.1	From Global Classification Error Bounds to Training Criteria	39
5.1.1	Examples for Discriminative Training Criteria Based on the f -Divergence .	41
5.1.2	Optimal Non-Parametric Solution for the Power Approximation f -Divergence	42
5.2	From Error Bounds to Training Criteria	43
5.2.1	Examples for Discriminative Training Criteria Based on the f -Divergence .	45
5.2.2	Optimal Non-Parametric Solution for the Power Approximation f -Divergence	47
5.3	Joint Work	51
6	Application to Frame-Wise Neural Network Training for Automatic Speech Recognition	53
6.1	Empirical Results for the Frame-Wise f -Divergence Training	53
6.2	Noisy f -Divergence Criteria and Minimum f -Divergence Training	56
6.3	Joint Work	62
7	Application to Sequence Training for Automatic Speech Recognition	63
7.1	Log-Linear Mixture Model	63
7.2	Transducer-Based Framework Implementation	64
7.3	Gradient of the Objective Function	64
7.4	Experimental Setup	64
7.5	Experimental Results	65
7.6	Joint Work	66
8	German QUAERO Evaluation Campaign	67
8.1	Language Resources	67
8.1.1	Training Data	68
8.1.2	Lexicon Modeling	68
8.1.3	Language Modeling	68
8.2	Acoustic Modeling	69
8.2.1	Speaker Normalization and Adaptation	69
8.2.2	MLP features	70
8.2.3	Discriminative training	70
8.3	Segmentation	70
8.4	System Development in 2009	71
8.4.1	Recognition Setup	71
8.4.2	Experiments	72
8.5	System Development in 2010	72
8.5.1	Recognition setup	72
8.5.2	Experiments	73
8.6	Evaluations	73
8.7	Joint Work	74
8.8	Acknowledgements	74
9	Scientific Achievements	75
10	Appendix	79
10.1	On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition	79
10.2	Statistical Learning Theory [Vapnik 98, p.30]	80
10.3	Automatic Speech Recognition is NP-Complete	82
10.4	Validation of f -Divergence Properties	83

List of Figures	93
List of Tables	95
Bibliography	97

1. INTRODUCTION

Automatic Speech Recognition has simplified the life of many people. Nowadays, automatic speech recognition is available on smartphones, smartwatches, televisions, or cars, just to name a few of many examples. Many people use it to perform their daily routine: to dictate, to perform commands using their favorite personal assistant, to play music, or as an intermediate step for speech-to-speech translation. Automatic speech recognition is especially helpful for disabled persons who are not able to use their hands to communicate or to perform commands in a smart home, for example.

The task of automatic speech recognition is to convert speech signals into written text. For this purpose, the recognition result for a given speech signal is the most probable sentence. The word error rate measures the performance of the recognition result based on the *Levenshtein* loss between the spoken and recognized sentence. This loss determines the minimum number of insertions, deletions, and substitutions to transform the spoken into the recognized sentence.

Traditionally an automatic speech recognition system is composed of *signal analysis*, an *acoustic* model, a *language* model, and the *global search*, as shown in Figure 1.1.

- The *signal analysis* performs a short-term spectral analysis to convert the continuous speech signal into low-dimensional features (Section 1.1).
- The *acoustic* model computes a probability of an acoustic state given a sequence of acoustic features (Section 1.2).
- The *language* model computes the probability of a word given a sequence of previous words (Section 1.3).
- The *global search* computes the most probable sentence. A combination of *acoustic* and *language* model probabilities composes the sentence probabilities (Section 1.4).

The *acoustic* and *language* model have free parameters. We learn those from training data. Traditionally generative training has been applied to the models followed by a discriminative fine-tuning. To test the approaches developed in this thesis in practice, we apply them to discriminative training of the *acoustic* model. Well-known examples of training criteria are the cross-entropy, the Maximum-Mutual-Information (MMI), and the MPE criterion. Section 2.1 gives an overview of the state-of-art in discriminative training.

Bayes decision rule is the basis of automatic speech recognition. This rule chooses the maximum probability sentence according to the true distribution. However, this distribution is the prior of speech signals and spoken sentences and is unknown in practice. Therefore, a model substitutes the true distribution instead. The corresponding choice of the maximum probability sentence according to the model distribution is called the model-based decision rule. In Section 2.2 a more general introduction to *Bayes* decision theory covers automatic speech recognition as a special case. By definition, *Bayes* decision rule minimizes the sentence error rate, which does guarantee

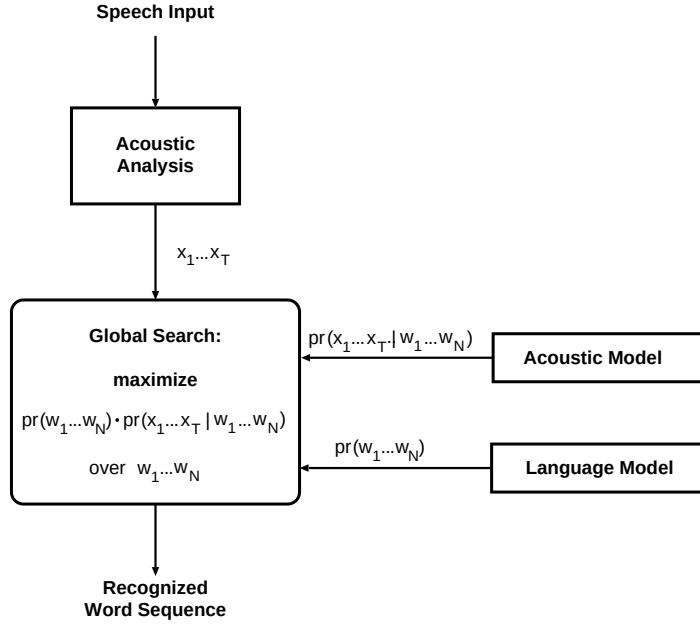


Figure 1.1: Architecture of an automatic speech recognition system adopted from [Ney 05].

to minimize the word error rate, and therefore the performance measure. This contradiction introduces a fundamental mismatch between the performance measure and the *Bayes* decision rule in automatic speech recognition. A detailed overview of the mismatch is given in Section 2.3.1. One solution to anticipate this mismatch is to include the *Levenshtein* loss into *Bayes* decision rule by choosing the recognition result according to the minimum expected *Levenshtein* loss. Nevertheless, in practice, this approach is too time- and memory-consuming and only feasible as an approximation through a post-processing step to the maximum probability decision rule.

Another approach to anticipate the mismatch is to fit the model towards the performance measure — the minimum word error rate — through the training criterion. MPE is such a criterion minimizing the posterior-expected *Levenshtein* loss of the spoken phoneme sequence. That is why MPE performs in practice better than other discriminative training criteria. However, the theoretical implications of this criterion are vague concerning the mismatch of performance measure and decision rule in automatic speech recognition.

Apart from its superior practical performance, the MPE criterion has a lack of theoretical justification. In contrast to MPE, the cross-entropy criterion can be derived from a formal derivation by the *Kullback-Leibler* divergence comparing the true and model distribution. In this case, the *Kullback-Leibler* divergence is an upper bound on the error difference between the *Bayes* and model-based decision rule. Nevertheless, for MPE no such derivation scheme exists relating the criterion to an upper bound on the error difference. We close this gap by developing a theoretical justification for the MPE criterion.

In this thesis, we contribute a scheme to derive discriminative training criteria from bounds on the error difference between the *Bayes* and model-based decision rule. The *f-Divergence* is the basis of these bounds, which is a generalization of the *Kullback-Leibler* divergence to compare two distributions. We first show proof for classification error bounds, which we extended to error bounds based on more general loss functions. These error bounds also include the case of the *Levenshtein* loss and are therefore relevant for the mismatch condition in automatic speech recognition. Our statements do not apply to automatic speech recognition only but address a wider scope of pattern recognition problems with loss functions different from the *Levenshtein*

loss. Furthermore, this thesis presents a specific type of explicit bounds which is suited to derive novel discriminative training criteria. We test the usefulness of these criteria in automatic speech recognition experiments on real-life data. Our experiments include frame-wise training of neural network training as well as sequence criteria of log-linear mixture models. The novel *f-Divergence* training criteria achieve competitive or even better performance than the conventional cross-entropy and MPE criterion.

1.1 Signal Analysis/Feature Extraction

The goal of the signal analysis is to reduce a continuous speech signal into a sequence of low-dimensional features. This transformation should ideally remove all irrelevant information from the speech signal, such as background noise, speaker identity, reverberation, and other effects.

Traditional speech recognition systems apply a short term spectral analysis [Rabiner & Schafer 78], based on a Fourier analysis. Two of the more widely used procedures for signal processing are Mel Frequency Cepstral Coefficients (MFCC) [Davis & Mermelstein 80] and Perceptual Linear Prediction (PLP) [Hermansky 90]. [Pitz 05] gives a more detailed description of the signal analysis.

A commonly used method to include dynamic information is augmenting the original feature vector with the first and second derivatives yielding a high dimensional vector. Linear Discriminant Analysis (LDA) is a more general approach concatenating feature vectors of neighboring time frames [Fisher 36, Duda & Hart⁺ 01]. The LDA is a linear transformation that projects a feature space into a lower-dimensional subspace. This projection maximizes the class separability for distributions with equal variances.

Several methods have been developed to overcome the speaker dependency of the acoustic feature vectors. Speaker normalization tries to reduce the speaker dependency by transforming the acoustic feature vectors. Speaker adaptation tries to adjust the model parameters of the speech recognition system to the characteristics of the given speaker. A unified view of speaker-dependent transformations is presented in [Pitz 05].

1.2 Acoustic Modeling

The acoustic model provides a statistical model $q(x_1^T | w_1^N)$ for the realization of a sequence of acoustic vectors x_1^T given a word sequence w_1^N . The acoustic model is a concatenation of the acoustic models for the basic sub-word units that the speech recognition system utilizes according to a pronunciation lexicon.

Depending on the amount of training data and the desired model complexity, the sub-word units are whole words, syllables, phonemes, or phonemes in context, or even words and therefore not using sub-units. Smaller units than words enable the speech recognition system to recognize words that have not occurred in the training data and to ensure that enough instances of each unit have been observed in training to allow a reliable parameter estimation. In Large Vocabulary Speech Recognition (LVCSR), the most commonly used sub-word units are phonemes in the context of one or two adjacent phonemes, so-called tri-phones, and quin-phones, respectively. Context-dependent phonemes (allophones) account for the different pronunciations of a phoneme depending on the surrounding phonemes.

The acoustic realizations of a sub-word unit differ significantly depending on the speaking rate. Variations in speaking rate are modeled by the Hidden Markov models (HMMs). HMMs have become the standard approach to automatic speech recognition systems [Baker 75, Rabiner 89]. An HMM is a stochastic finite state automaton consisting of many states and transitions between the states. The probability $q(x_1^T | w_1^N)$ is extended by unobservable (hidden) random variables

representing the states:

$$q(x_1^T | w_1^N) = \sum_{s_1^T} q(x_1^T, s_1^T | w_1^N).$$

The sum is over all possible state sequences s_1^T for a given word sequence w_1^N . Using Bayes' identity, this can be rewritten as

$$q(x_1^T | w_1^N) = \sum_{s_1^T} \prod_{t=1}^T q(x_t | x_1^{t-1}, s_1^t, w_1^N) \cdot q(s_t | x_1^{t-1}, s_1^{t-1}, w_1^N).$$

Applying the first-order Markov assumption [Duda & Hart⁺ 01] simplifies this equation. The probabilities $q(x_t | x_1^{t-1}, s_1^t, w_1^N)$ and $q(s_t | x_1^{t-1}, s_1^{t-1}, w_1^N)$ are assumed not to depend on previous observations but only on the states (Markov) and on the immediate predecessor state only (first-order):

$$q(x_1^T | w_1^N) = \sum_{s_1^T} \prod_{t=1}^T q(x_t | s_t, w_1^N) \cdot q(s_t | s_{t-1}, w_1^N). \quad (1.1)$$

Thus, the probability $q(x_1^T | w_1^N)$ splits into the *emission probability* $q(x_t | s_t, w_1^N)$ denoting the probability to observe an acoustic vector x_t while being in state s_t , and the *transition probability* $q(s_t | s_{t-1}, w_1^N)$ for a transition from state s_{t-1} to state s_t . Usually, the probability in (1.1) is evaluated in an expression that maximizes over all word sequences w_1^N . The value of the state sequence that maximizes the product in (1.2) approximates the value of the sum in (1.1).

$$q(x_1^T | w_1^N) \approx \max_{s_1^T} \left\{ \prod_{t=1}^T q(x_t | s_t, w_1^N) \cdot q(s_t | s_{t-1}, w_1^N) \right\}. \quad (1.2)$$

This approximation is called Viterbi or maximum approximation [Ney 90]. Equations (1.1) and (1.2) can be solved efficiently using the forward-backward algorithm [Baum 72, Rabiner & Juang 86], which is an example of dynamic programming [Bellman 57, Viterbi 67, Ney 84].

The topology used in this work has been introduced by Bakis [Bakis 76]: the basic HMM consists of six subsequent states where every two successive states are identical. Only transitions from a state to itself (*loop*), the next state (*forward*), and the next to next state (*skip*) are allowed. Using a frame-shift of 10ms, the path through the HMM with forward transitions only amounts to 60ms. This duration is close to the average duration of phonemes for most languages. This 6-state HMM has a minimum duration of 30ms (only skip transitions). Such an HMM is too long for fast conversational speech e.g., on the Verbmobil II corpus [Molau 03]. In this case, a 3-state model merges two identical states into a single one. This change reduces the minimum length of the HMM.

The emission probabilities $q(x_t | s_t, w_1^N)$ of an HMM can be modeled by discrete probabilities [Bahl & Baker⁺ 76], semi-continuous probabilities [Huang & Jack 89] or continuous probability distributions [Levinson & Rabiner⁺ 83]. A commonly used model for a continuous probability distribution is the Gaussian Mixture Model (GMM). Assuming GMMs, the emission probabilities read

$$q(x | s, w_1^N) = \sum_{l=1}^{L_s} c_{sl} \mathcal{N}(x | \mu_{sl}, \Sigma, w_1^N) \quad (1.3)$$

where c_{sl} denotes the non-negative mixture weights for HMM state s and mixture index $l \in \{1, \dots, L_s\}$ subject to the constraint $\sum_{l=1}^{L_s} c_{sl} = 1$, and $\mathcal{N}(x | \mu_{sl}, \Sigma, w_1^N)$ denotes the Gaussian density with mean μ_{sl} and covariance matrix Σ . The RWTH system uses a single globally pooled

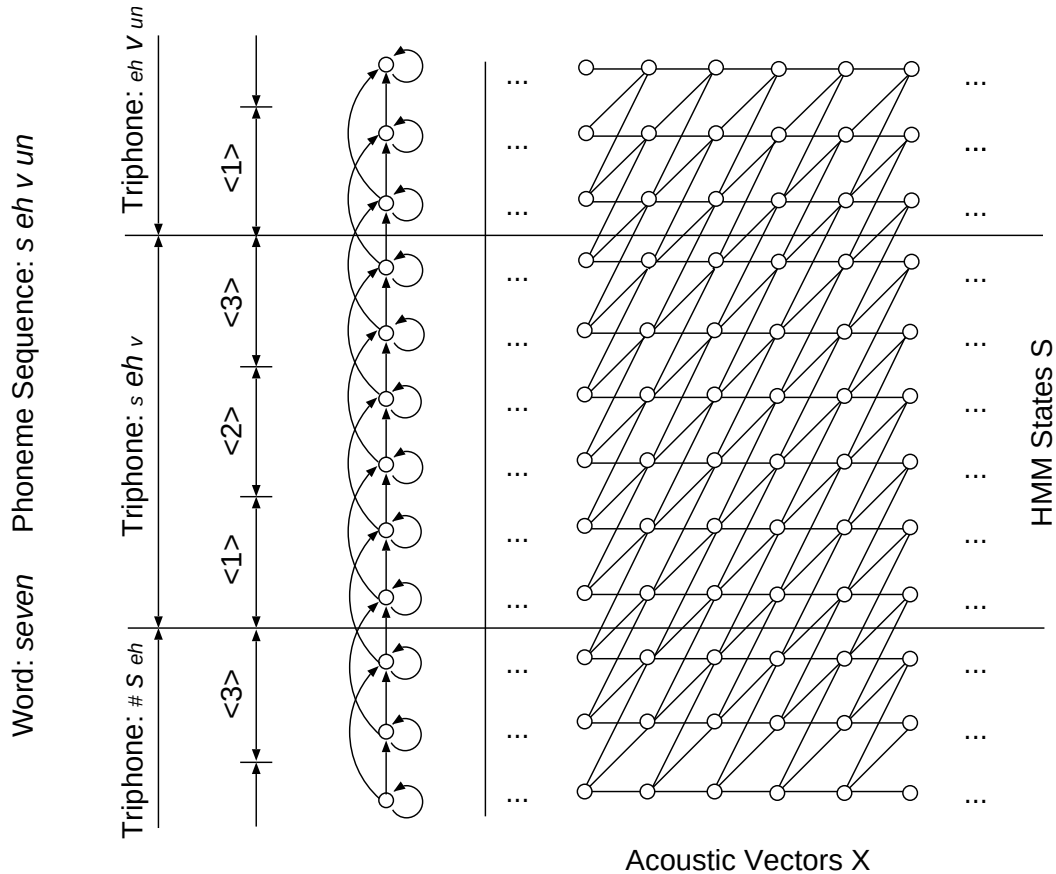


Figure 1.2: 6-state hidden Markov model in Bakis topology for the triphone $s eh_v$ in the word “seven”. The HMM segments are denoted by $\langle 1 \rangle$, $\langle 2 \rangle$, and $\langle 3 \rangle$.

and diagonal covariance matrix. This choice avoids problems caused by data sparseness, but it is also more efficient. Diagonal covariances assume decorrelated features. The feature decorrelation can be done, for instance, by LDA in a preprocessing step. Conventionally, the set of parameters $\Lambda = \{\{\mu_{sl}\}, \{c_{sl}\}, \Sigma\}$ is estimated according to the maximum likelihood (ML) training criterion in combination with the expectation-maximization (EM) algorithm [Dempster & Laird⁺ 77].

The number of distinct allophone states as basic sub-word units increases exponentially with the context length. Thus, a large number of allophones will have no or too few observations for reliable parameter estimation. Therefore, several states are tied together yielding *generalized* allophone models [Young 92]. Decision tree-based state clustering (e.g., CART) is part of almost all LVCSR systems. The main advantage of this top-down clustering method is that no backoff models need to be trained, and unseen allophones will be assigned to an appropriate HMM state. [Beulen & Ortmanns⁺ 99] described the state clustering of the RWTH system in detail. The pronunciation of a phoneme depends on the surrounding phonemes. Especially a phoneme at a word boundary is pronounced differently depending on the predecessor and successor words. Explicit modeling of this coarticulation uses *across-word* allophones [Hon & Lee 91, Odell & Valtchev⁺ 94]. This modeling takes respectively into account the ending and beginning phonemes of the adjacent words as a left and right context. Details of the across-word model implementation for the RWTH system can be found in [Sixtus 03].

1.3 Language Modeling

The language model $q(w_1^N)$ provides the probability for a word sequence $w_1^N = w_1, \dots, w_N$. This statistical model implicitly covers the syntax, semantics, and pragmatics of the language to be recognized. Due to the unlimited number of possible word sequences, further model assumptions help to estimate a reliable model. For LVCSR, m -gram language models [Bahl & Jelinek⁺ 83] have become widely accepted. The m -gram language models assume that the word sequence follows an $(m-1)$ -th order Markov process, i.e., the probability of the word w_n only depends on the $(m-1)$ predecessor words. Thus, the probability $q(w_1^N)$ factorizes into

$$\begin{aligned} q(w_1^N) &= \prod_{n=1}^N q(w_n | w_1^{n-1}) \\ &\stackrel{\text{model assumption}}{=} \prod_{n=1}^N q(w_n | w_{n-m+1}^{n-1}). \end{aligned} \quad (1.4)$$

The word sequence $h_n = w_{n-m+1}^{n-1}$ denotes the history of length m of the word w_n with the definitions $h := w_1^{n-1}$ if $n < m$ and $h := \emptyset$ if $n-1 < n-m+1$, e.g., at the boundary $q(w_1 | w_1^0) = p(w_1)$.

A commonly used measure for the evaluation of language models is the test set *perplexity* PP

$$PP = \left[\prod_{n=1}^N q(w_n | w_{n-m+1}^{n-1}) \right]^{-1/N}.$$

The log-perplexity is equal to the entropy of the model and perplexity can be interpreted as the average number of choices to continue a word sequence w_{n-m+1}^{n-1} at position n . When using the perplexity as optimization criterion for training the language model, closed-form solutions for $p(w|h)$ can be derived, which are equal to the relative frequency of the word sequence on the training corpus. The number of possible m -grams increases exponentially with the history length m . Thus, for a large vocabulary V , a considerable amount of m -grams will not be seen in training or has too few observations for a reliable estimation of $q(w|h)$, even for very large training corpora. Therefore, smoothing methods have to be applied. The smoothing is based on discounting in combination with backing-off or interpolation [Katz 87, Ney & Essen⁺ 94, Genet & Ney⁺ 95, Ney & Martin⁺ 97]. Discounting subtracts probability mass from seen events, which is then distributed over all unseen events (backing-off) or over all events (interpolation), usually in combination with a language model with a shorter history. The parameters of the smoothed language model can be estimated using a cross-validation scheme like leaving-one-out [Ney & Essen⁺ 94].

1.4 Search

The search module of the speech recognition system combines the acoustic model and language model, as depicted in Figure 1.1. The objective of the search is to find the word sequence that maximizes the *a posteriori* probability for a given sequence x_1^T of acoustic feature vectors according to (1.5)

$$\begin{aligned} c_{0-1}(x_1^T) &= \operatorname{argmax}_{w_1^N} \{q(w_1^N | x_1^T)\} \\ &= \operatorname{argmax}_{w_1^N} \{q(w_1^N) \cdot q(x_1^N | w_1^N)\}. \end{aligned} \quad (1.5)$$

If the language model is an m -gram model as given in (1.4) and the acoustic model is an HMM as given in (1.1), the following optimization problem has to be solved by the search module:

$$c_{0-1}(x_1^T) = \operatorname{argmax}_{w_1^N} \left\{ \left[\prod_{n=1}^N q(w_n | w_{n-m+1}^{n-1}) \right] \cdot \left[\sum_{s_1^T: w_1^N} \prod_{t=1}^T q(x_t | s_t, w_1^N) \cdot q(s_t | s_{t-1}, w_1^N) \right] \right\}$$

$$\stackrel{\text{Viterbi approx.}}{=} \operatorname{argmax}_{w_1^N} \left\{ \left[\prod_{n=1}^N q(w_n | w_{n-m+1}^{n-1}) \right] \cdot \left[\max_{s_1^T: w_1^N} \prod_{t=1}^T q(x_t | s_t, w_1^N) \cdot q(s_t | s_{t-1}, w_1^N) \right] \right\}. \quad (1.6)$$

In the second step, the Viterbi approximation simplifies the HMM. This approximation reduces the complexity of the optimization problem significantly. Equation (1.6) can be solved efficiently using dynamic programming [Bellman 57]. Dynamic programming exploits the mathematical structure and divides the problem into sub-instances. Like in all search problems, the search can be organized in two different ways: a depth-first and breadth-first search. The A^* -search or stack-decoding algorithm uses a depth-first search. Here, the state hypotheses are expanded time-asynchronously depending on a heuristic estimate of the costs to complete the path [Jelinek 69, Paul 91].

The *Viterbi* search uses the breadth-first search design where all state hypotheses expand time-synchronously [Vintsyuk 71, Baker 75, Sakoe 79, Ney 84]. This approach computes the probabilities of all hypotheses up to a given time frame. These hypotheses compare to each other then time-synchronously. This time-synchronous comparison allows reducing the search space significantly by pruning unlikely hypotheses early in the search process. Especially in the breadth-first approach, an efficient pruning is necessary as the number of possible word sequences with maximum length N grows exponentially with N . Thus, full optimization of (1.6) is only feasible for small vocabulary sizes $|W|$. For large vocabulary sizes, approximations have been made. Instead of finding the exact optimal solution of (1.6), the goal is to find a sufficiently good solution with much less effort. In the so-called beam-search, only that fraction of the hypotheses expands that likelihood, which is sufficiently close to that of the best hypothesis of the given time frame [Lowerre 76, Ney & Mergel⁺ 87, Ortmanns & Ney 95]. Beam-search does not guarantee to find the globally best word sequence. This optimal sequence may get pruned at an intermediate search stage due to a poor likelihood. However, if the pruning parameters are adjusted properly, no significant search errors occur, and the search effort reduces considerably.

Several other methods can be applied to reduce further the computational complexity of the Viterbi or beam-search, including lexical prefix tree [Ney & Hüb-Umbach⁺ 92], look-ahead [Steinbiss & Ney⁺ 93, Hüb-Umbach & Ney 94, Odell & Valtchev⁺ 94, Alleva & Huang⁺ 96, Ortmanns & Ney⁺ 96], and fast likelihood computation [Ramasubramanian & Paliwal 92, Bocchieri 93, Fritsch 97, Ortmanns & Ney⁺ 97b, Ortmanns 98, Kanthak & Schütz⁺ 00]. More advanced algorithms involving search (e.g., discriminative training) use N -best lists [Schwartz & Chow 90, Schwartz & Austin 91, Steinbiss 91] or word lattices [Ney & Aubert 94, Ortmanns & Ney⁺ 97a, Macherey 10] to reduce the search space.

2. STATE-OF-THE-ART IN TRAINING CRITERIA AND ERROR BOUNDS

In this section, we give an overview of the state-of-the-art related to this work. The relevant topics include state-of-the-art in discriminative training of *acoustic* models, an introduction to *Bayes* decision theory based on the classification error, the state-of-the-art in classification error bounds, and a discussion of the mismatch between performance measure and the decision rule in automatic speech recognition. At first we discuss the state-of-the-art of discriminative training criteria that are relevant for the frame-wise and sequence criteria developed in this thesis.

2.1 Discriminative Training Criteria

In the past, Gaussian Hidden Markov Models (GHMMs) were used. Traditionally the free parameters of GHMMs are estimated with generative training using the Maximum-Likelihood (ML) approach [Rabiner & Juang 86, Rabiner 89]. The maximum-likelihood criterion maximizes the class-conditional probability during training.

But in recent years GHMMs have been outperformed by neural networks which achieved a large word error rate gain [Mohamed & Dahl⁺ 09, Dahl & Ranzato⁺ 10, Seide & Li⁺ 11, Kingsbury & Sainath⁺ 12, Deng & Acero⁺ 12, Graves & Jaitly⁺ 13]. In this context, log-linear models are a special type of neural network with just one softmax output layer and no hidden layers. By default, neural network models with a softmax output layer are discriminative. The model is referred to as discriminative since the numerator competes (or discriminates) against the denominator re-normalization terms. For a generative model, discrimination is achievable by defining the class posterior probability explicitly. In the following, all discriminative criteria include the class posterior probability as an argument.

2.1.1 Frame-Wise Discriminative Training

The most popular discriminative training criterion is the Cross-Entropy (CE) criterion. This criterion is also known as the MMI criterion [Bahl & Brown⁺ 86, Bridle 90]. In automatic speech recognition cross-entropy usually is a frame-wise discriminative training [Heigold & Ney⁺ 12]. However, it is also known as frame-discriminative MMI [Povey & Woodland 99, Povey & Woodland 02a]. Given training data, a sequence of most likely HMM emission states s_1^T corresponding to the features x_1^T ; the model formulates a state posterior $q(s_t|x_t)$ per frame. For training, the HMM states are aligned to the features using a preliminary model. Then the cross-entropy criterion for frame-wise discriminative training is defined as:

$$\mathcal{F}_{\text{CE}}(q) = - \sum_{t=1}^T \log q(s_t|x_t).$$

At inference time in LVCSR the state posterior is applied as emission probability using the hybrid approach [Bourlard & Morgan 93]. The emission probability is expressed from the state posterior probability $q(s|x)$ and the state prior $q(s)$ using *Bayes* rule:

$$q(x|s) = \frac{q(s|x)q(x)}{q(s)} \propto \frac{q(s|x)}{q(s)}.$$

The re-normalization term $q(x)$ in *Bayes* rule is dropped during inference since it is constant with respect to the argument of maximization at inference time. The prior $q(s)$ can easily be estimated from relative frequencies of the given state sequence s_1^T .

2.1.2 Sequence Discriminative Training

In the case of automatic speech recognition, the word sequences $w_1^N = w_1, \dots, w_N$ correspond to HMM state sequences $s_1^T = s_1, \dots, s_T$ for given features $x_1^T = x_1, \dots, x_T$. In the following the notation $\tilde{s}_1^T : \tilde{w}_1^M$ refers to the HMM state sequence $\tilde{s}_1^T = \tilde{s}_1, \dots, \tilde{s}_T$ for the word sequence $\tilde{w}_1^M = \tilde{w}_1, \dots, \tilde{w}_M$ for given features x_1^T . The model-based posterior of the word sequence involving HMMs is formulated by

$$q(w_1^N | x_1^T) = \sum_{s_1^T : w_1^N} \frac{[q(x_1^T, s_1^T, w_1^N)]^\gamma}{Z(x_1^T | \gamma)}$$

with the posterior re-normalization

$$Z(x_1^T | \gamma) = \sum_{M, \tilde{w}_1^M} \sum_{\tilde{s}_1^T : \tilde{w}_1^M} [q(x_1^T, \tilde{s}_1^T, \tilde{w}_1^M)]^\gamma$$

and the posterior scale γ .

The term MMI is usually used to name the sequence criterion corresponding the cross-entropy criterion [Bahl & Brown⁺ 86, Chow 90, Normandin & Morgera 91, Kapadia & Valtchev⁺ 93, Cardin & Normandin⁺ 93, Bahl & Padmanabhan⁺ 96, Schlüter 00, Woodland & Povey 00, Woodland & Povey 02]. For a given set of spoken sentences $W_1^N := W_1, \dots, W_N$ and the corresponding acoustic feature sequences $X_1^N := X_1, \dots, X_N$ the sequence MMI criterion is defined as:

$$\mathcal{F}_{\text{MMI}}(q) = - \sum_{n=1}^N \log q(W_n | X_n).$$

Another type of sequence criterion directly involves the *Levenshtein* loss of the performance measure. In this case, the classes are word, phoneme, or state sequences, and the criterion is purely sequence-based. The criterion involves the *Levenshtein* accuracy between word, phoneme, or state sequences. The MPE or Minimum Word Error (MWE) criteria maximize the model-based posterior expected *Levenshtein* accuracy of the spoken phoneme or word sequence [Kaiser & Horvat⁺ 00, Kaiser & Horvat⁺ 02, Doumpiotis & Byrne 04, Schlüter 00, Povey & Woodland 02a, Povey 04, Macherey & Haferkamp⁺ 05, Zheng & Stolcke 05, Heigold 10, Wiesler 16]. While [Kaiser & Horvat⁺ 00, Kaiser & Horvat⁺ 02] uses N -best lists to approximate the search space of the competing word sequences for the sequence criterion, [Doumpiotis & Byrne 04] applies pinched lattices. Most other methods use lattices to model this search space [Schlüter 00, Povey & Woodland 02a, Macherey & Haferkamp⁺ 05, Povey 04, Zheng & Stolcke 05, Heigold 10, Wiesler 16].

2.1.3 Optimization

In general, the optimum model of a discriminative training criterion has no closed-form solution. Instead, gradient-based methods optimize the criterion and the model iteratively. The most popular method to optimize criteria using neural network models is Stochastic Gradient Descent (SGD) [Rumelhart & Hinton⁺ 86]. The stochastic gradient descent method applies the error back-propagation algorithm computing forward and backward passes. First, in the forward pass, activations of the neural network are calculated. Second, in the backward pass, the gradients of the training criterion and model are calculated. Starting from this gradient subsequently, the neural network gradient is calculated layer-wise recursively from output to input by applying the chain rule. The update rule of the model parameters is applied in a step-wise fashion in which the previous model parameters and the model gradient are scaled by the learning rate. The parameters are initialized by random or by pre-training.

2.2 Bayes Decision Theory

This section introduces the *Bayes* decision theory [Duda & Hart⁺ 00, Berger 85], which is a theoretical framework to analyze decision making. A statistical decision problem exists whenever, for a given observation $x \in \mathcal{X}$, there is a set of classes $c \in \mathcal{C}$, and the preference of one class over another depends on an unknown probability distribution and the error incurred by a choosing one class compared to another is determined by a loss function. *Bayes* decision rule is the error-optimal decision rule to such a statistical decision problem that prefers the class that minimizes the posterior expected loss. But the posterior is based on an unknown distribution. Since the underlying distribution is unknown, a model-based decision rule approximates the *Bayes* decision problem involving a statistical model with free parameters. The optimization step estimates the model parameters from training samples of the unknown distribution by minimizing a training criterion. The goal of the estimation process is to deliver a model-based decision rule that minimizes the error. Under ideal conditions the model-based decision rule is identical to the *Bayes* decision rule.

From now on, we refer to this unknown distribution as the true distribution. A loss function determines the cost incurred by choosing one class over another.

2.2.1 Bayes Decision Theory for Decision Problems Using the 0-1 Loss

In this section, we focus on the most common type of decision problem based on the 0-1 loss function (cost 0 for a match and 1 otherwise), which results in the classification error as a performance measure. We start by defining some fundamental ingredients of *Bayes* decision theory.

A *decision rule* $r : \mathcal{X} \rightarrow \mathcal{C}$ categorizes *observations* $x \in \mathcal{X}$ into *classes* $c \in \mathcal{C}$. Observations and classes occur according to the *true distribution* $pr(x, c)$. Further, the *loss function* $\mathcal{L} : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}^+$ measures the cost of confusing two classes. Lower costs indicate that the two compared classes have a higher similarity. The 0-1 loss is defined by:

$$\mathcal{L}_{0-1}(c, \tilde{c}) = 1 - \delta(c, \tilde{c}) \quad \text{based on Kronecker delta} \quad \delta(c, \tilde{c}) = \begin{cases} 0, & c \neq \tilde{c} \\ 1, & c = \tilde{c} \end{cases}.$$

Section 2.2.2 introduces the classification error which, is based on the 0-1 loss function.

2.2.2 Classification Error

The local and global classification error to a decision rule $r : \mathcal{X} \rightarrow \mathcal{C}$ is the posterior and joint expected loss of the corresponding decision rule. Therefore, the local and global classification error evaluate to the following equations:

$$\begin{aligned} E_{0-1}(c|x) &:= 1 - pr(c|x), && \text{"local classification error of class } c\text{"} \\ E_{0-1}(r) &:= \int pr(x)E_{0-1}(r(x)|x) dx. && \text{"global classification error"} \end{aligned}$$

By definition, the *Bayes* decision rule minimizes the global classification error. This decision rule is also identical to the Maximum-A-Posteriori (MAP) decision rule:

$$\begin{aligned} c_{0-1}(x) &:= \operatorname{argmin}_{c \in \mathcal{C}} \{E_{0-1}(c|x)\} \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \{pr(c|x)\}. \end{aligned}$$

2.2.3 Model-Based Decision Rule

The true distribution is unknown in practice and only accessible through samples. Therefore, a *model-based distribution* $q(x, c)$ approximates this distribution. The model-based distribution learns the free parameters from training data. The free parameters are trained with samples $(x_n, c_n), n = 1, \dots, N$, drawn according to the true distribution, to minimize the discrepancy between the model and the true distribution. Analogous to the case of the true distribution, we can also define a model-based classification error for the model distribution. The model-based posterior expected loss defines the model-based local classification error.

$$E_{0-1}^q(c|x) := 1 - q(c|x).$$

In the model-based decision rule the true is replaced by the model-based posterior.

$$c_{0-1}^q(x) := \operatorname{argmax}_{c \in \mathcal{C}} \{q(c|x)\}. \quad (2.1)$$

One interpretation is that the decision rule should approximate the *Bayes* decision rule based on the 0-1 loss as close as possible. Another interpretation is to train the model error in an oriented way and to shift the decision boundaries e.g. in favor of the performance measure.

The next section introduces the quality measure used for our analysis of error bounds and training criteria.

2.2.4 The Classification Error Difference as a Quality Measure

A quality measure of a decision rule is the error difference of a decision rule $r : \mathcal{X} \rightarrow \mathcal{C}$ and the *Bayes* decision rule. We distinguish between the local error difference for an observation x and the global error difference integrated over all observations.

$$\begin{aligned} \Delta_{0-1}(x) &:= pr(c_{0-1}(x)|x) - pr(r(x)|x), && \text{"local classification error difference"} \\ \Delta_{0-1} &:= \int pr(x)\Delta_{0-1}(x) dx. && \text{"global classification error difference"} \end{aligned}$$

From now on the decision rule $r : \mathcal{X} \rightarrow \mathcal{C}$ will mostly be the model-based decision rule using the 0-1 loss from (2.1). If the error difference uses the 0-1 loss also the loss function index is ignored in favor of the following notation:

$$\begin{aligned} \Delta_{0-1}(x) &\rightsquigarrow \Delta(x) \\ \Delta_{0-1} &\rightsquigarrow \Delta. \end{aligned}$$

Section 2.3 discusses state-of-the art error bounds and training criteria.

2.3 State-of-the-Art in Error Bounds

A part of *Bayes* decision theory is concerned with the relation of the error, error bounds, and the training criterion, as discussed in [Vapnik 98, p.30] and [Ney 03, Guntuboyina 11]. Such a relation is known for decision problems based on the 0-1 loss function (error cost one and zero caused by incorrect and correct classification). In summary, this relates the error to an error bound expressed through the *Kullback-Leibler* divergence, simultaneously the error-bound relates to the cross-entropy criterion. Unlike other error bounds, the reviewed error bounds, in the remainder of this section, also consider the model-based decision rule in addition to the *Bayes* decision rule. In this thesis, we investigate these statements found for decision problems based on the 0-1 loss function to a more general type of loss function and establish a new relationship between the error, error bounds, and the training criterion. These novel error bounds also relate the *Bayes* and model-based decision rule. We subsequently show the connection of this work to discriminative training criteria.

In machine learning, in the context of density estimation, the Bretagnolle-Huber bound is known [Vapnik 98, p.30]. In this reviewed work, the total variational mismatch between the joint true and model distribution [Cover & Thomas 06, p.369] is the central value to be bounded.

$$V := \int \sum_{c \in \mathcal{C}} |pr(x, c) - q(x, c)| dx. \quad (2.2)$$

The following relation to the global classification error mismatch, the total variational mismatch, and the *Kullback-Leibler* divergence $D_{\text{KL}}(pr||q)$ can be established [Vapnik 98, p.30].

$$\begin{aligned} \Delta^2 &\leq V^2 \leq 4(1 - \exp(-D_{\text{KL}}(pr||q))) \\ \text{with } D_{\text{KL}}(pr||q) &= - \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \frac{q(c|x)}{pr(c|x)} dx \\ \text{and } \Delta &= \int pr(x, c_{0-1}(x)) - pr(x, c_{0-1}^q(x)) dx. \end{aligned} \quad (2.3)$$

A proof of (2.3) only partially exists in [Vapnik 98, p.30]. Since the complete proof is missing in the literature, we reconstructed it in Appendix 10.2.

Also, the *Pinsker* inequality connects total variational distance and *Kullback-Leibler* distance [Cover & Thomas 06, p.370] in the following way:

$$\Delta^2 \leq V^2 \leq 2D_{\text{KL}}(pr||q). \quad (2.4)$$

An improvement of *Pinsker's* inequality was presented by *Vajda* [Fedotov & Harremos⁺ 03]

$$2D_{\text{KL}}(pr||q) \geq \log \left(\frac{2+V}{2-V} \right) - \frac{2V}{2+V}. \quad (2.5)$$

Vajda also proposed a tight bound, for which *Fedotov* presented a parametrized form in [Fedotov & Harremos⁺ 03].

In [Ney 03], the *Kullback-Leibler* divergence establishes an upper bound on the classification error difference. In this publication, also a framework is presented on how to derive empirical training criteria from those error bounds. Local error bounds are derived starting from the local error difference by deriving basic inequalities between the local model-based and *Bayes* classification error. Subsequently, using relations to the L_∞ and L_2 norm, those bounds are

widened.

$$\Delta(x) \leq \sum_{c \in \mathcal{C}} |q(c|x) - pr(c|x)|, \quad (2.6)$$

$$\Delta(x) \leq 2 \max_{c \in \mathcal{C}} \{|q(c|x) - pr(c|x)|\} \leq 2 \sqrt{\sum_{c \in \mathcal{C}} (q(c|x) - pr(c|x))^2}. \quad (2.7)$$

Global error bounds are then derived by computing the expectation from these local error bounds and these error bounds are then expanded to:

$$\begin{aligned} \Delta^2 &\leq 4 \int pr(x) \sum_{c \in \mathcal{C}} (q(c|x) - pr(c|x))^2 dx, \\ \Delta^2 &\leq -2 \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \frac{q(c|x)}{pr(c|x)} dx. \end{aligned}$$

The empirical distribution is formulated by the *Dirac* delta $\delta(x - x_n)$ and the *Kronecker* delta $\delta(c, c_n)$, which result once substituted into the derived bounds in training criteria by using the sifting property [Bracewell 99] of the Dirac delta.

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

The Squared-Error (SE) criterion

$$\mathcal{F}_{SE}(q) = \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} (q(c|x_n) - \delta(c, c_n))^2,$$

and cross-entropy or MMI criterion are derived from

$$\mathcal{F}_{MMI}(q) = -\frac{1}{N} \sum_{n=1}^N \log q(c_n|x_n)$$

by evaluating the true distribution $pr(x, c)$ through the empirical distribution $pr_N(x, c)$ at the sampling points $(x_n, c_n), n = 1, \dots, N$. A more detailed derivation of this scheme is given in Appendix 10.1.

The work in [Guntuboyina 11] is not directly related to bounds on the error difference but establishes lower bounds on the minimax risk in the context of density estimation. Even though this paper focuses on a different problem, one of the main contributions of the paper is an implicit bound on the variational distance between two density distributions based on the *f-Divergence*, which is a generalization of the *Kullback-Leibler* divergence. The *f-Divergence* is further investigated in more detail in Section 4.1 while only introduced briefly here. The key statement in [Guntuboyina 11] builds around the *f-Divergence*. Consider the variational distance between two distributions pr and $\overline{pr}$

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

The *f-Divergence* for densities pr , $\overline{pr}$, and q for a convex function with $f(1) = 0$ is defined as

$$\begin{aligned} D_f(pr||q) &:= \int \sum_{c \in \mathcal{C}} q(c, x) f\left(\frac{pr(c, x)}{q(c, x)}\right) dx, \\ D_f(\overline{pr}||q) &:= \int \sum_{c \in \mathcal{C}} q(c, x) f\left(\frac{\overline{pr}(c, x)}{q(c, x)}\right) dx. \end{aligned}$$

Based on the *f-Divergence*, a novel implicit bound on the variational distance is introduced [Guntuboyina 11, p.3, Corollary II.3]:

$$\begin{aligned} D_f(pr||q) + D_f(\overline{pr}||q) &\geq \inf_{\pi} \{D_f(pr||\pi) + D_f(\overline{pr}||\pi)\} \\ &\geq f\left(1 - \frac{1}{2}V\right) + f\left(1 + \frac{1}{2}V\right). \end{aligned}$$

As shown in Section 4.1.4 this implicit bound can be made explicit under certain conditions.

2.3.1 Loss Function Mismatch in Automatic Speech Recognition

The *Bayes* decision rule minimizes the sentence error rate and is not guaranteed to minimize the word error rate. The word error rate is the performance measure of automatic speech recognition. An exact evaluation for automatic speech recognition should minimize the *Bayes* decision rule based on the posterior expected *Levenshtein* loss instead. However, in Appendix 10.3 of this thesis, we show that this *Bayes* decision rule based on the *Levenshtein* loss is NP-complete. It can only be evaluated approximatively as a post-processing step to the MAP decision rule [Stolcke & König⁺ 97, Goel & Byrne⁺ 98, Mangu & Brill⁺ 00, Evermann & Woodland 00b, Wessel & Schlüter⁺ 01, Hoffmeister & Klein⁺ 06]. An exact evaluation is too time and memory consuming, and in practice, instead, the model-based decision rule based on the 0-1 loss is evaluated efficiently. So in automatic speech recognition, there is a fundamental mismatch between the loss function used in the performance measure – the *Levenshtein* loss – and the loss function used in the model-based decision rule – the 0-1 loss.

Another way to anticipate this mismatch is through discriminative training of the model. If the model uses a training criterion that minimizes the model-based posterior expected *Levenshtein* loss, it generally is assumed that the decision boundaries of the model are shifted to match the word error rate closer. In practice, the superior performance of MPE [Woodland & Povey 02, Povey & Woodland 02a, Povey 04, Povey & Kingsbury 07, Heigold & Deselaers⁺ 08, Heigold 10] and related criteria emphasizes this assumption, but in theory, these criteria are not justified. It is not clear how the estimation, e.g., using the MPE criterion based on the *Levenshtein* loss, affects the model. Specifically, what effect has the model-based decision rule using the MPE model on the performance measure? These questions have not been answered by the existing *Bayes* decision theory yet.

2.4 Outline of this Thesis

This thesis investigates the relation between the error, error bounds, and training criteria for automatic speech recognition. Nevertheless, the relations found do also apply to statistical decision problems based on different losses in general. Section 3 lays out the scientific goals of this work. Our contributions start from Section 4 onwards. In Section 4 we establish novel *f-Divergence* error bounds on the error difference for decision problems based on the 0-1-loss and a more general type of loss. This type of loss also includes the *Levenshtein* loss of automatic speech recognition. From those novel error bounds in Section 5 empirical training criteria are derived. In Section 6 and 7, the empirical training criteria are interpreted and evaluated for frame-wise training of neural networks and sequence-based training of log-linear mixture models for automatic speech recognition.

In the final Section 9 the scientific achievements of this work are discussed.

2.5 Previously Published

During my thesis, the following papers have been published at peer-reviewed conferences. For each paper, we distinguish my individual contributions apart from those of the other contributors.

- Error Bounds and Discriminative Training Criteria:
 - [Nußbaum-Thom & Tüske⁺ 12] "Posterior-Scaled MPE: Novel Discriminative Training Criteria": My individual contribution in this paper is the discovery of novel training criteria from bounds on the error difference. I also implemented, trained all criteria and evaluated them experimentally on the TIMIT corpus. Zoltan Tüske helped me to set up the baseline maximum likelihood trained system for TIMIT in this paper. Georg Heigold implemented the transducer-based discriminative training framework in the RWTH speech recognition system. I am also grateful to my co-authors for all useful discussions and the proof-reading of this paper.
 - [Schlüter & Nußbaum-Thom⁺ 13] "Novel Tight Classification Error Bounds under Mismatch Conditions based on the f -Divergence": Ralf Schlüter discovered the general proof of the error bounds based on the f -Divergence presented in this paper. My individual contribution in this work is an earlier discovery of a proof for the reversed *Kullback-Leibler* bound, which gave inspiration to the more general proof. In addition to the proof presented in this paper, I also discovered a general proof based on [Guntuboyina 11, p.3, Corollary II.3]. In the paper, this proof is only mentioned briefly. In this thesis the complete proof is presented. Eugen Beck implemented and evaluated the simulation framework as part of his master thesis. I am also grateful to my co-authors for all useful discussions and the proof-reading of this paper.
 - [Nußbaum-Thom & Beck⁺ 13] "Relative error bounds for statistical classifiers based on the f -divergence": Ralf Schlüter discovered the general proof of the error bounds based on the f -Divergence presented in this paper. My individual contribution in this work is an earlier discovery of a proof for the reversed *Kullback-Leibler* bound, which gave inspiration to the more general proof. In addition to the proof presented in this paper, I also discovered a general proof based on [Guntuboyina 11, p.3, Corollary II.3]. In this thesis the complete proof is presented. Eugen Beck implemented and evaluated the simulation framework as part of his master thesis. I am also grateful to my co-authors for all useful discussions and the proof-reading of this paper.
 - [Nußbaum-Thom & Cui⁺ 14] "A Family of Discriminative Training Criteria based on the f -Divergence for Deep Neural Networks": I am very thankful to my co-authors for proof-reading my paper. The Babel baseline systems were built by the IBM Babel speech team at the IBM Watson Yorktown Heights Research Center. We modified these baseline systems to conduct the experiments for frame-wise training criteria. From the IBM Watson team, in building the baselines Xiaodong Cui, Jia Cui, Bhuvana Rhamabhadran, and Brian Kingsbury were highly involved. The remaining setups, baseline systems, theoretical derivations, and experiments connected to scientific work are my individual contribution.
 - [Nußbaum-Thom & Schlüter⁺ 17] "Noisy Objective Functions based on the f -divergence": This work is completely part of my individual contribution. I am very thankful to my co-authors for discussing, correcting and proof-reading my paper.
- Bayes Decision Theory:
 - [Schlüter & Nußbaum-Thom⁺ 10] "On the Relationship Between Bayes Risk and Word Error Rate in ASR": Ralf Schlüter discovered the theory in this paper and helped

me with the experimental design. My individual contribution to this paper is the experimental setup, evaluation, and implementation of the developed methods.

- [Schlüter & Nußbaum-Thom⁺ 11] ”On the Relation of Bayes Risk, Word Error, and Word Posteriors in ASR”: Ralf Schlüter discovered the theory in this paper and helped me with the experimental design. My individual contribution to this paper is the experimental setup, evaluation and implementation of the developed methods.
- [Schlüter & Nußbaum-Thom⁺ 12] ”Does the Cost Function Matter in Bayes Decision Rule?”: Ralf Schlüter discovered the theory in this paper and helped me with the experimental design. My individual contribution to this paper is the experimental setup, evaluation and implementation of the developed methods.
- Deep Learning:
 - [Cui & Kingsbury⁺ 15] ”Multilingual Representations for Low Resource Speech Recognition and Keyword Search”: Jia Cui implemented all methods in the IBM speech recognition system and evaluated those in automatic speech recognition experiments. The other co-authors and I contributed baseline systems in several languages which were part of the BABEL evaluation campaign.
 - [Nußbaum-Thom & Cui⁺ 16a] ”Acoustic Modeling Using Bidirectional Gated Recurrent Convolutional Units”: All contributions in this paper are based on my individual work. I am very thankful to my co-authors for discussing, correcting and proof-reading my paper.
 - [Cui & Kingsbury⁺] ”Knowledge Distillation Across Ensembles of Multilingual Models for Low-Resource Languages”: Jia Cui implemented all methods in the IBM speech recognition system. She also evaluated these methods in automatic speech recognition experiments. I and the other co-authors contributed baseline systems in several languages which were part of the BABEL evaluation campaign.
- International Evaluation Campaigns:
 - [Nußbaum-Thom & Wiesler⁺ 10a] ”The RWTH 2009 QUAERO ASR Evaluation System for English and German”: This paper is a joint effort. My individual contribution in this paper is the development and evaluation of the German QUAERO system. Christian Plahl developed and trained the bottleneck features used in this paper. Simon Wiesler developed and evaluated the English QUAERO system. All authors helped in proof-reading this paper.
 - [Nußbaum-Thom & Beck⁺ 13] ”Compound Word Recombination for German LVCSR”: All contributions in this paper are based on my individual work. I am very thankful to my co-authors for discussing, correcting and proof-reading my paper.
 - [Lamel & Courcinous⁺ 11] ”Speech Recognition for Machine Translation in Quaero”: My individual contribution in this work is the development and evaluation of the German QUAERO system. This work is a joint effort by all authors.
 - [Sundermeyer & Nußbaum-Thom⁺ 11] ”The RWTH 2010 Quaero ASR Evaluation System for English, French, and German”: This paper is a joint effort. My individual contribution in this paper is the development and evaluation of the German QUAERO system. Christian Plahl developed and trained the bottleneck features used in this paper. Simon Wiesler developed and evaluated the English QUAERO system. Martin Sundermeyer developed and evaluated the French QUAERO system. All authors helped in proof-reading this paper.

- [Peitz & Mansour⁺ 12] "The RWTH Aachen Speech Recognition and Machine Translation System for IWSLT 2012": This paper is a joint effort. My individual contribution in this paper is the development and evaluation of all speech recognition systems in this paper. All authors helped in proof-reading this paper.
- [Shaik & Tüske⁺ 14] "RWTH LVCSR Systems for QUAERO and EU-Bridge: German, Polish, Spanish and Portuguese": This paper is a joint effort. My individual contribution in this paper is the development and evaluation of the German, Spanish, and Portuguese EU-bridge system. Zoltan Tuske developed and trained the bottleneck features used in this paper. Basha Shaik developed and evaluated the Polish EU-bridge speech recognition system.
- [Shaik & Tüske⁺ 15] "Improvements in RWTH LVCSR Evaluation Systems for Polish, Portuguese, English, Urdu, and Arabic": This paper is a joint effort. My individual contribution in this paper is the development and evaluation of the baseline systems for English, Portuguese. Zoltan Tuske developed and trained the bottleneck features used in this paper. Basha Shaik developed and evaluated all novel methods developed in this paper.
- Other Contributions:
 - [Wiesler & Heigold⁺ 10] "A Discriminative Splitting Criterion for Phonetic Decision Trees": My contribution in this paper is to setup the Wall Street Journal (WSJ) baseline speech recognition system. Simon Wiesler discovered, implemented and evaluated all novel methods developed in this paper.
 - [Heigold & Wiesler⁺ 10] "Discriminative HMMS, Log-Linear Models, and CRFS: What is the Difference?": Georg Heigold and Simon Wiesler discovered, implemented and evaluated all novel methods developed in this paper. My contribution in this paper is the setup of the wall street journal baseline speech recognition system. All authors helped in proof-reading this paper.
 - [Peitz & Mansour⁺ 12] "Spoken Language Translation using Automatically Transcribed Text in Training": Stephan Peitz discovered, implemented and evaluated all machine translation related methods in this paper. Simon Wiesler trained the baseline English automatic speech recognition system. I supported Stephan Peitz to evaluate the English speech recognition system on the IWSLT data.
 - [Tahir & Nußbaum-Thom⁺ 12] "Simultaneous Discriminative Training and Mixture Splitting of HMMs for Speech Recognition": Mohammad Tahir discovered, implemented and evaluated all novel splitting methods developed in this paper. I supported the author in setting up the discriminative training in this work. Specifically the log-MMI and the log-MPE discriminative training criteria in this work are based on my implementation. All authors helped in proof-reading this paper.

3. SCIENTIFIC GOALS

Bayes decision theory provides a mathematical framework for statistical decision problems to analyze decision making and learning of statistical models. Although this theory covers a wide range of topics, some aspects still need further investigation.

Error Bounds One aspect is the relation between the error, error bound, and the training criterion. The publications in [Vapnik 98, p.30] and [Ney 03, Guntuboyina 11] established such a connection for statistical decision problems based on the 0-1 loss. In summary, there it is shown that the *Kullback-Leibler* divergence bounds the error difference between the model-based and *Bayes* decision rule. The author of [Ney 03] also derives the cross-entropy/maximum-mutual-information criterion from the *Kullback-Leibler* divergence bound. A further investigation of this relation between the error difference, bounds, and training criteria perhaps leading to new bounds and criteria which are not only based on the 0-1 loss but also depend on more general losses is still missing and is an essential topic in this thesis.

The first goal of this thesis is to investigate novel bounds on the error difference between the *Bayes* and the model-based maximum probability decision rule. This investigation includes both: Decision problems based on the classification error difference as well as decision problems based on a more general loss like the *Levenshtein* loss. The latter case also reflects the mismatch situation of automatic speech recognition. Furthermore, we look at the implicit and explicit formulation of bounds (on the error difference) as well as their ability to derive discriminative training criteria.

Error-Bound-Based Training Criteria Automatic speech recognition is such a decision problem where the performance measure relies on a more general loss — the *Levenshtein* loss. At the same time, the *Bayes* decision rule based on the 0-1 loss determines the recognition result. By definition, this decision rule minimizes the sentence error rate, which is not guaranteed to minimize the word error rate. Therefore automatic speech recognition faces a fundamental mismatch of the loss used in the performance measure and the *Bayes* decision rule. Unfortunately, the evaluation of the exact *Bayes* decision rule — minimizing the posterior expected *Levenshtein* loss — is too time and memory consuming and only feasible as an approximation by a post-processing step to the maximum probability decision rule.

In practice, word-error-based training of the model tries to anticipate this mismatch. Even though there is no clear theoretical proof, the corresponding type of criterion is assumed to shift the decision boundaries of the model closer to the performance measure — the word error rate. MPE is such a criterion minimizing the posterior-expected *Levenshtein* loss of the spoken phoneme sequence, which in practice performs better than other discriminative training criteria. However, how are the theoretical implications of this training criterion for the mismatch in automatic speech recognition?

The second goal in this thesis is to investigate novel discriminative training criteria derived from suitable error bounds. This investigation includes both: Classification error bounds as well as

bounds on the error difference based on a more general loss like the *Levenshtein* loss. The first case is suitable to derive frame-wise discriminative training criteria for automatic speech recognition. We plan to formulate sequence-based training criteria for automatic speech recognition from the second case. Furthermore, the non-parametric solution of the model posterior learned with such a criterion should be investigated.

Application to Frame-Wise Neural Network Training for Automatic Speech Recognition The third goal in this thesis is to evaluate the frame-wise training criteria derived from error bounds on the classification error difference in practical experiments. In this evaluation frame-wise discriminative criteria are used to train acoustic neural networks models for automatic speech recognition.

Application to Sequence Training for Automatic Speech Recognition The fourth goal in this thesis is to evaluate the criteria derived from bounds on the error difference based on the *Levenshtein* loss in practical experiments. In this evaluation the training criteria are interpreted as sequence-based discriminative training of acoustic log-linear models for automatic speech recognition.

German QUAERO Evaluation Campaign The fifth goal in this thesis is to describe the German QUAERO Evaluation Campaign in which we contributed for six years top scoring German ASR systems.

4. ERROR BOUNDS

Previously the *Kullback-Leibler* divergence between the true and model distribution could be shown to relate to the classification error difference in the form of different upper bounds [Vapnik 98, Fedotov & Harremos⁺ 03, Ney 03, Cover & Thomas 06, Guntuboyina 11]. However, all these bounds lack specific properties. The upper bounds derived in [Vapnik 98, Fedotov & Harremos⁺ 03, Cover & Thomas 06, Guntuboyina 11], for example, are not tightly bound with classification error difference, which is one of the desired properties. We are interested in bounds similar to those in [Ney 03]. Nevertheless, the author in [Ney 03] only presents two types of bounds restricted to decision problems based on the 0-1 loss only. Our interest includes new classification error bounds and bounds for decision problems based on more general losses — like the *Levenshtein* loss. The investigation of such bounds is crucial for the mismatch in automatic speech recognition. It is also essential to us if the bounds can be formulated explicitly concerning the error difference. Similar to [Ney 03], explicit bounds are useful to derive discriminative training criteria.

Section 4 introduces an approach to establish novel bounds on the error difference. Section 5 establishes empirical training criteria from these error bounds. Our scientific contribution in this section splits into two parts. First, in Section 4.1, we focus on finding a derivation scheme for error bounds from decision problems based on the 0-1 loss. Second, we generalize this scheme in Section 4.2 to decision problems based on a more general loss.

A construct that compares two probability distributions is the *f-Divergence* — a generalization of the *Kullback-Leibler* divergence to compare two probability distributions. The *f-Divergence* is used in the following to derive error bounds. In Section 4.1, two analytical proofs of bounds on the classification error difference are presented based on the *f-Divergence*. The mathematical properties of the *f-Divergence* are helpful to derive this family of tight classification error bounds. We also discuss under which conditions the bound is tight with the classification error difference. Random simulations of the *Bayes* and model-based decision rule visually show the tightness of the new classification error bounds.

In Section 4.2, we focus on error bounds derived for decision problems based on a more general loss. In general, an accuracy defines this loss. We only consider decision problems with a positive posterior-expected accuracy. Based on this restriction, we can find error bounds using a more general loss. To derive this particular case of bounds, we revisit the classification error bounds from Section 4.1. We transform the decision problems based on a more general loss into an equivalent decision problem based on the 0-1 loss. Then we reuse the classification error bounds from Section 4.1 for the transformed decision problem.

4.1 Derivation of Classification Error Bounds

In this section, we develop bounds on the classification error difference. To derive these bounds, we introduce the *f-Divergence*. Intuitively the *f-Divergence* is a family of functions to compare probability distributions. Subsequently, we show that this relative measure between distributions

can be related to the classification error difference.

As introduced in Section 2.2.1, consider the true and model-based distribution $pr(x, c)$ and $q(x, c)$ for continuous observations $x \in \mathcal{X}$ and classes $c \in \mathcal{C}$. Now, the *f-Divergence* is defined to introduce the main theorem of this section.

Global f-Divergence: The definition of the *f-Divergence* assumes a convex function $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ with $f(1) = 0$. Then according to [Österreicher 02, Csiszár & Shields 04, Liese & Vajda 06], the global *f-Divergence* is defined by (4.1).

$$D_f(pr||q) := \int \sum_{c \in \mathcal{C}} q(x, c) f\left(\frac{pr(x, c)}{q(x, c)}\right) dx. \quad (4.1)$$

A very common *f-Divergence* is the *Kullback-Leibler* divergence, which is also called relative entropy.

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

The *Kullback-Leibler* divergence is also closely related to the perplexity difference between distribution pr and q .

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

The fundamental theorem of this work is formulated using the global *f-Divergence* theorem for error bounds in the following theorem.

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

$$2D_f(pr||q) \geq f(1 + \Delta) + f(1 - \Delta).$$

Equality is obtained with class-conditional probabilities shared between the model and the true distribution,

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

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

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

and

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

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

In the remaining part of this section, a partial proof of Theorem 1 is presented, which is continued and finished in two independent alternative proofs in Section 4.1.1 and Section 4.1.2. Properties that are applied multiple times in this proof are *Jensens* inequality and the Aggregation property.

Jensens inequality *Jensens inequality* states that for a convex function f and a random variable over $p(x)$ that the following inequality is fulfilled concerning the expectation.

$$f\left(\int xp(x) \, dx\right) \leq \int p(x) f(x) \, dx.$$

Aggregation property The aggregation property is an analog of the *f-Divergence* to the log-sum inequality of the *Kullback-Leibler* divergence. Using this, many of the properties of the *Kullback-Leibler* divergence $D_{\text{KL}}(p||q)$ extend to general *f-Divergences*, as shown next. Assume $a_1, \dots, a_I, b_1, \dots, b_I \geq 0$ and $a = \sum_{i=1}^I a_i, b = \sum_{i=1}^I b_i$ (for $a_i = b_i = 0$ the L'Hospital rule applies). Then due to convexity and *Jensens inequality*, the requirements are met for

$$\begin{aligned} \sum_{i=1}^I b_i f\left(\frac{a_i}{b_i}\right) &= \sum_{i=1}^I b \frac{b_i}{b} f\left(\frac{a \frac{a_i}{a}}{b \frac{b_i}{b}}\right) && \text{Jensens inequality: } E(f(\frac{a}{b}X)) \geq f(E(\frac{a}{b}X)) \text{ for convex } f \\ &\geq b f\left(\frac{a}{b} \sum_{i=1}^I \frac{b_i}{b} \frac{\frac{a_i}{a}}{\frac{b_i}{b}}\right) \\ &= b f\left(\frac{a}{b} \sum_{i=1}^I \frac{a_i}{a}\right) \\ &= b f\left(\frac{a}{b}\right). \end{aligned}$$

The aggregation property does also apply for continuous integrals instead of discrete sums. Assume $a(x), b(x) \geq 0$ for $x \in \mathcal{X}$ and $a = \int a(x) \, dx, b = \int b(x) \, dx$. Using the analogous proof to the discrete case then due to convexity and *Jensens inequality* the requirements are met for

$$\begin{aligned} \int b(x) f\left(\frac{a(x)}{b(x)}\right) \, dx &= \int b \frac{b(x)}{b} f\left(\frac{a \frac{a(x)}{a}}{b \frac{b(x)}{b}}\right) \, dx \\ &\geq b f\left(\frac{a}{b} \int \frac{b(x)}{b} \frac{\frac{a(x)}{a}}{\frac{b(x)}{b}} \, dx\right) && \text{Jensens inequality: } E(f(\frac{a}{b}X)) \geq f(E(\frac{a}{b}X)) \text{ for convex } f \\ &= b f\left(\frac{a}{b} \int \frac{a(x)}{a} \, dx\right) \\ &= b f\left(\frac{a}{b}\right). \end{aligned}$$

The true and model-based distributions $pr(x, c)$ and $q(x, c)$ are aggregated into $\overline{pr}(c_{0-1})$, $\overline{pr}(c_{0-1}^q)$, $\overline{q}(c_{0-1})$, and $\overline{q}(c_{0-1}^q)$ in order to define a two-class distribution in the step following to this definition.

$$\begin{aligned}\overline{pr}(c_{0-1}) &= \int pr(x, c_{0-1}(x)) \, dx, & \overline{pr}(c_{0-1}^q) &= \int pr(x, c_{0-1}^q(x)) \, dx, \\ \overline{q}(c_{0-1}^{pr}) &= \int q(x, c_{0-1}(x)) \, dx, & \text{and} & \overline{q}(c_{0-1}^q) &= \int q(x, c_{0-1}^q(x)) \, dx.\end{aligned}$$

Consider the definitions $R = \{c_{0-1}, c_{0-1}^q\}$ and $R(x) = \{c_{0-1}(x), c_{0-1}^q(x)\}$ for $x \in \mathcal{X}$. In summary the repeated application of the aggregation property leads to the following inequalities bounding a two-class f -Divergence.

$$\begin{aligned}D_f(pr||q) &= \int \sum_{c \in \mathcal{C}} q(x, c) f\left(\frac{pr(x, c)}{q(x, c)}\right) \, dx \\ &= \int \sum_{c \in R(x)} q(x, c) f\left(\frac{pr(x, c)}{q(x, c)}\right) \, dx + \int \sum_{c \in \mathcal{C} \setminus R(x)} q(x, c) f\left(\frac{pr(x, c)}{q(x, c)}\right) \, dx \\ &\geq \int \sum_{c \in R(x)} q(x, c) f\left(\frac{pr(x, c)}{q(x, c)}\right) \, dx \\ &\quad + \int \left(\sum_{c \in \mathcal{C} \setminus R(x)} q(x, c) \right) f\left(\frac{\sum_{c \in \mathcal{C} \setminus R(x)} pr(x, c)}{\sum_{c \in \mathcal{C} \setminus R(x)} q(x, c)}\right) \, dx && \text{using Aggregation} \\ &\geq \left[\sum_{r \in R} \left(\int q(x, r(x)) \, dx \right) f\left(\frac{\int pr(x, r(x)) \, dx}{\int q(x, r(x)) \, dx}\right) \, dx \right] && \text{using Aggregation} \\ &\quad + \left(\int \sum_{c \in \mathcal{C} \setminus R(x)} q(x, c) \, dx \right) f\left(\frac{\int \sum_{c \in \mathcal{C} \setminus R(x)} pr(x, c) \, dx}{\int \sum_{c \in \mathcal{C} \setminus R(x)} q(x, c) \, dx}\right) \, dx && \text{using Aggregation} \\ &= \sum_{r \in R} \overline{q}(r) f\left(\frac{\overline{pr}(r)}{\overline{q}(r)}\right) + \left(1 - \sum_{r \in R} \overline{q}(r)\right) f\left(\frac{1 - \sum_{r \in R} \overline{pr}(r)}{1 - \sum_{r \in R} \overline{q}(r)}\right) \\ &= \sum_{r \in R} \overline{q}(r) f\left(\frac{\frac{\overline{pr}(r)}{2}}{\frac{\overline{q}(r)}{2}}\right) + 2 \frac{\left(1 - \sum_{r \in R} \overline{q}(r)\right)}{2} f\left(\frac{\frac{1}{2} \left(1 - \sum_{r \in R} \overline{pr}(r)\right)}{\frac{1}{2} \left(1 - \sum_{r \in R} \overline{q}(r)\right)}\right). \tag{4.2}\end{aligned}$$

After extending (4.2), the summation terms in (4.3) are aggregated into (4.4).

$$\begin{aligned}
 &= \bar{q}(c_{0-1}) f\left(\frac{\bar{p}r(c_{0-1})}{\bar{q}(c_{0-1})}\right) + \frac{\left(1 - \sum_{r \in R} \bar{q}(r)\right)}{2} f\left(\frac{\frac{1}{2} \left(1 - \sum_{r \in R} \bar{p}r(r)\right)}{\frac{1}{2} \left(1 - \sum_{r \in R} \bar{q}(r)\right)}\right) \\
 &\quad + \bar{q}(c_{0-1}^q) f\left(\frac{\bar{q}(c_{0-1}^q)}{\bar{q}(c_{0-1}^q)}\right) + \frac{\left(1 - \sum_{r \in R} \bar{q}(r)\right)}{2} f\left(\frac{\frac{1}{2} \left(1 - \sum_{r \in R} \bar{p}r(r)\right)}{\frac{1}{2} \left(1 - \sum_{r \in R} \bar{q}(r)\right)}\right)
 \end{aligned} \tag{4.3}$$

$$\begin{aligned}
 &\geq \left(\bar{q}(c_{0-1}) + \frac{1 - \sum_{r \in R} \bar{q}(r)}{2}\right) f\left(\frac{\bar{p}r(c_{0-1}) + \frac{1}{2} \left(1 - \sum_{r \in R} \bar{p}r(r)\right)}{\bar{q}(c_{0-1}) + \frac{1}{2} \left(1 - \sum_{r \in R} \bar{q}(r)\right)}\right) \\
 &\quad + \left(\bar{q}(c_{0-1}^q) + \frac{1 - \sum_{r \in R} \bar{q}(r)}{2}\right) f\left(\frac{\bar{p}r(c_{0-1}^q) + \frac{1}{2} \left(1 - \sum_{r \in R} \bar{p}r(r)\right)}{\bar{q}(c_{0-1}^q) + \frac{1}{2} \left(1 - \sum_{r \in R} \bar{q}(r)\right)}\right).
 \end{aligned} \tag{4.4}$$

The values of this two-class distribution λ , $1 - \lambda$ and β , $1 - \beta$ are defined as follow,

$$\lambda = \frac{1}{2} + \frac{1}{2} \bar{p}r(c_{0-1}) - \frac{1}{2} \bar{p}r(c_{0-1}^q) = \frac{1}{2} + \frac{1}{2} \Delta, \quad 1 - \lambda = \frac{1}{2} - \frac{1}{2} \Delta, \tag{4.5}$$

$$\beta = \frac{1}{2} + \frac{1}{2} \bar{q}(c_{0-1}) - \frac{1}{2} \bar{q}(c_{0-1}^q) = \frac{1}{2} - \frac{1}{2} \Delta^q, \quad \text{and} \quad 1 - \beta = \frac{1}{2} + \frac{1}{2} \Delta^q \tag{4.6}$$

using the following definition of the model-based error difference (which is always non-negative):

$$\begin{aligned}
 \Delta^q &:= \int [q(x, c_{0-1}^q(x)) - q(x, c_{0-1}(x))] \, dx \\
 &= \int \left[\max_{c \in \mathcal{C}} \{q(x, c)\} - q(x, c_{0-1}(x)) \right] \, dx \\
 &\geq 0.
 \end{aligned}$$

Then (4.4) can equivalently be transformed to (4.7) and (4.8).

$$\begin{aligned}
 &= \left(\frac{1}{2} + \frac{1}{2} \bar{q}(c_{0-1}) - \frac{1}{2} \bar{q}(c_{0-1}^q)\right) f\left(\frac{\frac{1}{2} + \frac{1}{2} \bar{p}r(c_{0-1}) - \frac{1}{2} \bar{p}r(c_{0-1}^q)}{\frac{1}{2} + \frac{1}{2} \bar{q}(c_{0-1}) - \frac{1}{2} \bar{q}(c_{0-1}^q)}\right) \\
 &\quad + \left(\frac{1}{2} + \frac{1}{2} \bar{q}(c_{0-1}^q) - \frac{1}{2} \bar{q}(c_{0-1})\right) f\left(\frac{\frac{1}{2} + \frac{1}{2} \bar{p}r(c_{0-1}^q) - \frac{1}{2} \bar{p}r(c_{0-1})}{\frac{1}{2} + \frac{1}{2} \bar{q}(c_{0-1}^q) - \frac{1}{2} \bar{q}(c_{0-1})}\right)
 \end{aligned} \tag{4.7}$$

$$\begin{aligned}
 &= \left(\frac{1}{2} - \frac{1}{2} \Delta^q\right) f\left(\frac{\frac{1}{2} + \frac{1}{2} \Delta}{\frac{1}{2} - \frac{1}{2} \Delta^q}\right) + \left(\frac{1}{2} + \frac{1}{2} \Delta^q\right) f\left(\frac{\frac{1}{2} - \frac{1}{2} \Delta}{\frac{1}{2} + \frac{1}{2} \Delta^q}\right) \\
 &= \beta f\left(\frac{\lambda}{\beta}\right) + (1 - \beta) f\left(\frac{1 - \lambda}{1 - \beta}\right).
 \end{aligned} \tag{4.8}$$

■

In Section 4.1.1 and Section 4.1.2, the partial proof presented in this section and resulting in (4.8) continues with two analytical proofs of the implicit *f-Divergence* bound of Theorem 1. The first proof in Section 4.1.1 shows a direct derivation while the second proof in Section 4.1.2 uses a theorem from statistical information theory to prove the bound. Then a proof of explicit bounds on the squared error difference follows for a specific type of *f-Divergences*. These explicit bounds are extended to global bounds and used to derive discriminative training criteria in Section 5 for classification problems based on the 0-1 loss.

4.1.1 Implicit Classification Error Bounds

Due to (4.5) and (4.6) in Section 4.1, there is a direct relation between λ and β to the classification error difference Δ and the model-based classification error difference Δ^q . So the requirements are met for:

$$\begin{aligned} 2\lambda - 1 &= \lambda - (1 - \lambda) = \Delta, \\ 2\beta - 1 &= \beta - (1 - \beta) = -\Delta^q. \end{aligned}$$

With $\Delta \geq 0$ and $\Delta^q \geq 0$, it follows that: $\beta \leq \frac{1}{2} \leq \lambda$. Consider 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.

$$\begin{aligned} \frac{1}{2} (f(2\lambda) + f(2[1 - \lambda])) &= \frac{1}{2} \left(f\left(\frac{\lambda}{1/2}\right) + f\left(\frac{1 - \lambda}{1/2}\right) \right) \\ &\geq \left(\frac{1}{2} + \frac{1}{2} \right) \cdot f\left(\frac{\lambda + 1 - \lambda}{\frac{1}{2} + \frac{1}{2}} \right) \\ &= f(1) = 0. \end{aligned}$$

Then, starting from (4.8), the following simplification is carried out.

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

This finishes the proof of Theorem 1. The next Section gives an alternative proof of Theorem 1 using *Guntuboyinas* Theorem.

4.1.2 Implicit Classification Error Bounds and Guntuboyinas Theorem

The next theorem is a result of theoretical statistics [Guntuboyina 11, p.3, Corollary II.3] adopted to our problem formulation to prove Theorem 1.

Theorem 2 Consider the variational distance between probabilities p and $\bar{p}$

$$V = \sum_{c \in C} |p(c) - \bar{p}(c)|,$$

then the following relation exists [Guntuboyina 11, p.3, Corollary II.3].

$$\inf_{\pi} \{D_f(p||\pi) + D_f(\bar{p}||\pi)\} \geq f\left(1 - \frac{V}{2}\right) + f\left(1 + \frac{V}{2}\right).$$

Starting from (4.8) we define a two-class subspace based on two abstract classes $c_1, c_2, c_1 \neq c_2$ (without loss of generality) with distributions p and $\bar{p}$.

$$\begin{aligned} p(c_1) &= \lambda, & p(c_2) &= 1 - \lambda \\ \bar{p}(c_1) &= 1 - \lambda, & \bar{p}(c_2) &= \lambda. \end{aligned}$$

Guntuboyinas theorem allows proving Theorem 1 applied to the two-class subspace distribution p and $\bar{p}$.

$$\begin{aligned} 2D_f(p||q) &= 2 \int \sum_{c \in C} q(x, c) f\left(\frac{pr(x, c)}{q(x, c)}\right) dx \\ &\geq 2\beta f\left(\frac{\lambda}{\beta}\right) + 2(1 - \beta)f\left(\frac{1 - \lambda}{1 - \beta}\right) && \text{starting from (4.8)} \\ &= \beta f\left(\frac{p(c_1)}{\beta}\right) + (1 - \beta)f\left(\frac{p(c_2)}{1 - \beta}\right) + \beta f\left(\frac{\bar{p}(c_2)}{\beta}\right) + (1 - \beta)f\left(\frac{\bar{p}(c_1)}{1 - \beta}\right) \\ &\geq \min_{\pi} \left\{ \pi f\left(\frac{p(c_1)}{\pi}\right) + (1 - \pi)f\left(\frac{p(c_2)}{1 - \pi}\right) + \pi f\left(\frac{\bar{p}(c_1)}{\pi}\right) + (1 - \pi)f\left(\frac{\bar{p}(c_2)}{1 - \pi}\right) \right\} \\ &&& \text{applying Guntuboyina} \\ &\geq f\left(1 - \frac{V}{2}\right) + f\left(1 + \frac{V}{2}\right) \\ &= f\left(1 - \sum_{c \in \{c_1, c_2\}} \frac{|p(c) - \bar{p}(c)|}{2}\right) + f\left(1 + \sum_{c \in \{c_1, c_2\}} \frac{|p(c) - \bar{p}(c)|}{2}\right) \\ &= f\left(1 - \frac{|p(c_1) - \bar{p}(c_1)|}{2} - \frac{|p(c_2) - \bar{p}(c_2)|}{2}\right) + f\left(1 + \frac{|p(c_1) - \bar{p}(c_1)|}{2} + \frac{|p(c_2) - \bar{p}(c_2)|}{2}\right) \\ &= f(1 - |2\lambda - 1|) + f(1 + |2\lambda - 1|) \\ &= f(1 - |\Delta - 1 + 1|) + f(1 + |\Delta - 1 + 1|) \\ &= f(1 - \Delta) + f(1 + \Delta). \end{aligned} \quad \blacksquare$$

This proves Theorem 1. In the next section, some examples of error bounds based on the f -Divergence are presented.

4.1.3 Implicit Classification Error Bounds with Closed Form Solution

The following f -Divergences are often used in statistics and resolve to a closed-form classification error bound. They are called reversed *Kullback-Leibler*, *Chi-squared*, *Neyman-Pearson*, and *Vajda* divergence. More possible f -Divergences are presented in [Österreicher 02].

Reversed Kullback-Leibler The reversed *Kullback-Leibler* divergence is obtained with setting $f_{\text{RKL}}(u) = -\log u$:

$$\Rightarrow \Delta \leq \sqrt{1 - \exp(-2D_{\text{KL}}(q||pr))}. \quad (4.9)$$

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

$$\begin{aligned} D_{f_{\chi^2}}(pr||q) &= D_{\chi^2}(q||pr) \\ &\geq \frac{1}{2}((1 + \Delta)^2 - 1) + ((1 - \Delta)^2 - 1)) \\ &= \Delta^2 \\ \Rightarrow \Delta &\leq \sqrt{D_{\chi^2}(q||pr)} = \sqrt{\int \sum_{c \in \mathcal{C}} \frac{pr^2(x, c)}{q(x, c)} dx - 1}. \end{aligned} \quad (4.10)$$

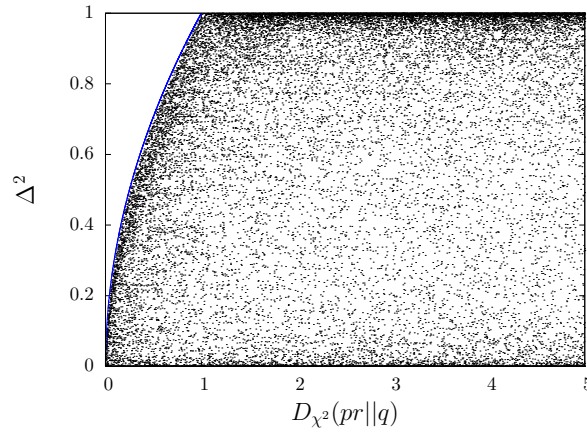


Figure 4.1: *Chi-Squared Bound*.

Neyman-Pearson Assume the *f-Divergence* with $f_{\text{NP}}(u) = \frac{(u-1)^2}{u}$. The associated bound is then:

$$\begin{aligned} D_{f_{\text{NP}}}(pr||q) &\geq \frac{1}{2} \left(\frac{(1 - \Delta - 1)^2}{1 - \Delta} + \frac{(1 + \Delta - 1)^2}{1 + \Delta} \right) \\ &= \frac{1}{2} \left(\frac{\Delta^2(1 + \Delta) + \Delta^2(1 - \Delta)}{1 - \Delta^2} \right) \\ &= \frac{\Delta^2}{1 - \Delta^2} \\ \Rightarrow \Delta &\leq \sqrt{\frac{D_{f_{\text{NP}}}(pr||q)}{D_{f_{\text{NP}}}(pr||q) + 1}} \\ \Delta &\leq \sqrt{\frac{D_{f_{\text{NP}}}(pr||q)}{1 + D_{f_{\text{NP}}}(pr||q)}}. \end{aligned} \quad (4.11)$$

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

$$\begin{aligned}
 D_{f_V}(pr||q) &\geq \frac{1}{2}(|1 - \Delta - 1|^\alpha + |1 + \Delta - 1|^\alpha) \\
 &= \Delta^\alpha \\
 &\Leftrightarrow \Delta \leq D_{f_V}^{\frac{1}{\alpha}}(pr||q) \\
 \Delta &\leq D_{f_V}^{\frac{1}{\alpha}}(pr||q).
 \end{aligned} \tag{4.12}$$

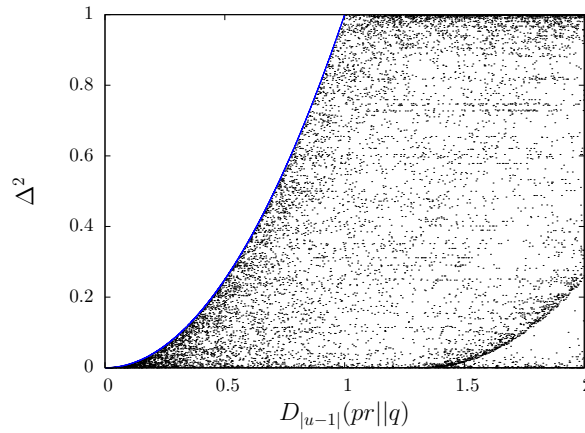


Figure 4.2: Vajda Bound for $f_V(u) = |u - 1|$.

Unfortunately, the bound derived from the *Kullback-Leibler* does not result in a closed-form expression in terms of Δ , in which case numerical approximations can be used. For $\alpha = 1$, the Vajda divergence results in the total variational distance and is interestingly a bound on the classification error difference itself. The bound derived from the reversed *Kullback-Leibler* divergence takes values in $[0, 1]$ and might give a reasonable upper bound on the classification error difference. On the other hand, the bounds derived by the Chi-Squared and Vajda divergences are not limited to the classification error difference domain, which makes them useless for those cases where the trivial bound $\Delta \leq 1$ is tighter.

The simulations in Figure 4.1 through 4.2 were performed by generating a shared class conditional and two (true and model) class prior distributions, where three classes and two observations were assumed. The same tendency was confirmed in simulations using more classes and observations. In general, several million distributions were generated, and filtered afterwards to achieve better visualization.

4.1.4 Explicit Classification Error Bounds

In this section, we investigate explicit local bounds based on the global bounds found in Section 4.1 to support a subsequent derivation of training criteria. To that purpose, the local classification error bounds are considered by reinterpreting the *f-Divergence* and Theorem 1 for class posterior probabilities $pr(c|x)$ and $q(c|x)$ instead of the joint distributions $pr(x, c)$ and $q(x, c)$.

Local f-Divergence Let $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ be a convex function with $f(1) = 0$ then the local f -Divergence is defined by:

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

Reinterpreting Theorem 1 in terms of the local f -Divergence results in:

$$2D_f^x(pr||q) \geq f(1 + \Delta(x)) + f(1 - \Delta(x)). \quad (4.13)$$

Without loss of generality, the proof for the local f -Divergence classification error bounds is analogous to the case of the global f -Divergence classification error bounds.

Taylor's theorem Let $k \in \mathbb{N}$ and the function $f : \mathbb{R} \rightarrow \mathbb{R}$ be k times differentiable at point $y_0 \in \mathbb{R}$. Then, for every $y \geq y_0$ 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 $f(y) = \sum_{n=0}^{k-1} \frac{f^{(n)}(y_0)(y-y_0)^n}{n!} + R_k(y)$.

Theorem 3 If $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ is convex, $f(1) = 0$, f''' exists, and is monotonically increasing, then the following bound on the global classification error difference is valid.

$$f''(1)\Delta^2(x) \leq 2D_f^x(pr||q).$$

Proof We start from (4.13). Locally in $y_0 = 1$, the application of *Taylor's Theorem* results in:

$$\begin{aligned} 2D_f^x(pr||q) &\geq f(1 + \Delta(x)) + f(1 - \Delta(x)) \\ &= f(1) + f'(1)\Delta(x) + \frac{f''(1)\Delta^2(x)}{2!} + \frac{f'''(\mu_{1+\Delta(x)})\Delta^3(x)}{3!} \\ &\quad + f(1) - f'(1)\Delta(x) + \frac{f''(1)\Delta^2(x)}{2!} - \frac{f'''(\mu_{1-\Delta(x)})\Delta^3(x)}{3!} \\ &= \underbrace{2f(1)}_{=0} + f''(1)\Delta^2(x) + \frac{\Delta^3(x)}{3!}(f'''(\mu_{1+\Delta(x)}) - f'''(\mu_{1-\Delta(x)})) \\ &\geq f''(1)\Delta^2(x) + \frac{\Delta^3(x)}{3!} \left(\underbrace{\min_{a \in [1, 1+\Delta(x)]} f'''(a)}_{\geq f'''(1)} + \underbrace{\min_{b \in [1-\Delta(x), 1]} -f'''(b)}_{\geq -f'''(1)} \right) \\ &\geq f''(1)\Delta^2(x) + \frac{\Delta^3(x)}{3!}(f'''(1) - f'''(1)) \\ &= \underbrace{f''(1)\Delta^2(x)}_{\geq 0}. \end{aligned}$$

This leads to the bound on the squared local classification error difference.

$$\Delta^2(x) \leq \frac{2}{f''(1)} D_f^x(pr||q). \quad \blacksquare$$

Global Explicit Error Bound The global classification error bound is obtained by integration of all observations and the application of Jensens inequality.

$$\begin{aligned} &\frac{2}{f''(1)} \int pr(x) D_f^x(pr||q) \, dx \\ &\geq \int pr(x) \Delta^2(x) \, dx \quad \text{Jensens inequality: } E(f(X)) \geq f(E(X)) \text{ for convex } f \\ &\geq \Delta^2. \end{aligned} \quad (4.14)$$

Tightness: The derived classification error bound is tightly bounded with the squared error difference only if 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 [Ney 03].

4.2 Derivation of Error Bounds for General Loss Functions

In this section, we investigate bounds based on a more general loss, e.g., the *Levenshtein* loss. These bounds are relevant to the mismatch described in Section 2.3.1 for automatic speech recognition and are the basis for deriving new error-based sequence criteria. These error bounds are derivable for decision problems based on a more general loss by extending the classification error bounds presented in Section 4.1. To this end, we reduce decision problems based on a more general loss to an equivalent decision problem based on the 0-1 loss. The local error is transformable into a class posterior, which maintains the *Bayes* decision rule of the decision problem using the more general loss. As a consequence, the classification error bound based on the 0-1 loss is reusable for deriving error bounds based on the more general loss.

Section 4.2.1 revisits *Bayes* decision theory for decision problems based on a general loss. These definitions are used in Section 4.2.2 to describe the transformation of the local error to a class posterior probability. With the relation to classification error bounds in Section 4.2.3, implicit local error bounds are derivable for this class posterior, which in Section 4.2.4 results in a closed-form solution.

4.2.1 Bayes Decision Theory for General Loss Functions

Consider statistical decision problems based on a general loss of the form:

$$\mathcal{L} : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}_0^+.$$

Terms and definitions introduced in Section 2.2 for *Bayes* decision theory for decision problems based on the 0-1 loss are now extended to general loss functions. For a decision rule $r : \mathcal{X} \rightarrow \mathcal{C}$, the local and global error are the expected posterior and joint loss concerning the true distribution.

$$\begin{aligned} E_{\mathcal{L}}(c|x) &:= \sum_{k \in \mathcal{C}} pr(k|x) \mathcal{L}(k, c), & \text{“local error of hypothesis } c\text{”} \\ E_{\mathcal{L}}(r) &:= \int pr(x) E_{\mathcal{L}}(r(x)|x) dx. & \text{“global error of decision } r\text{”} \end{aligned}$$

The *Bayes* decision rule minimizes the local error over all classes.

$$c_{\mathcal{L}}(x) := \operatorname{argmin}_{c \in \mathcal{C}} \{E_{\mathcal{L}}(c|x)\}. \quad (4.15)$$

In the following, we focus on the relation between the *Bayes* decision rule and the model-based maximum probability decision rule based on a posterior $q(c|x)$ defined in (2.1).

$$c_{0-1}^q(x) := \operatorname{argmax}_{c \in \mathcal{C}} \{q(c|x)\}.$$

Therefore as a quality measure, we look at the error difference of the model-based and the *Bayes* decision rule. We distinguish between the local error for an observation x and the global error integrated over all observations considering a fixed model-based decision rule c_{0-1}^q .

$$\begin{aligned} \Delta_{\mathcal{L}}(x) &:= E_{\mathcal{L}}(c_{0-1}^q(x)|x) - E_{\mathcal{L}}(c_{\mathcal{L}}(x)|x), & \text{“local error difference”} \\ \Delta_{\mathcal{L}} &:= \int pr(x) \Delta_{\mathcal{L}}(x) dx. & \text{“global error difference”} \end{aligned}$$

In the next section, the considered statistical decision problems are restricted to a specific type of loss. Then the local error using this type of loss is reduced to a class posterior.

4.2.2 From the Local Error to Class Posteriors

At first sight, the classification error difference and error difference based on a general loss have no connection. However, we will show that the opposite is the case. By considering a specific type of loss – the classification error difference and error difference is shown to be closely connected. Classification error bounds are reusable for deriving error bounds for decision problems based on this specific type of loss – different from the 0-1 loss.

In the following, consider statistical decision problems where the loss is defineable from an accuracy function $\mathcal{A} : \mathcal{C} \times \mathcal{C} \rightarrow \mathbb{R}$.

$$\mathcal{L}(k, c) := \mathcal{A}(k, k) - \mathcal{A}(k, c).$$

Besides, the accuracy function is needed to meet certain requirements important for the derivations in subsequent proofs. The length of a class $c \in \mathcal{C}$ is defined by

$$\mathcal{A}(c, c) \geq 1,$$

and is required to be larger or equal to one in the identity case. For example, if the classes are word sequences the length of a class is the number of words in the word sequence. The expected length is required to converge to

$$\sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, k) < \infty.$$

Finally, the local reward, defined by the posterior-expected accuracy, is required to be positive.

$$R_{\mathcal{A}}(c|x) := \sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, c) \geq 0. \quad (4.16)$$

The local reward is the corresponding opposite concept to the local error considering the maximization of the posterior-expected accuracy instead of the minimization of the local error. The local error and reward are connected by:

$$\begin{aligned} E_{\mathcal{L}}(c|x) &= \sum_{k \in \mathcal{C}} pr(k|x) \mathcal{L}(k, c) \\ &= \sum_{k \in \mathcal{C}} pr(k|x) [\mathcal{A}(k, k) - \mathcal{A}(k, c)] \\ &= \left[\sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, k) \right] - \left[\sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, c) \right] \\ &= \left[\sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, k) \right] - R_{\mathcal{A}}(c|x). \end{aligned} \quad (4.17)$$

Another essential property of the reward is that the *Bayes* decision rule is reformulated from the minimization of the local error into the maximization of the local reward.

$$\begin{aligned} c_{\mathcal{L}}(x) &= \operatorname{argmin}_{c \in \mathcal{C}} \{E_{\mathcal{L}}(c|x)\} \\ &= \operatorname{argmin}_{c \in \mathcal{C}} \left\{ \left[\sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, k) \right] - R_{\mathcal{A}}(c|x) \right\} \quad \text{according to (4.17)} \\ &= \operatorname{argmin}_{c \in \mathcal{C}} \{-R_{\mathcal{A}}(c|x)\} \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \{R_{\mathcal{A}}(c|x)\}. \end{aligned}$$

Consequently, the *Bayes* decision rule achieves the maximum local reward among all classes $R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)$. Also, the local error difference can be reformulated in terms of the local reward.

$$\begin{aligned}
 \Delta_{\mathcal{L}}(x) &= E_{\mathcal{L}}(c_{0-1}^q(x)|x) - E_{\mathcal{L}}(c_{\mathcal{L}}(x)|x) \\
 &= \left[\sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, k) \right] - R_{\mathcal{A}}(c_{0-1}^q(x)|x) - \left[\left[\sum_{k \in \mathcal{C}} pr(k|x) \mathcal{A}(k, k) \right] - R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x) \right] \\
 &\quad \text{according to (4.17)} \\
 &= R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x) - R_{\mathcal{A}}(c_{0-1}^q(x)|x). \tag{4.18}
 \end{aligned}$$

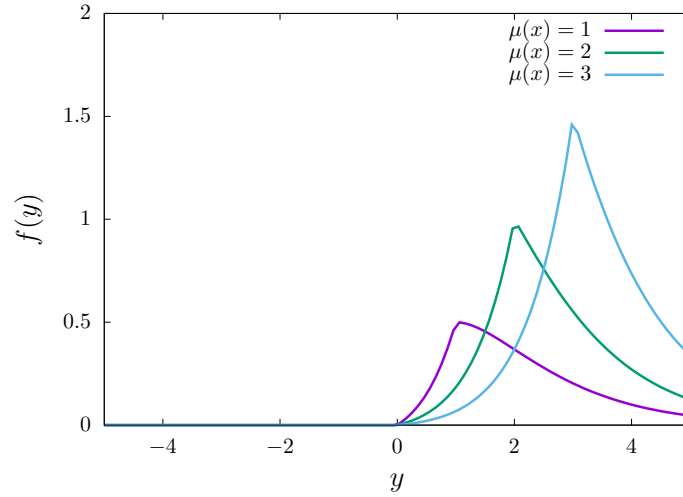


Figure 4.3: Example of the distribution $\phi(y|x)$ for mean $\mu(x) = 1, 2, 4$.

Subsequently, we show that the local error based on the specified type of loss is reducible to a posterior based on the 0-1 loss. This posterior maintains the *Bayes* decision rule based on the more general loss. The posterior is defined from the function $\phi(y|x)$ related to the *Laplace* distribution with $\mu(x) = R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)$ and $\sigma = 1$.

$$\begin{aligned}
 \phi(y|x) &= \frac{y}{2\sigma} \exp \left(- \left| \frac{y - \mu(x)}{\sigma} \right| \right) \\
 &= \frac{y}{2} \exp (-|y - R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)|). \tag{4.19}
 \end{aligned}$$

It can easily be shown that the global maximum of $\phi(y|x)$ for each observation x is attained at $R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)$.

$$\operatorname{argmax}_{y \in \mathbb{R}} \{ \phi(y|x) \} = R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x). \tag{4.20}$$

From the lower bound (4.16) and the upper bound (4.20) with $c \in \mathcal{C}$, the relation can be derived

$$0 \leq R_{\mathcal{A}}(c|x) \leq R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x) \Rightarrow 0 \leq \phi(R_{\mathcal{A}}(c|x)|x) \leq \phi(R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)|x). \tag{4.21}$$

Another interesting property of $\phi(y|x)$ is derived from the fact that

$$\frac{1}{2\sigma} \exp \left(- \left| \frac{y - \mu(x)}{\sigma} \right| \right) \tag{4.22}$$

is a *Laplace* distribution. This is the reason why the expected value of this distribution carries over to the normalization of $\phi(y|x)$.

$$\int \frac{y}{2\sigma} \exp\left(-\left|\frac{y - \mu(x)}{\sigma}\right|\right) dy = \mu(x). \quad (4.23)$$

Since the local reward $R_{\mathcal{A}}(c|x)$ for each class c is assumed to be positive in (4.16) the true reward posterior $\Pr_{\mathcal{A}}(c|x)$ is reducible by

$$\Pr_{\mathcal{A}}(c|x) = \frac{\phi(R_{\mathcal{A}}(c|x)|x)}{Z(x)} \quad \text{with the re-normalization} \quad Z(x) = \sum_{k \in \mathcal{C}} \phi(R_{\mathcal{A}}(k|x)|x). \quad (4.24)$$

Please bare in mind that the true reward posterior $\Pr_{\mathcal{A}}(c|x)$ is different from the true posterior $pr(c|x)$ and that those two posteriors should not be confused. Due to the maximality of the *Bayes* reward, there is no value attained larger than $y = R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)$. Therefore the maximum posterior is attained at $R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)$.

$$\begin{aligned} c_{\mathcal{L}}(x) &= \operatorname{argmin}_{c \in \mathcal{C}} \{E_{\mathcal{L}}(c|x)\} \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \{R_{\mathcal{A}}(c|x)\} \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \{\phi(R_{\mathcal{A}}(c|x)|x)\} \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \left\{ \frac{\phi(R_{\mathcal{A}}(c|x)|x)}{Z(x)} \right\} \\ &= \operatorname{argmax}_{c \in \mathcal{C}} \{\Pr_{\mathcal{A}}(c|x)\}. \end{aligned}$$

This proof reduces the *Bayes* decision rule of a decision problem based on the considered type of loss to a decision problem using the simple 0-1 loss decision rule. In other words, the decision problem fulfilling these requirements is equivalent to a decision problem using the 0-1 loss. We conclude that the classification error bounds from Section 4.1 is reusable to derive error bounds for the specified type of loss.

4.2.3 Implicit Local Error Bounds

As outlined in Section 4.2.2, the error of a statistical decision problem based on a general loss is reducible to a decision problem based on the 0-1 loss using a posterior probability that conserves the *Bayes* decision rule. The decision problem using a general loss has been reduced to an equivalent problem using a simple 0-1 loss. For this equivalent decision problem based on the 0-1 loss, the bounds and training criteria derived in Section 4.1 remain valid.

Now, the local error bounds are considered by reinterpreting the *f-Divergence* and Theorem 1 for class posterior probabilities $\Pr_{\mathcal{A}}(c|x)$ and $q(c|x)$. Theorem 1 initially was formulated for joint distributions of observations and classes. However, as outlined in Section 4.1.4, the theorem can be reinterpreted for class posterior probabilities and the local error of a decision problem based on the 0-1 loss. The local error difference concerning posterior $\Pr_{\mathcal{A}}$ and the model-based decision rule $c_{0,1}^q$ is defined by:

$$\Delta_{0,1}^{\Pr_{\mathcal{A}}}(x) := \Pr_{\mathcal{A}}(c_{\mathcal{L}}(x)|x) - \Pr_{\mathcal{A}}(c_{0,1}^q(x)|x). \quad (4.25)$$

Since the error probability $\Pr_{\mathcal{A}}(c|x)$ is also a class posterior probability, this theorem applies here as well. Therefore the following implicit local error bound can be derived from (4.13).

$$2D_f^x(\Pr_{\mathcal{A}}||q) \geq f(1 + \Delta_{0,1}^{\Pr_{\mathcal{A}}}(x)) + f(1 - \Delta_{0,1}^{\Pr_{\mathcal{A}}}(x)). \quad (4.26)$$

Next, a discussion of implicit local error bounds with a closed-form solution and their extension to global error bounds follows. The closed-form solutions originate from the classification error bounds from Section 4.1.3.

4.2.4 Implicit Error Bounds with Closed Form Solutions

Consider the implicit classification error bounds from Section 4.1.3 for the reversed *Kullback-Leibler*, Chi-squared, Neyman-Pearson, and Vajda divergence. As argued earlier, the derivation of global classification error bounds for the joint distribution also implies local error bounds using the class posteriors without loss of generality. Therefore, the closed-form error bounds from Section 4.1.3 are also valid for $\Pr_{\mathcal{A}}(c|x)$ and the following bounds are derived.

Reversed Kullback-Leibler By setting $f_{\text{RKL}}(u) = -\log u$ the reversed *Kullback-Leibler* divergence is obtained. The associated bound becomes then:

$$\Delta_{0-1}^{\Pr_{\mathcal{A}}}(x) \leq \sqrt{1 - \exp(-2D_{f_{\text{RKL}}}^x(\Pr_{\mathcal{A}}||q))}. \quad (4.27)$$

Chi-Squared By setting $f_{\chi^2}(u) = u^2 - 1$, the distance D_{χ^2} is obtained. The associated bound becomes then:

$$\Delta_{0-1}^{\Pr_{\mathcal{A}}}(x) \leq \sqrt{D_{f_{\chi^2}}^x(\Pr_{\mathcal{A}}||q)} = \sqrt{\sum_{c \in \mathcal{C}} \frac{[\Pr_{\mathcal{A}}(c|x)]^2}{q(c|x)} - 1}. \quad (4.28)$$

Neyman-Pearson Assume the *f-Divergence* with $f_{\text{NP}}(u) = \frac{(u-1)^2}{u}$. The associated bound is then:

$$\Delta_{0-1}^{\Pr_{\mathcal{A}}}(x) \leq \sqrt{\frac{D_{f_{\text{NP}}}^x(\Pr_{\mathcal{A}}||q)}{1 + D_{f_{\text{NP}}}^x(\Pr_{\mathcal{A}}||q)}}. \quad (4.29)$$

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

$$\Delta_{0-1}^{\Pr_{\mathcal{A}}}(x) \leq \sqrt{D_{f_V}^{x, \frac{1}{\alpha}}(\Pr_{\mathcal{A}}||q)}. \quad (4.30)$$

To simplify the following derivation of global error bounds notice that all the local error bounds have the following form

$$\Delta_{0-1}^{\Pr_{\mathcal{A}}}(x) \leq B(x)$$

where $B(x)$ denotes the above bounds derived from the reversed *Kullback-Leibler*, *Chi-squared*, *Neyman-Pearson* and *Vajda* divergence.

Then the local classification error difference based on $\Pr_{\mathcal{A}}$ relates to the local error difference

as follows:

$$\begin{aligned}
 & \Delta_{0-1}^{\text{Pr}_A}(x) \\
 &= \Pr_A(c_{\mathcal{L}}(x)|x) - \Pr_A(c_{0-1}^q(x)|x) \\
 &= \frac{\phi(R_A(c_{\mathcal{L}}(x)|x)|x)}{Z(x)} - \frac{\phi(R_A(c_{0-1}^q(x)|x)|x)}{Z(x)} \\
 &= \frac{\phi(R_A(c_{\mathcal{L}}(x)|x)|x) - \phi(R_A(c_{0-1}^q(x)|x)|x)}{Z(x)} \\
 &= \frac{\frac{R_A(c_{\mathcal{L}}(x)|x)}{2} \exp(-|R_A(c_{\mathcal{L}}(x)|x) - R_A(c_{\mathcal{L}}(x)|x)|)}{Z(x)} \\
 &\quad - \frac{\frac{R_A(c_{0-1}^q(x)|x)}{2} \exp(-|R_A(c_{0-1}^q(x)|x) - R_A(c_{\mathcal{L}}(x)|x)|)}{Z(x)} \\
 &= \frac{R_A(c_{\mathcal{L}}(x)|x)}{2} \frac{\exp(-0)}{Z(x)} \\
 &\quad - \frac{\frac{R_A(c_{0-1}^q(x)|x)}{2} \exp(-|R_A(c_{0-1}^q(x)|x) - R_A(c_{\mathcal{L}}(x)|x)|)}{Z(x)} \quad \text{according to (4.18)} \\
 &= \frac{R_A(c_{\mathcal{L}}(x)|x) - R_A(c_{0-1}^q(x)|x) \exp(-|\Delta_{\mathcal{L}}(x)|)}{2Z(x)} \quad \text{where } \Delta_{\mathcal{L}}(x) \geq 0 \Rightarrow \exp(-|\Delta_{\mathcal{L}}(x)|) \leq 1 \\
 &\geq \frac{R_A(c_{\mathcal{L}}(x)|x) - R_A(c_{0-1}^q(x)|x)}{2Z(x)} \\
 &= \frac{\Delta_{\mathcal{L}}(x)}{2Z(x)}
 \end{aligned}$$

Using the error bound $\Delta_{0-1}^{\text{Pr}_A}(x) \leq B(x)$ with an assumed error bound $B(x)$ this can be reformulated to a bound on the local reward difference.

$$\Delta_{\mathcal{L}}(x) \leq 2Z(x)\Delta_{0-1}^{\text{Pr}_A}(x) \quad (4.31)$$

$$\leq 2Z(x)B(x). \quad (4.32)$$

By integration over all observations, the local implicit error bound becomes a global implicit error bound.

$$\Delta_{\mathcal{L}} \leq 2 \int pr(x)Z(x)B(x) \, dx.$$

The next section interprets (4.32) to derive explicit error bounds on the squared global error difference.

4.2.5 Explicit Global Error Bounds

We reuse the explicit classification error bounds from Section 4.1.4 to derive explicit bounds on the error difference. Theorem 3 can be reinterpreted for $\Pr_A(c|x)$ and $\Delta_{0-1}^{\text{Pr}_A}(x)$ as follows.

$$\left[\Delta_{0-1}^{\text{Pr}_A}(x) \right]^2 \leq \frac{2}{f''(1)} D_f^x(\Pr_A||q). \quad (4.33)$$

Theorem 4 Consider f -Divergences with such that f''' is monotonically increasing (in compliance with Theorem 3). Then, the following global error bound exists on the error difference.

$$\Delta_{\mathcal{L}}^2 \leq \frac{8}{f''(1)} \int Z^2(x) D_f^x(\Pr_A||q) \, dx. \quad (4.34)$$

Proof Combining with (4.32) results in an error bound on the local error difference.

$$\begin{aligned}
\Delta_{\mathcal{L}}^2(x) &\leq \left[2Z(x)\Delta_{0-1}^{\text{Pr}_{\mathcal{A}}}(x) \right]^2 && \text{according to (4.31)} \\
&= 4Z^2(x) \left[\Delta_{0-1}^{\text{Pr}_{\mathcal{A}}}(x) \right]^2 && \text{according to (4.33)} \\
&\leq 4Z^2(x) \frac{2}{f''(1)} D_f^x(\text{Pr}_{\mathcal{A}}||q) && (4.35)
\end{aligned}$$

By integration of the result of (4.35) over all observations, the local error bound becomes global.

$$\begin{aligned}
\Delta_{\mathcal{L}}^2 &= \left(\int pr(x) \Delta_{\mathcal{L}}(x) dx \right)^2 \\
&\quad \text{since } y^2 \text{ is convex and due to Jensen's inequality: } [E(y)]^2 \leq E(y^2) \\
&\leq \int pr(x) \Delta_{\mathcal{L}}^2(x) dx && \Delta_{\mathcal{L}}^2(x) \leq 4 \frac{Z^2(x)}{f''(1)} D_f^x(\text{Pr}_{\mathcal{A}}||q) \\
&\leq 8 \int pr(x) \frac{Z^2(x)}{f''(1)} D_f^x(\text{Pr}_{\mathcal{A}}||q) dx.
\end{aligned}$$

Tightness From the *f-Divergence* property $D_f^x(\text{Pr}_{\mathcal{A}}||q) = 0 \Leftrightarrow \text{Pr}_{\mathcal{A}}(c|x) = q(c|x)$ follows that presented bound is tightly bounded with the squared global error difference only if the model $q(c|x)$ approaches $\text{Pr}_{\mathcal{A}}(c|x)$ for all observations.

Conclusion A derivation scheme was presented to derive novel bounds on the error difference. The scheme distinguishes between bounds for decision problems based on the 0-1 loss in Section 4.1 and decision problems based on more general losses in Section 4.2. An implicit classification error bound based on the *f-Divergence* was established in Theorem 1. Two proofs of this bound were presented. While the first proof in Section 4.1.1 establishes the classification error bound directly in subsequent steps, the second proof in Section 4.1.2 applies a shortcut from theoretical statistics [Guntuboyina 11, p.3, Corollary II.3] to arrive at the resulting theorem. Different from previous bounds in literature these novel classification error bounds are tight with the classification error difference even for the non-trivial case, i.e. the model is not identical to the true distribution. Implicit bounds with a closed form solution with respect to the classification error difference were discussed in Section 4.1.3, e.g. the Reversed *Kullback-Leibler*, *Chi-Squared*, *Neyman-Pearson* and *Vajda* bound. Unfortunately those explicit bounds are not suitable to derive training criteria. More suited for this task are the explicit bounds on the squared classification error difference which were derived using a *Taylor* expansion in Section 4.1.4.

In Section 4.2 the classification error bounds from Section 4.1 were reused to derive error bounds based on a more general loss, e.g., the Levenshtein loss. This was achieved by reducing decision problems based on a more general loss to an equivalent decision problem based on the 0-1 loss. The local error is transformable into a class posterior, which maintains the Bayes decision rule of the decision problem using the more general loss. As a consequence, the classification error bound based on the 0-1 loss is reusable for deriving error bounds based on the more general loss. Also in the case of a more general loss, implicit and explicit bounds on the error difference are discussed as well as explicit error bounds on the squared error difference which are suitable to derive training criteria.

4.3 Joint Work

The derivation of classification error bounds in Section 4.1 is joint work with Ralf Schlüter and Eugen Beck [Nußbaum-Thom & Tüske⁺ 12, Schlüter & Nußbaum-Thom⁺ 13, Nußbaum-

Thom & Cui⁺ 14]. Ralf Schlüter discovered the direct proof of this theorem in Section 4.1.1. The simulation for Figure 4.1 and Figure 4.2 of implicit classification error bounds in Section 4.1.3 were contributed by Eugen Beck. Eugen Beck contributed the simulation for Figure 4.1 and Figure 4.2 of implicit classification error bounds in Section 4.1.1. The alternative proof in Section 4.1.2 is my individual contribution. The remaining work, including the explicit error bounds in Section 4.1.4 and Section 4.2, is also my individual contribution.

5. ERROR-BOUND-BASED TRAINING CRITERIA

In this section, the explicit error bounds from Section 4 are widened to establish novel empirical training criteria. We distinguish between training criteria derived from classification error bounds for decision problems based on the 0-1 loss from Section 4.1.4 and training criteria derived from error bounds for decision problems based on a more general loss from Section 4.2.5. Each derivation of training criteria is followed by practical examples of discriminative training criteria derived for different *f-Divergences* and an analysis of the non-parametric solution for the optimal model for training criteria based on the power approximation. We first focus on the derivation of training criteria from classification error bounds followed by an investigation of training criteria derived from error bounds using a more general loss. The analysis of the non-parametric solution for the optimal model for training criteria based on a more general loss surprisingly leads to a monotonic transformation of the true reward posterior $\text{Pr}_{\mathcal{A}}$.

5.1 From Global Classification Error Bounds to Training Criteria

In this section, we widen the explicit classification error bounds presented in Section 4.1.4 to derive new criteria. To this end, an expected value, based on the joint true distribution of a monotonic decreasing function of the model posterior will express the upper bound. Using this algebraic form, the true distribution is replaceable with the empirical distribution in order to derive training criteria using properties of the *Kronecker* and *Dirac* delta functions.

In the following we will focus on specific *f-Divergences* based on a function $f(u) = ug(1/u)$ where $g(u)$ is convex and $g(1) = 0$. The first derivative, of $f(u)$ is:

$$f'(u) = g\left(\frac{1}{u}\right) - \frac{1}{u}g'\left(\frac{1}{u}\right).$$

Based on the first derivative the second derivative evaluates to:

$$\begin{aligned} f''(u) &= -\frac{1}{u^2}g'\left(\frac{1}{u}\right) + \frac{1}{u^2}g'\left(\frac{1}{u}\right) + \frac{1}{u^2}g''\left(\frac{1}{u}\right) \\ &= \frac{1}{u^2}g''\left(\frac{1}{u}\right). \end{aligned}$$

Therefore the requirements are met for the identity $f''(1) = g''(1)$.

We derive novel empirical discriminative training criteria. Starting from the explicit classification error bounds in (4.14) presented in Section 4.1.4. The next theorem applies the monotonicity property of the *f-Divergence* demanded in Theorem 3 to widen the error bounds from (4.14). These error bounds are expressable with the joint probability $pr(x, c)$, which subsequently is replaced with the empirical distribution to derive the discriminative training criteria.

Theorem 5 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 3).

Then, the global classification error bound from Theorem 3 based on the f -Divergence can be extended to:

$$\Delta^2 \leq \frac{2}{g''(1)} \int pr(x) D_{ug(1/u)}^x(pr||q) dx \leq \frac{2}{g''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g(q(c|x)) dx.$$

Proof Then the global classification error bounds in (4.14) become:

$$\begin{aligned} \Delta^2 &\leq \frac{2}{f''(1)} \int pr(x) D_f^x(pr||q) dx && \text{with } f(u) = ug\left(\frac{1}{u}\right) \text{ and } f''(1) = g''(1) \\ &= \frac{2}{g''(1)} \int pr(x) D_{ug(1/u)}^x(pr||q) dx \\ &= \frac{2}{g''(1)} \int pr(x) \sum_{c \in \mathcal{C}} q(c|x) \frac{pr(c|x)}{q(c|x)} g\left(\frac{q(c|x)}{pr(c|x)}\right) dx \\ &= \frac{2}{g''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g\left(\frac{q(c|x)}{pr(c|x)}\right) dx \\ &\leq \frac{2}{g''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g\left(\frac{q(c|x)}{1}\right) dx && \text{where } g \text{ is monotonically decreasing} \\ &= \frac{2}{g''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g(q(c|x)) dx. \quad \blacksquare \end{aligned}$$

According to the scheme of [Ney 03, p.642], practical training criteria derive from this extended bound using the empirical distribution of the training data. The empirical distribution of the labeled samples $(x_n, c_n), n = 1, \dots, N$ of observations and labels is defined by:

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

Here, $\delta(x)$ denotes the continuous *Dirac* delta and,

$$\delta(c, \tilde{c}) = \begin{cases} 1 & , c = \tilde{c} \\ 0 & , c \neq \tilde{c} \end{cases}$$

is the discrete *Kronecker* delta. The *Dirac* delta has the useful sifting property [Bracewell 99] for integrable functions

$$\int h(x) \delta(x - x_0) dx = h(x_0),$$

which is used in the following to derive the training criterion. By evaluating the true distribution

through the empirical distribution, the training criterion is derived by:

$$\begin{aligned}
 F_g(q) &= \frac{2}{g''(1)} \int \sum_{c \in \mathcal{C}} pr(x, c) g(q(c|x)) \, dx \\
 &\quad \text{evaluate } pr(x, c) \text{ through } pr_N(x, c) = \frac{1}{N} \sum_{n=1}^N \delta(x - x_n) \delta(c, c_n) \\
 &= \frac{2}{g''(1)} \int \sum_{c \in \mathcal{C}} \frac{1}{N} \sum_{n=1}^N \delta(x - x_n) \delta(c, c_n) g(q(c|x)) \, dx \\
 &= \frac{2}{g''(1)} \frac{1}{N} \sum_{n=1}^N g(q(c_n|x_n)). \quad \text{sifting property}
 \end{aligned}$$

This proof finishes the derivation scheme of empirical discriminative training criteria from classification error bounds. In the next section, we show that the appropriate choices of *f-Divergences* results in novel training criteria.

5.1.1 Examples for Discriminative Training Criteria Based on the f-Divergence

This section presents three examples of training criteria derived from classification error bounds for *f-Divergences* fulfilling Theorem 5. We discuss the discriminative training criteria derived from the *Kullback-Leibler*, Lin, and power-approximation *f-Divergence*. Theorem 5 demands that the *f-Divergence* function $f(u) = ug(1/u)$ of the *f-Divergence* meets the following properties:

- f is decomposable into $f(u) = ug(1/u)$ with g convex and $g(1) = 0$ (decomposition),
- if $f''(u) > 0$ and $f(1) = 0$ then f is convex (convexity),
- $f'''(u)$ exists and is monotonic increasing on the interval $[1, 2]$ (monotonicity).

In the Appendix in Section 10.4, the validity of these properties for the applied *f-Divergences* is discussed in detail.

Kullback-Leibler divergence A natural example fulfilling Theorem 5 is the Kullback-Leibler divergence.

$$g_{\text{KL}}(u) = -\log(u). \quad (5.1)$$

The training criterion derived from the *Kullback-Leibler f-Divergence* becomes the well-known cross-entropy or maximum-mutual-information criterion:

$$F_{g_{\text{KL}}}(q) = -2 \frac{1}{N} \sum_{n=1}^N \log q(c_n|x_n). \quad (\text{CE})$$

Power-Approximation f-Divergence The power approximation for $\alpha \rightarrow 0$ expresses the natural logarithm as:

$$-\log(u) = \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} (1 - u^\alpha). \quad (5.2)$$

The power-approximation *f-Divergence* is defined by:

$$g_{\alpha\text{-PA}}(u) = \frac{1}{\alpha} (1 - u^\alpha). \quad (5.3)$$

The training criterion derived from the power approximation *f-Divergence* becomes the power approximation criterion:

$$F_{g_{\alpha\text{-PA}}}(q) = \frac{2}{\alpha(1-\alpha)} \frac{1}{N} \sum_{n=1}^N (1 - q^\alpha(c_n|x_n)). \quad (\alpha\text{-PA})$$

Lin f-Divergence The function $f_{\alpha\text{-LIN}}(u) = ug_{\alpha\text{-LIN}}(1/u)$ derives the *Lin f-Divergence*

$$g_{\alpha\text{-LIN}}(u) = \log \left(\frac{1}{1+\alpha} (\alpha + u) \right). \quad (5.4)$$

The training criterion derived from the *Lin f-Divergence* becomes the *Lin* criterion:

$$F_{g_{\alpha\text{-LIN}}}(q) = -2(1+\alpha)^2 \frac{1}{N} \sum_{n=1}^N \log \left(\frac{1}{1+\alpha} (\alpha + q(c_n|x_n)) \right). \quad (\alpha\text{-LIN})$$

Due to simplicity, we refer to the previous criteria as the CE, $\alpha\text{-LIN}$, and $\alpha\text{-PA}$ or in general *f-Divergence* criteria. The next section discusses the non-parametric model solution learned by the power approximation criterion for decision problems based on the 0-1 loss.

5.1.2 Optimal Non-Parametric Solution for the Power Approximation f-Divergence

In this section, the optimal non-parametric solution of the model $q(c|x)$ learned from the criterion corresponding to the power-approximation *f-Divergence* is discussed. The empirical distribution converges to the true distribution for an infinite amount of samples $N \rightarrow \infty$:

$$\begin{aligned} pr(x, c) &= \lim_{N \rightarrow \infty} pr_N(x, c) \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \delta(c, c_n) \delta(x - x_n). \end{aligned} \quad (5.5)$$

Consider the training criterion $\mathcal{F}_{g_{\alpha\text{-PA}}}(q)$ with $g_{\alpha\text{-PA}}(u) = \frac{1-u^\alpha}{\alpha}$ with Lagrange multipliers $\mu(x)$ such that $q(c|x)$ fulfills the probability normalization of $q(c|x)$ in each observation $x \in \mathcal{X}$.

$$\begin{aligned} \bar{\mathcal{F}}_{g_{\alpha\text{-PA}}}(q) &= \mathcal{F}_{\alpha\text{-PA}}(q) + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx \\ &= \frac{2}{g''(1)} \frac{1}{N} \sum_{n=1}^N g(q(c_n|x_n)) + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx \\ &\quad \text{using the sifting property } h(x_n) = \int \delta(x - x_n) h(x) dx \\ &= \frac{2}{g''(1)} \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \int g(q(c|x)) \delta(c, c_n) \delta(x - x_n) dx + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx \\ &= \frac{2}{g''(1)} \sum_{c \in \mathcal{C}} \int g(q(c|x)) \frac{1}{N} \sum_{n=1}^N \delta(c, c_n) \delta(x - x_n) dx + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx. \end{aligned}$$

The derivatives concerning the Lagrange multiplier $\mu(x)$, the model $q(k|y)$ and $\mu(y)$ for observation $y \in \mathcal{X}$ and class $k \in \mathcal{C}$ results in:

$$\begin{aligned} \nabla_{q(k|y)} \bar{\mathcal{F}}_{g_{\alpha\text{-PA}}}(q) &= -\frac{2}{g''(1)} q^{\alpha-1}(k|y) \frac{1}{N} \sum_{n=1}^N \delta(k, c_n) \delta(y - x_n) + \mu(y) \stackrel{!}{=} 0, \\ \nabla_{\mu(y)} \bar{\mathcal{F}}_{\alpha\text{-PA}}(q) &= \left(\sum_{c \in \mathcal{C}} q(c|y) - 1 \right) \stackrel{!}{=} 0. \end{aligned} \quad (5.6)$$

With an infinite amount of data, i.e., $N \rightarrow \infty$ the equation evolves to:

$$\begin{aligned}
 \lim_{N \rightarrow \infty} [\nabla_{q(k|y)} \bar{\mathcal{F}}_{\alpha\text{-PA}}(q)] &= \lim_{N \rightarrow \infty} \left[-\frac{2}{g''(1)} q^{\alpha-1}(k|y) \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \delta(k, c_n) \delta(y - x_n) + \mu(y) \right] \\
 &= \frac{2}{g''(1)} q^{\alpha-1}(k|y) \lim_{N \rightarrow \infty} \left[-\frac{1}{N} \sum_{n=1}^N \delta(k, c_n) \delta(y - x_n) \right] + \mu(y) \\
 &\quad \text{according to (5.5)} \\
 &= -\frac{2}{g''(1)} q^{\alpha-1}(k|y) pr(y, k) + \mu(y) \stackrel{!}{=} 0.
 \end{aligned}$$

This can be reformulated to the following equation:

$$q(k|y) = \sqrt[1-\alpha]{\frac{\frac{1}{g''(1)} pr(y, k)}{\mu(y)}}.$$

Combined with the renormalization constraint in (5.20) this results in:

$$\begin{aligned}
 1 &= \sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{\frac{\frac{2}{g''(1)} pr(y, c)}{\mu(y)}}, \\
 \Rightarrow \sqrt[1-\alpha]{\mu(y)} &= \sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{\frac{2}{g''(1)} pr(y, c)}.
 \end{aligned}$$

Then the non-parametric optimal solution for the model in case of infinite training data formulate to:

$$\begin{aligned}
 q(k|y) &= \frac{\sqrt[1-\alpha]{\frac{2}{g''(1)} pr(y, k)}}{\sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{\frac{2}{g''(1)} pr(y, c)}} \\
 &= \frac{\sqrt[1-\alpha]{pr(k|y)}}{\sum_{c \in \mathcal{C}} \sqrt[1-\alpha]{pr(c|y)}}.
 \end{aligned}$$

Since $\sqrt[1-\alpha]{\cdot}$ is monotonically increasing the model-based decision rule results in posterior maximizing class which is identical to the *Bayes* decision rule.

$$c_{0-1}^q(x) = c_{0-1}(x).$$

The next section establishes empirical training criteria from error bounds for decision problems based on more general losses.

5.2 From Error Bounds to Training Criteria

Starting from the explicit error bounds for decision problems based on more general losses from Theorem 4, empirical discriminative training criteria are derived. The next theorem widens this bound with the monotonicity property of the *f-Divergence* demanded in Theorem 4 and expresses the widened error bound in terms of the joint true distribution $pr(x, c)$. By replacing the joint true distribution with the empirical distribution the discriminative training criteria are derived.

Theorem 6 Consider f -Divergences 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 3).

Let us assume $\kappa(x, c)$ as the Laplace distribution:

$$\kappa(x, c) = \frac{\exp(-|R_{\mathcal{A}}(c|x) - R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)|)}{2} \leq 1.$$

Then, the error difference is bounded by the global error bound:

$$\Delta_{\mathcal{L}}^2 \leq \frac{8}{g''(1)} \int Z(x) \sum_{c \in \mathcal{C}} R_{\mathcal{A}}(c|x) \kappa(x, c) g(q(c|x)) \, dx. \quad (5.7)$$

Proof If we start from Theorem 4, an error bound on the global error difference is formulated by integrating over all observations.

$$\begin{aligned} \Delta_{\mathcal{L}}^2 &\leq \frac{8}{g''(1)} \int pr(x) Z^2(x) D_f^x(\text{Pr}_{\mathcal{A}}||q) \, dx \\ &= \frac{8}{g''(1)} \int pr(x) Z^2(x) D_{ug(1/u)}(\text{Pr}_{\mathcal{A}}||q) \, dx && \text{with } f(u) = ug\left(\frac{1}{u}\right) \\ &= \frac{8}{g''(1)} \int pr(x) Z^2(x) \sum_{c \in \mathcal{C}} q(c|x) \frac{\text{Pr}_{\mathcal{A}}(c|x)}{q(c|x)} g\left(\frac{q(c|x)}{\text{Pr}_{\mathcal{A}}(c|x)}\right) \, dx \\ &= \frac{8}{g''(1)} \int pr(x) Z^2(x) \sum_{c \in \mathcal{C}} \text{Pr}_{\mathcal{A}}(c|x) g\left(\frac{q(c|x)}{\text{Pr}_{\mathcal{A}}(c|x)}\right) \, dx && \text{with } \text{Pr}_{\mathcal{A}}(c|x) \text{ def. in (4.24)} \\ &= \frac{8}{g''(1)} \int pr(x) Z^2(x) \sum_{c \in \mathcal{C}} \frac{\phi(R_{\mathcal{A}}(c|x))}{Z(x)} g\left(\frac{q(c|x)}{\text{Pr}_{\mathcal{A}}(c|x)}\right) \, dx && \text{with } \phi \text{ def. in (4.19)} \\ &= \frac{8}{g''(1)} \int pr(x) Z(x) \sum_{c \in \mathcal{C}} \phi(R_{\mathcal{A}}(c|x)) g\left(\frac{q(c|x)}{\text{Pr}_{\mathcal{A}}(c|x)}\right) \, dx \\ &= \frac{8}{g''(1)} \int pr(x) Z(x) \sum_{c \in \mathcal{C}} \frac{R_{\mathcal{A}}(c|x)}{2} \exp(-|R_{\mathcal{A}}(c|x) - R_{\mathcal{A}}(c_{\mathcal{L}}(x)|x)|) g\left(\frac{q(c|x)}{\text{Pr}_{\mathcal{A}}(c|x)}\right) \, dx \\ &= \frac{8}{g''(1)} \int pr(x) Z(x) \sum_{c \in \mathcal{C}} R_{\mathcal{A}}(c|x) \kappa(x, c) g\left(\frac{q(c|x)}{\text{Pr}_{\mathcal{A}}(c|x)}\right) \, dx \\ &\quad \text{where } g \text{ is monotonically decreasing} \\ &\leq \frac{8}{g''(1)} \int pr(x) Z(x) \sum_{c \in \mathcal{C}} R_{\mathcal{A}}(c|x) \kappa(x, c) g\left(\frac{q(c|x)}{1}\right) \, dx \\ &= \frac{8}{g''(1)} \int pr(x) Z(x) \sum_{c \in \mathcal{C}} R_{\mathcal{A}}(c|x) \kappa(x, c) g(q(c|x)) \, dx. \end{aligned} \quad (5.8)$$

■

Now from this widened error bound training criteria are derived by evaluating the true distribution through the empirical distribution. According to the scheme [Ney 03, p.642] practical training criteria derive from this bound by using the empirical distribution of the training data. By substituting the true distribution with the empirical distribution the error bound from (5.8)

is rewritable as:

$$\begin{aligned}
 & \frac{8}{g''(1)} \int Z(x) \sum_{c \in \mathcal{C}} R_{\mathcal{A}}(x, c) \kappa(x, c) g(q(c|x)) \, dx \\
 &= \frac{8}{g''(1)} \int Z(x) \sum_{c \in \mathcal{C}} \sum_{\tilde{c} \in \mathcal{C}} pr(x, \tilde{c}) \mathcal{A}(\tilde{c}, c) \kappa(x, c) g(q(c|x)) \, dx \\
 & \quad \text{evaluate } pr(x, c) \text{ through } pr_N(x, c) = \frac{1}{N} \sum_{n=1}^N \delta(x - x_n) \delta(c, c_n) \\
 &= \frac{8}{g''(1)} \int Z(x) \sum_{c \in \mathcal{C}} \sum_{\tilde{c} \in \mathcal{C}} \frac{1}{N} \sum_{n=1}^N \delta(x - x_n) \delta(\tilde{c}, c_n) \mathcal{A}(\tilde{c}, c) \kappa(x, c) g(q(c|x)) \, dx \\
 &= \frac{8}{g''(1)} \frac{1}{N} \sum_{n=1}^N \int \delta(x - x_n) Z(x) \sum_{c \in \mathcal{C}} \sum_{\tilde{c} \in \mathcal{C}} \delta(\tilde{c}, c_n) \mathcal{A}(\tilde{c}, c) \kappa(x, c) g(q(c|x)) \, dx \\
 &= \frac{8}{g''(1)} \frac{1}{N} \sum_{n=1}^N \int \delta(x - x_n) Z(x) \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x, c) g(q(c|x)) \, dx \quad \text{sifting property} \\
 &= \frac{8}{g''(1)} \frac{1}{N} \sum_{n=1}^N Z(x_n) \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) g(q(c|x_n)) \\
 &\leq \frac{8}{g''(1)} \max_{n=1, \dots, N} \{Z(x_n)\} \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) g(q(c|x_n)) \quad (5.9)
 \end{aligned}$$

since $\kappa(x_n, c) \leq 1$

$$\leq \frac{8}{g''(1)} \max_{n=1, \dots, N} \{Z(x_n)\} \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) g(q(c|x_n)). \quad (5.10)$$

Both (5.9) and (5.10) are suitable to derive a training criterion. In the remainder of this section (5.9) will be discussed. Moreover, later sections analyze the training criterion in (5.10) since no appropriate empirical approximation of $\kappa(x, c)$ exists yet.

The following equations show a simplified version of the training criterion of (5.9) independent from constant factors that do not affect the minimization of the free parameters during training.

$$\begin{aligned}
 \hat{q} &= \arg \min_q \left\{ \frac{8}{g''(1)} \max_{n=1, \dots, N} \{Z(x_n)\} \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) g(q(c|x_n)) \right\} \\
 &= \arg \min_q \left\{ \frac{1}{g''(1)} \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) g(q(c|x_n)) \right\} \\
 &= \arg \min_q \{ \mathcal{F}_g(q) \}. \quad (5.11)
 \end{aligned}$$

The following empirical training criterion evolves from the minimization.

$$\mathcal{F}_g(q) = \frac{1}{g''(1)} \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) g(q(c|x_n)). \quad (5.12)$$

If we want to combine different *f-Divergence* criteria, the factor $\frac{1}{g''(1)}$ is vital in order to keep the ratio between the criteria balanced concerning the error difference.

5.2.1 Examples for Discriminative Training Criteria Based on the f-Divergence

This section reconsiders the *Kullback-Leibler*, *Lin*, and *Power-Approximation f-Divergences* in the context of Theorem 6. As demonstrated in Section 5.1.1, these *f-Divergences* fulfill the

requirements of Theorem 5. Therefore, the requirements of Theorem 6 are then fulfilled by the *Kullback-Leibler*, *Lin*, and *Power-Approximation f-Divergence* as well. Those training criteria are derived from a bound on the error difference in (5.9) and (5.10). Therefore, now discriminative training criteria can be derived for the more general losses as an extension to the 0-1 loss case.

Kullback-Leibler divergence A natural example fulfilling Theorem 6 is the Kullback-Leibler divergence.

$$g_{\text{KL}}(u) = -\log u. \quad (5.13)$$

The training criterion derived from the *Kullback-Leibler f-Divergence* becomes:

$$F_{g_{\text{KL}}}(q) = -\frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) \log q(c|x_n). \quad (\text{WMMI})$$

This criterion is a weighted version of the cross-entropy or maximum-mutual-information criterion. Therefore this criterion is referred to as the Weighted Maximum-Mutual-Information (WMMI) criterion.

Power-Approximation f-Divergence From the power approximation for $\alpha \rightarrow 0$, the natural logarithm derives:

$$-\log(u) = \lim_{\alpha \rightarrow 0} \frac{1}{\alpha} (1 - u^\alpha). \quad (5.14)$$

The power-approximation *f-Divergence* is defined by:

$$g_{\alpha\text{-PA}}(u) = \frac{1}{\alpha} (1 - u^\alpha). \quad (5.15)$$

The training criterion derived from the power approximation *f-Divergence* becomes the power approximation criterion:

$$\begin{aligned} & \underset{q}{\operatorname{argmin}} \left\{ \frac{1}{\alpha(1-\alpha)} \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) (1 - q^\alpha(c|x_n)) \right\} \\ &= \underset{q}{\operatorname{argmin}} \left\{ -\frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) q^\alpha(c|x_n) \right\} \\ &= \underset{q}{\operatorname{argmax}} \left\{ \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) q^\alpha(c|x_n) \right\}. \\ & \rightsquigarrow F_{g_{\alpha\text{-PA}}}(q) = \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) q^\alpha(c|x_n). \quad (\alpha\text{-MPE}) \end{aligned}$$

The functional form of this criterion is very similar to the well-known MPE criterion. Therefore this criterion is referred to as the α -MPE criterion.

Lin f-Divergence The function

$$g_{\alpha\text{-LIN}}(u) = -\log\left(\frac{1}{1+\alpha}(\alpha+u)\right) \quad (5.16)$$

derives the *Lin f-Divergence*. The training criterion derived from the *Lin f-Divergence* becomes the *Lin* criterion:

$$\mathcal{F}_{g_{\alpha\text{-LIN}}}(q) = -(1+\alpha)^2 \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) \log\left(\frac{1}{1+\alpha}(\alpha + q(c|x_n))\right). \quad (\alpha\text{-LIN})$$

5.2.2 Optimal Non-Parametric Solution for the Power Approximation f-Divergence

This section discusses the optimal non-parametric solution of the model $q(c|x)$ learned from the criterion derived from the power-approximation *f-Divergence*. The power-approximation *f-Divergence* is defined by:

$$g_{\alpha\text{-PA}}(u) = \frac{1}{\alpha} (1 - u^\alpha). \quad (5.17)$$

The training criterion derived from the power approximation *f-Divergence* becomes the α -MPE criterion:

$$\leadsto \mathcal{F}_{g_{\alpha\text{-PA}}}(q) = \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \kappa(x_n, c) q^\alpha(c|x_n).$$

For infinite training data, the true distribution is expressible as the sampling of observations and classes:

$$\begin{aligned} pr(x, c) &= \lim_{N \rightarrow \infty} pr_N(x, c) \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \delta(c, c_n) \delta(x - x_n). \end{aligned} \quad (5.18)$$

Consider the training criterion $\mathcal{F}_f(q)$ with the Lagrange multipliers which fulfill the probability normalization of $q(c|x)$.

$$\begin{aligned}
 \overline{\mathcal{F}}_{g_{\alpha\text{-PA}}}(q) &= \mathcal{F}_{\alpha\text{-PA}}(q) + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx \quad \text{with } h(x_n) = \int \delta(x - x_n) h(x) dx \\
 &= \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \int \delta(x - x_n) \kappa(x, c) g(q(c|x)) dx \\
 &\quad \text{with } \mathcal{A}(c_n, c) = \sum_{\tilde{c} \in \mathcal{C}} \mathcal{A}(\tilde{c}, c) \delta(\tilde{c}, c_n) \\
 &\quad + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx \\
 &= \int \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \sum_{\tilde{c} \in \mathcal{C}} \mathcal{A}(\tilde{c}, c) \delta(\tilde{c}, c_n) \delta(x - x_n) \kappa(x, c) g(q(c|x)) dx \\
 &\quad + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx \\
 &= \int \sum_{c \in \mathcal{C}} \kappa(x, c) g(q(c|x)) \sum_{\tilde{c} \in \mathcal{C}} \mathcal{A}(\tilde{c}, c) \frac{1}{N} \sum_{n=1}^N \delta(\tilde{c}, c_n) \delta(x - x_n) dx \\
 &\quad + \int \mu(x) \left(\sum_{c \in \mathcal{C}} q(c|x) - 1 \right) dx.
 \end{aligned}$$

The derivative for the model $q(k|y)$ and $\mu(y)$ for an observation $y \in \mathcal{X}$ results in:

$$\nabla_{q(k|y)} \overline{\mathcal{F}}_{g_{\alpha\text{-PA}}}(q) = -\kappa(y, k) q^{\alpha-1}(k|y) \sum_{\tilde{c} \in \mathcal{C}} \mathcal{A}(\tilde{c}, k) \frac{1}{N} \sum_{n=1}^N \delta(\tilde{c}, c_n) \delta(y - x_n) + \mu(y) \stackrel{!}{=} 0, \quad (5.19)$$

$$\nabla_{\mu(y)} \overline{\mathcal{F}}_{g_{\alpha\text{-PA}}}(q) = \left(\sum_{c \in \mathcal{C}} q(c|y) - 1 \right) \stackrel{!}{=} 0. \quad (5.20)$$

With an infinite amount of data $N \rightarrow \infty$ equation (5.19) can be reformulated to:

$$\begin{aligned}
 \lim_{N \rightarrow \infty} [\nabla_{q(k|y)} \overline{\mathcal{F}}_{g_{\alpha\text{-PA}}}(q)] &= \lim_{N \rightarrow \infty} \left[-\kappa(y, k) q^{\alpha-1}(k|y) \sum_{\tilde{c} \in \mathcal{C}} \mathcal{A}(\tilde{c}, k) \frac{1}{N} \sum_{n=1}^N \delta(\tilde{c}, c_n) \delta(y - x_n) + \mu(y) \right] \\
 &= -\kappa(y, k) q^{\alpha-1}(k|y) \sum_{\tilde{c} \in \mathcal{C}} \mathcal{A}(\tilde{c}, k) \lim_{N \rightarrow \infty} \left[\frac{1}{N} \sum_{n=1}^N \delta(\tilde{c}, c_n) \delta(y - x_n) \right] + \mu(y) \\
 &= -\kappa(y, k) q^{\alpha-1}(k|y) \sum_{\tilde{c} \in \mathcal{C}} \mathcal{A}(\tilde{c}, k) pr(y, \tilde{c}) + \mu(y) \\
 &= -\kappa(y, k) q^{\alpha-1}(k|y) R_{\mathcal{A}}(k|y) + \mu(y) \\
 &= -q^{\alpha-1}(k|y) R_{\mathcal{A}}(k|y) \kappa(y, k) + \mu(y) \quad \text{with } \phi(R_{\mathcal{A}}(k|y)|y) = R_{\mathcal{A}}(k|y) \kappa(y, k) \\
 &= -q^{\alpha-1}(k|y) \phi(R_{\mathcal{A}}(k|y)|y) + \mu(y) \stackrel{!}{=} 0.
 \end{aligned}$$

The last equation is equivalent to:

$$q(k|y) = {}^{1-\alpha}\sqrt{\frac{\phi(R_{\mathcal{A}}(k|y))}{\mu(y)}}.$$

Combined with the re-normalization constraint in (5.20) this results in:

$$\begin{aligned} 1 &= \sum_{c \in \mathcal{C}} {}^{1-\alpha}\sqrt{\frac{\phi(R_{\mathcal{A}}(c|y))}{\mu(y)}} \\ \Rightarrow {}^{1-\alpha}\sqrt{\mu(y)} &= \sum_{c \in \mathcal{C}} {}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(c|y))}. \end{aligned}$$

Then the non-parametric optimal solution for the model in case of infinite training data is:

$$q(k|y) = \frac{{}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(k|y))}}{\sum_{c \in \mathcal{C}} {}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(c|y))}}.$$

Since ${}^{1-\alpha}\sqrt{\cdot}$ is monotonically increasing the optimal solution results in the *Bayes* decision rule $c_{\mathcal{L}}(x)$. The general form of the optimal solution is a re-weighted version of $\Pr_{\mathcal{A}}(k|y)$.

Discussion Now we discuss the implications of the non-parametric model solution of the α -MPE criterion for decision problems based on a more general loss using the maximum posterior model-based decision rule. The correct model-based decision rule – minimizing the posterior expected loss – is too time consuming and therefore substituted with the 0-1 loss model-based decision rule – maximizing the model posterior – instead.

$$c_{0-1}^q(x) := \operatorname{argmax}_{c \in \mathcal{C}} \{q(c|x)\}. \quad (5.21)$$

Surprisingly, as shown in Section 5.2.2, the optimal non-parametric model under the α -MPE criterion results in a monotonically increasing transformation of the posterior $\Pr_{\mathcal{A}}(c|x)$:

$$q(k|x) = \frac{{}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(k|x))}}{\sum_{c \in \mathcal{C}} {}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(c|x))}}. \quad (5.22)$$

Therefore the model-based maximum posterior decision rule is identical with the *Bayes* decision rule of the decision problem based on the more general loss:

$$\begin{aligned}
 c_{0-1}^q(x) &:= \operatorname{argmax}_{c \in C} \{q(c|x)\} \\
 &= \operatorname{argmax}_{c \in C} \left\{ \frac{{}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(c|y))}}{\sum_{k \in C} {}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(k|y))}} \right\} \\
 &= \operatorname{argmax}_{c \in C} \left\{ {}^{1-\alpha}\sqrt{\phi(R_{\mathcal{A}}(c|y))} \right\} \\
 &= \operatorname{argmax}_{c \in C} \{\phi(R_{\mathcal{A}}(c|y))\} \\
 &= \operatorname{argmax}_{c \in C} \left\{ \frac{\phi(R_{\mathcal{A}}(c|y))}{\sum_{k \in C} \phi(R_{\mathcal{A}}(k|y))} \right\} \\
 &= \operatorname{argmax}_{c \in C} \{\Pr_{\mathcal{A}}(c|x)\} \\
 &= \operatorname{argmax}_{c \in C} \{R_{\mathcal{A}}(c|y)\} \\
 &= \operatorname{argmin}_{c \in C} \{E_{\mathcal{L}}(k|y)\} \\
 &= c_{\mathcal{L}}(x).
 \end{aligned} \tag{5.23}$$

So in summary, under ideal theoretical conditions of the non-parametric solution and an infinite amount of training data, the correct but time-consuming model-based decision rule – minimizing the posterior expected general loss – can be substituted with the time-efficient 0-1 loss model-based decision rule – maximizing the model posterior – instead.

What does this mean in practice? The non-parametric solution symbolizes the error optimal model, which can be learned by the model under ideal conditions. As shown in (5.22), under these conditions, the model’s solution becomes a monotonically increasing transformation of the true reward posterior $\Pr_{\mathcal{A}}(c|x)$. Therefore, as shown in (5.23), maximizing the model-based decision rule based on the 0-1 loss, i.e., choosing the most likely sentence, is identical to minimizing the *Bayes* decision rule based on a more general loss.

Transferring this to the case of automatic speech recognition means: By applying an error-based training criterion, e.g., the α -MPE criterion, involving the more general loss, the model is trained in such a way that at test time the most probable sentence implicitly tries to minimize the posterior expected *Levenshtein* loss. This result confirms our initial assumption that in automatic speech recognition, error-aware training criteria shift the decision boundaries of the model to better anticipate the mismatch between the performance measure and the model-based decision rule.

Conclusion We widened the explicit error bounds from Section 4 to establish novel empirical training criteria. Like in Section 4 we distinguish between training criteria derived from classification error bounds for decision problems based on the 0-1 loss from Section 4.1 and training criteria derived from error bounds for decision problems based on a more general loss from Section 4.2. Each derivation of training criteria is followed by practical examples of discriminative training criteria derived for different *f-Divergences* and an analysis of the non-parametric solution for the optimal model for training criteria based on the power approximation. We first focused on the derivation of training criteria from classification error bounds followed by an investigation

of training criteria derived from error bounds using a more general loss. The analysis of the non-parametric solution for the optimal model for training criteria based on a more general loss surprisingly leads to a monotonic transformation of the true reward posterior $\Pr_{\mathcal{A}}$.

5.3 Joint Work

The work in Section 5 is completely part of my individual contribution [Nußbaum-Thom & Tüske⁺ 12, Nußbaum-Thom & Cui⁺ 14, Nußbaum-Thom & Schlüter⁺ 17].

6. APPLICATION TO FRAME-WISE NEURAL NETWORK TRAINING FOR AUTOMATIC SPEECH RECOGNITION

In this section, we apply the empirical training criteria introduced in Section 5.1 to discriminative frame-wise training in automatic speech recognition. As introduced in Section 2.1.1 in frame-wise training, the posterior of an HMM emission state given an acoustic feature is modeled, i.e., in this section, a class is an HMM emission state given a feature observation. In the following, the model-based posterior corresponds to the softmax output layer of the neural networks.

6.1 Empirical Results for the Frame-Wise f -Divergence Training

This section evaluates the CE, 1-LIN, and α -PA training criteria derived in Section 5.1.1 for frame-wise training of deep neural networks on the Babel Vietnamese and Bengali speech recognition tasks. In the following, we describe the experimental setup.

Optimization The neural networks using the novel training criteria are optimized using the SGD framework [Rumelhart & Hinton⁺ 86]. If we implement the novel criteria based on the f -Divergence within this framework, only the gradient of the neural network for 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, the gradient has a canonical form:

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

We want to emphasize that the factor $1/g''(1)$ from Theorem 3 theoretically describes the correct weighting between different f -Divergence criteria because only by using these factors the corresponding bounds limit the classification error difference without bias.

Combination of Training Criteria Since a linear combination of f -Divergences is still an f -Divergence, the linear combination of the f -Divergence bounds and criteria results in valid bounds and training criteria along with the definition and proofs we covered so far. Furthermore, a linear combination of criteria $F_{g_1}(q), \dots, F_{g_I}(q)$ with weights $\lambda_1, \dots, \lambda_I \in \mathbb{R}$ comes with no (or little) 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 (at frame index n)

$$\sum_{i=1}^I \lambda_i g'_i(q(c_n|x_n))$$

on top of the standard cross-entropy derivative $\nabla q(c_n|x_n)$, which integrate efficiently using a matrix multiplication on GPUs. We ask the question: Why should a combination of criteria result in an improved model? As outlined before, the power approximation *f-Divergence* criterion discussed in Section 5.1.1 is consistent with a monotonically increasing transformation of the true posterior. In theory, during optimization, different *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.

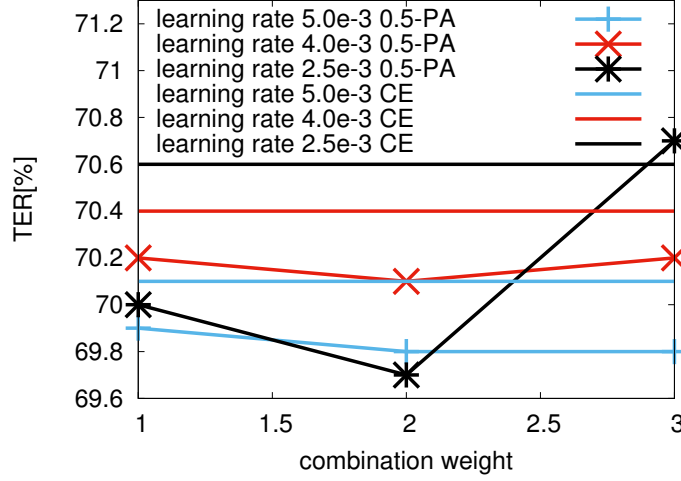


Figure 6.1: TER[%] progress vs. the linear combination of the 0.5-PA criterion with CE on the Bengali Babel test set.

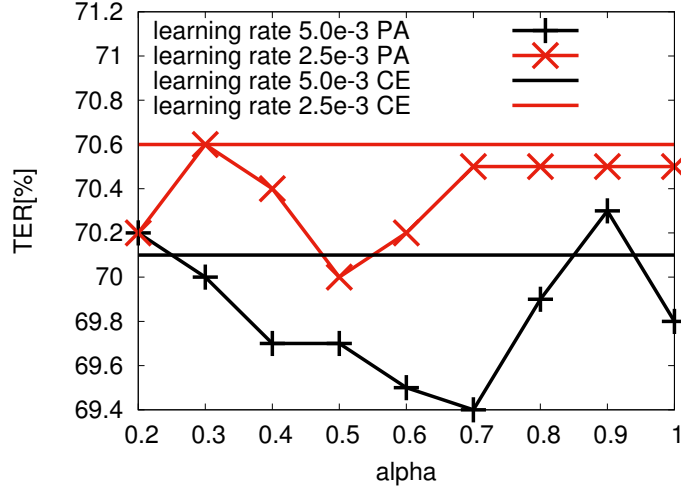


Figure 6.2: TER[%] progress vs. factor α of the α -PA criterion measured on the Bengali development set.

Experimental Setup We test the new training criteria on Babel Vietnamese and Bengali automatic speech recognition tasks. Both training and development sets are composed similar to the Babel tasks described in [Cui & Cui⁺ 13]. The Babel training and development sets for each language consist of about 20h conversational and scripted telephony speech each. For Vietnamese, 14h 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 comprises of Perceptive Linear Predictive (PLP) features. We concatenate nine consecutive frames, and use a Linear Discriminant Analysis (LDA) 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 initializes the DNN. The alignments used for the training of the neural network come from alternating optimization of DNN training and realignment. The deep neural network is optimized using the mini-batch based SGD algorithm [Seide & Li⁺ 11, Kingsbury & Sainath⁺ 12], and performed efficiently using GPUs. On top of the best frame-wise trained DNN models, a Hessian-free (HF) state-wise Minimum Bayes Risk (sMBR) optimization [Kingsbury & Sainath⁺ 12] 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* [Kneser & Ney 95] trigram for Vietnamese and bigram model for Bengali was estimated for recognition. The token error rate (TER) measures all improvements. This is the WER for language-dependent tokens instead of words.

Experimental Results In initial experiments, we used only single criteria to train the DNNs. It turned out, only the criteria resulting from the Lin and α -PA divergence (for small values of α) can produce comparable but slightly worse results than the CE criterion.

The next series of experiments test the linear combination using weights 1, 2, and 3 of Lin and 0.5-PA with CE. We show the TER of the baseline CE and the most successful combinations in Table 6.1. 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 Hessian-Free sMBR training.

Table 6.1: *TER[%] results for frame-wise training using the CE criterion in combination with the 0.5-PA and Lin (+1-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-PA	+HF	+1-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 the 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-PA criteria to figure out a proper weight across varying learning rates. Figure 6.1 shows the TER progress for different learning rates for combinations of weight 1, 2, and 3 between the 0.5-PA and CE criterion for the Bengali development set. We observe the same behavior for the Vietnamese development set as well. Interestingly, combination weight 2 is the ratio suggested by the factor $1/g''(1)$ of Theorem 3 for the 0.5-PA criterion. Although the sample of this measurement is not large enough to make a real statement about this assumption, this is an exciting direction for future work.

The third series of experiments measure the TER progress for a uniform combination of the α -PA and CE criteria for different values of $\alpha = 1, 0.9, \dots, 0.2$, and initial learning rates of the last layer. Figure 6.2 shows the corresponding results. The combination of both criteria leads to

improved results for most of the α values. Furthermore, a stable interval for α is observable from 0.4 through 0.7.

The next section discusses two different variants of the α -PA criterion by randomly or systematically changing the criterion throughout the training.

6.2 Noisy f-Divergence Criteria and Minimum f-Divergence Training

Deep neural networks have a vast number of free parameters and can learn classifiers for various problems. However, overfitting is known to be a common problem for deep neural networks. Dropout [Srivastava & Hinton⁺ 14] is an effective and popular method to avoid overfitting. Dropout, the random dropping out of activations according to a specified rate, is a straightforward but effective method to avoid overfitting of deep neural networks to the training data.

In this section, we approach regularization from the view of the objective function by dynamically changing the objective function to avoid local minima. The underlying theory bases on the training criteria introduced in Section 5.1. A novel family of training criteria is introduced based on the f-divergence. These criteria are a generalization of the cross-entropy criterion. We introduce two regularization schemes – the first approach minimizes over a family of training criteria in order to achieve the lowest possible criterion, and the second approach randomly chooses a criterion from a family of criteria according to a Gaussian distribution.

In practical experiments on the WSJ-5K corpus, the proposed schemes are successfully evaluated compared to dropout for deep neural networks and Bidirectional Gated Recurrent Units (BGRUs), both as standalone approaches and in combination with dropout.

A Family of Discriminative Training Criteria based on the f-Divergence The power approximation (α -PA) training criterion associated with the function $g(u) = (1 - u^\alpha)$ and $g''(1) = 1 - \alpha$ fulfills the above conditions and results in

$$\mathcal{F}_{g_{\alpha\text{-PA}}}(q) = \frac{1}{N} \sum_{n=1}^N \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1 - \alpha)}. \quad (6.1)$$

For $\alpha \rightarrow 0$ the power approximation converges to the logarithm $-\log(u) = \lim_{\alpha \rightarrow 0} \frac{1-u^\alpha}{\alpha}$, while the corresponding *f-Divergence* converges to the *Kullback-Leibler f-Divergence*, and the criterion converges to the cross-entropy criterion.

$$\mathcal{F}_{g_{\text{KL}}}(q) = \mathcal{F}_{\text{CE}}(q) = -\frac{1}{N} \sum_{n=1}^N \log q(c_n|x_n) \quad (6.2)$$

As convex functions are closed under addition, this is also valid for the corresponding *f-Divergences* and derived training criteria.

In the next section, we introduce the minimum power approximation training criterion, which minimizes over α to derive a smaller bound, and therefore a better training criterion.

Minimum Conjugate Power Approximation All training criteria are derived according to the scheme from Section 5.1 from a classification error bound on the classification error difference presented in Section 4.1. By choosing a tighter bound we can expect a better training criterion. According to this argument, Figure 6.3 shows that for different model posteriors, different α -PA criteria are minimal. Therefore, by minimizing the corresponding α -PA classification error bound over α will result in a tighter bound. The corresponding training criteria for a sample or

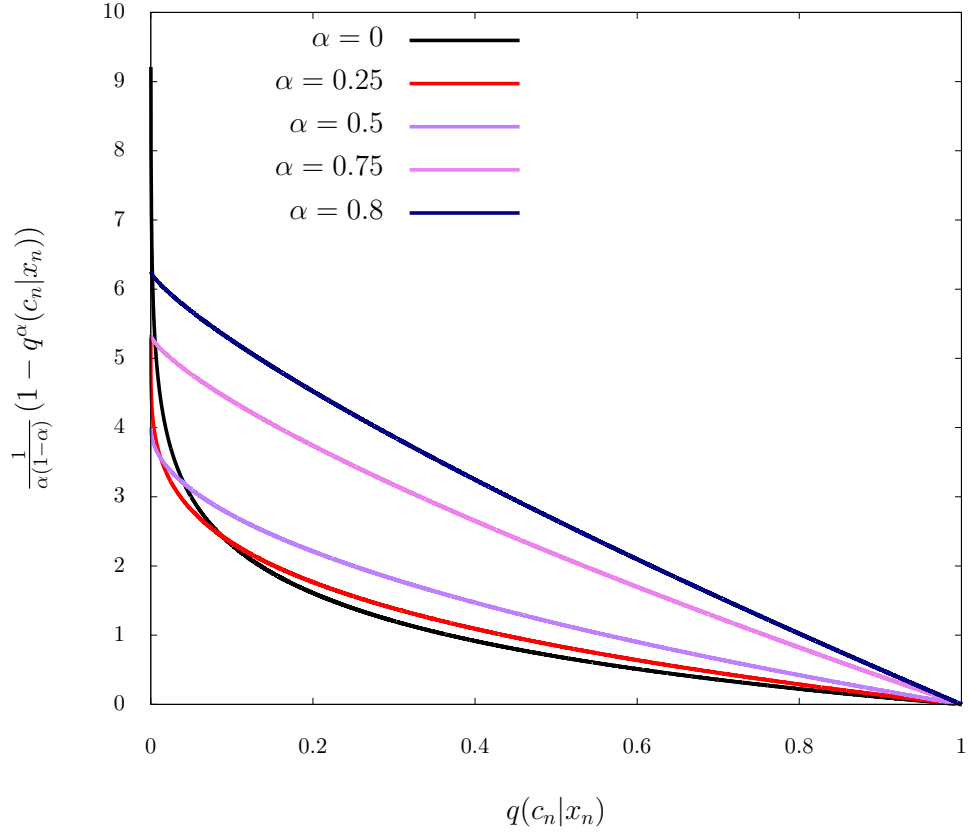


Figure 6.3: Function $\frac{1-q^\alpha(c_n|x_n)}{\alpha(1-\alpha)}$ corresponding to one sample of the α -PA criterion which converges to the cross-entropy criterion for $\alpha \rightarrow 0$.

batch-wise minimization over α results in

$$\mathcal{F}_{\text{MIN-SAMP-PA},\beta}(q) = \frac{1}{N} \sum_{n=1}^N \min_{\alpha \in [0,\beta]} \left\{ \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1-\alpha)} \right\}, \text{ and}$$

$$\mathcal{F}_{\text{MIN-BATCH-PA},\beta}(q) = \min_{\alpha \in [0,\beta]} \left\{ \frac{1}{N} \sum_{n=1}^N \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1-\alpha)} \right\}.$$

For practical purposes, we have included a parameter $\beta \in [0, 1]$ to limit the choice of α . The practical implementation uses a golden section search for the minimal α within the interval $[0, \beta]$.

The next section introduces the other proposed approach of noisy objective functions which chooses the parameter α randomly drawn according to a Gaussian distribution.

Noisy Objective Function In this section, we introduce the noisy power approximation criterion. By randomly choosing the parameter $\alpha \sim \mathcal{N}(\mu, \sigma^2)$ according to a Gaussian distribution with specific mean μ and variance σ^2 the resulting training criterion should be less sensitive to local minima. Therefore, this method is suitable for regularization towards a better local optimum.

The resulting training criterion randomly chooses α either per sample or batch-wise

$$\mathcal{F}_{\text{RAND-SAMP-PA},\mu,\sigma^2}(q) = \frac{1}{N} \sum_{n=1}^N \text{rand}_{\alpha \sim \mathcal{N}(\mu,\sigma^2)} \left\{ \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1 - \alpha)} \right\}, \text{ and}$$

$$\mathcal{F}_{\text{RAND-BATCH-PA},\mu,\sigma^2}(q) = \text{rand}_{\alpha \sim \mathcal{N}(\mu,\sigma^2)} \left\{ \frac{1}{N} \sum_{n=1}^N \frac{(1 - q^\alpha(c_n|x_n))}{\alpha(1 - \alpha)} \right\}.$$

Experimental Setup The Gaussian mixture Hidden Markov Model (GHMM) baseline recognition system for WSJ0 uses 1500 generalized triphone states, which were top-down clustered using a decision tree, plus one silence state. We show the corpus statistics for WSJ0 in Table 6.2. Gaussian

Table 6.2: Corpus statistics.

Corpus	Train/ Dev/ Eval		
	Data[h]	#Sentences	#Words
WSJ0	15:17/ 0:46/ 0:4	7k/ 410/ 330	130k/ 6k/ 5k

mixture distributions with a total of about 200k densities model the emission probabilities. The raw acoustic features are 19-dimensional PLP features. Temporal context is included by splicing nine successive frames of PLP features into super-vectors, then projecting to forty dimensions using linear discriminant analysis (LDA). The recognizer utilizes a 5k lexicon and trigram language model for testing on the WSJ0 corpus.

All neural network experiments use an acoustic front-end that comprises 40 Log-Mel features augmented with delta and double delta. We evaluate the neural networks as hybrid acoustic models for automatic speech recognition. On many tasks, more complex Long Short-Term Memory (LSTM) RNNs [Graves & Jaitly⁺ 13, Sak & Senior⁺ 14] outperformed simple recurrent neural networks (RNNs) and DNNs. The Gated Recurrent Unit (GRU) recently introduced in [Cho & Van Merriënboer⁺ 14] has a structure similar to the LSTM. The GRU makes each recurrent unit to capture dependencies of different time scales. Similarly to the LSTM unit, the GRU has gating units that modulate the flow of information inside the unit, however, without having a separate memory cell. Interestingly, the studies in [Chung & Güleşhre⁺ 14, Józefowicz & Zaremba⁺ 15] indicate that the GRU matches the LSTM performance. GRUs also consume a smaller number of parameters compared to LSTMs for the same hidden layer size due to a smaller gating mechanism and missing peepholes. In initial experiments, we verified that BGRUs matched the performance of bidirectional LSTMs for automatic speech recognition and decided to use BGRUs in the remaining experiments. Due to these arguments, we decided in our experiments to use GRUs instead of LSTMs.

We trained DNNs and BGRUs using the architectures and recipes described in [Nußbaum-Thom & Cui⁺ 16b].

Experimental Results This section describes the result of a series of experiments that combine and compare different training criteria. In the following, we use the notation A+B in case criterion A is combined with criterion B, i.e., the training uses (A+B)/2 effectively. Since the addition of multiple *f-Divergences* remains an *f-Divergence*, the training criteria derived here still come from error bounds based on the f-Divergence, as *f-Divergences* are closed under addition. We choose in the subsequent experiments, the best model according to the best WER on the DEV corpus. As a contrastive result, the overall best model on the EVAL corpus achieves a WER of 1.9%. This model is a BGRU with dropout 0.1 and trained using the CE+RAND-SAMP-PA criterion. This model, however, achieves a WER of 2.6% on the DEV corpus.

Table 6.3: The WER[%] as a function of different training criteria working as a regularization scheme for DNN, BGRU and BGRU(0.1) (BGRU with dropout rate 0.1) model on the WSJ0 test corpora.

MODEL	CRITERION	WER[%]	
		DEV	EVAL
DNN	CE	3.1	3.3
	MIN-SAMP-PA	3.0	3.5
	CE+MIN-SAMP-PA	2.9	3.0
	MIN-BATCH-PA	3.1	3.4
	CE+MIN-BATCH-PA	3.1	3.2
	RAND-SAMP-PA	3.0	3.3
	CE+RAND-SAMP-PA	2.9	3.0
	RAND-BATCH-PA	2.9	3.5
	CE+RAND-BATCH-PA	2.9	3.1
BGRU	CE	2.6	2.4
	MIN-SAMP-PA	2.7	2.0
	CE+MIN-SAMP-PA	2.7	2.2
	MIN-BATCH-PA	2.7	2.2
	CE+MIN-BATCH-PA	2.7	2.1
	RAND-SAMP-PA	2.6	2.2
	CE+RAND-SAMP-PA	2.6	2.0
	RAND-BATCH-PA	2.5	2.2
	CE+RAND-BATCH-PA	2.5	2.2
BGRU(0.1)	CE	2.6	2.3
	MIN-SAMP-PA	2.6	2.3
	CE+MIN-SAMP-PA	2.7	2.1
	MIN-BATCH-PA	2.6	2.2
	CE+MIN-BATCH-PA	2.7	2.2
	RAND-SAMP-PA	2.6	2.1
	CE+RAND-SAMP-PA	2.6	2.1
	RAND-BATCH-PA	2.5	2.3
	CE+RAND-BATCH-PA	2.5	2.2

In initial experiments, we first trained the DNNs using the CE, PA, and $CE+\alpha$ -PA for $\alpha \in \{0.001, 0.01, 0.1, 0.2, 0.3, 0.4\}$. The baseline CE criterion using the DNN model results in a WER of 3.1% on the DEV corpus, and 3.3% on the EVAL corpus.

Figure 6.4 shows the result for the α -PA criterion compared to the CE criterion only. We omitted the $CE+\alpha$ -PA criterion since it had a similar performance.

By choosing the optimal α according to the WER on the DEV corpus, a small improvement over the CE criterion is achievable while leading to a WER degradation on the EVAL corpus. For slightly different $\alpha \in \{0.01, 0.1\}$, the EVAL performance is better than the baseline CE criterion. The evaluation of the MIN-PA and $CE+MIN$ -PA criterion minimized on a sample or batch-wise, without restriction on β (i.e., $\beta = 1$), results in a WER ranging from 3.7 to 4.0. This setup does slightly worse than the CE baseline criterion. The analysis of these results, and with Figure 6.3 in mind, where the CE criterion is lower-bounded by α -PA for smaller values of α , suggests a different strategy – a smaller α close to zero can achieve a better criterion. Therefore, in the next set of experiments, α is constrained to smaller values either by choosing a small β or choosing a smaller mean and variance.

In the next set of experiments, we apply dropout to the intermediate layer outputs for DNNs and BGRUs using the dropout rates $\{0, 0.01, 0.1, 0.2, 0.3, 0.4, 0.5\}$. For BGRUs, we also found it useful to apply the same dropout rate to the weights simultaneously. Dropout did not help for DNNs. However, for BGRUs the best dropout rate is 0.1, which reduced the WER on the DEV

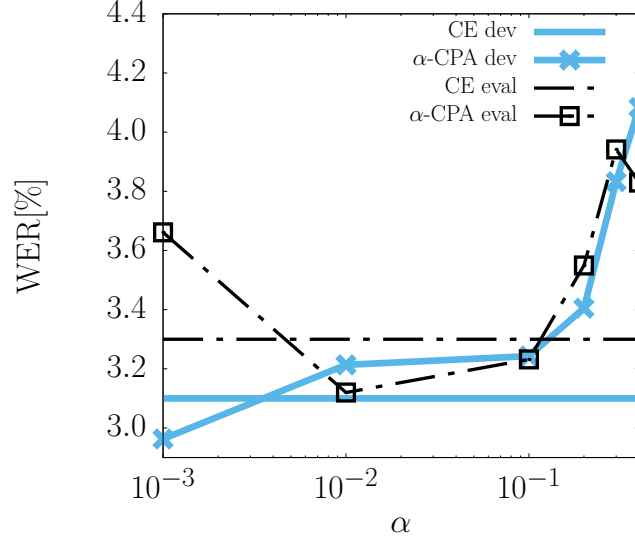


Figure 6.4: The WER[%] as a function of the parameter α of the α -PA criterion trained using DNN models.

corpus only slightly by a couple of error counts, but reduced the WER on the EVAL corpus from 2.4 to 2.3.

In the next set of experiments, we train DNN and BGRU models with a variety of combined criteria where the choice for α is constrained to values close to zero. We trained for $\beta \in \{10^{-i} | i \in \{1, 2, 3, 4, 5\}\}$ the DNN and BGRU models using the following criteria

- MIN-SAMP-PA,
- CE+MIN-SAMP-PA,
- MIN-BATCH-PA,
- and CE+MIN-BATCH-PA.

Also for $\mu \in \{10^{-i} | i \in \{1, 2, 3, 4, 5, 6\}\}$ and $\sigma^2 \in \{10^{-i} | i \in \{1, 2, 3, 4, 5, 6\}\}$ the DNN and BGRU models are trained using the following criteria

- RAND-SAMP-PA,
- CE+RAND-SAMP-PA,
- RAND-BATCH-PA,
- and CE+RAND-BATCH-PA.

Figures 6.5 and 6.6 show the WER curve of the CE+MIN-SAMP-PA criterion and the CE+RAND-SAMP-PA criterion for BGRU models with dropout 0.1, respectively.

Table 6.3 shows the result of different training criteria combinations for the various models. We highlighted the best results. In summary, using the combined CE+MIN-SAMP-PA, CE+MIN-BATCH-PA, CE+RAND-SAMP-PA, CE+RAND-BATCH-PA criteria for both DNN and BGRU models, result in a similar or lower WER on the DEV corpus, and a 0.2-0.4% WER improvement on the EVAL corpus. The combination of the new criteria with cross-entropy results in a 9-16% relative improvement on the EVAL corpus. For BGRU models, the influence of dropout in

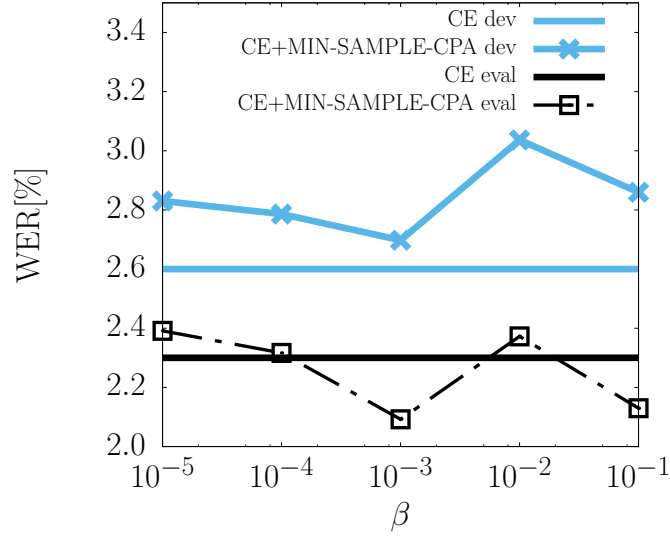


Figure 6.5: The $WER[\%]$ as a function of the constraint β for the $CE+MIN-SAMP-PA$ criterion for BGRU models with dropout 0.1 on the WSJ0 test corpora.

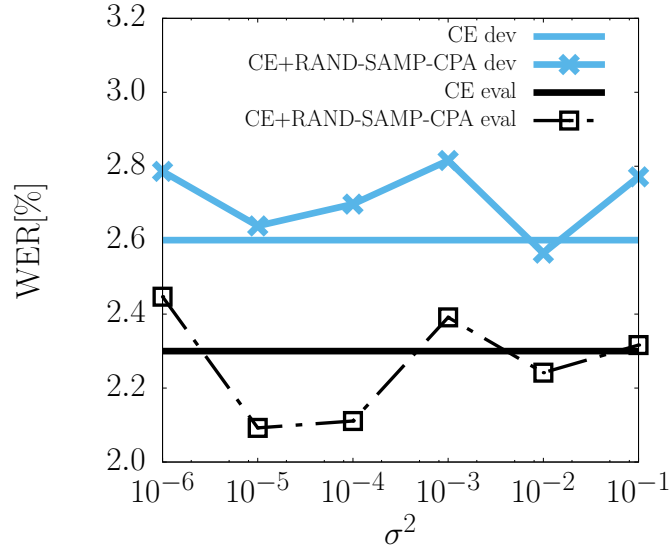


Figure 6.6: The $WER[\%]$ as a function of the variance σ^2 for the $CE+RAND-SAMP-PA$ criterion with $\mu = 10^{-6}$ for BGRU models with dropout 0.1 on the WSJ0 test corpora.

combination with the novel criteria has no major impact on the WER. However, we observed a small negative impact on the WER in cases where the combination does not include the CE criterion. Overall, the randomized criteria $CE+RAND-SAMP-PA$, and $CE+RAND-BATCH-PA$ perform more stably than the other criteria.

Conclusion Two novel regularization techniques were introduced based on new training criteria. Following a principled approach, we derived training criteria from f -Divergence bounds on the classification error difference between the model-based and *Bayes* decision rule. The first

technique minimizes over this family of training criteria. These criteria continuously change the functional form of the training criterion to avoid early local minima. The second technique dynamically changes the objective function by randomly drawing from a family of criteria according to a Gaussian distribution, which also avoids local minima. Both techniques successfully showed a WER improvement over the cross-entropy baseline criterion when tested in practical experiments on the WSJ-5K corpus for deep neural networks and bidirectional gated recurrent units, both in standalone implementations and in combination with dropout.

6.3 Joint Work

I modified these baseline systems to conduct the experiments for frame-wise training criteria. The main setups, baseline systems, theoretical derivations, and experiments are my individual contribution [Nußbaum-Thom & Tüske⁺ 12, Nußbaum-Thom & Cui⁺ 14, Nußbaum-Thom & Schlüter⁺ 17]. The IBM Babel speech team at the IBM Watson Yorktown Heights Research Center built the Babel baseline systems used in the experiments in Section 6.1. From the IBM Watson team, Xiaodong Cui, Jia Cui, Bhuvana Rhamabhadran, and Brian Kingsbury were highly involved in building the baselines.

7. APPLICATION TO SEQUENCE TRAINING FOR AUTOMATIC SPEECH RECOGNITION

This section is dedicated to discriminative sequence training, as introduced in Section 2.1. In the case of sequence training a class $c \in \mathcal{C}$ corresponds to a word sequence w_1^N and an observation $x \in \mathcal{X}$ corresponds to a feature sequence x_1^T .

We evaluate the WMMI and α -MPE training criteria from Section 5.2.1 for sequence-based training of Log-Linear Models on the TIMIT speech recognition task. If $\alpha \rightarrow 1$, the α -MPE criterion approaches the MPE criterion.

$$\mathcal{F}_{\alpha\text{-MPE}}(q) = \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) q^\alpha(c|x_n). \quad (\alpha\text{-MPE})$$

If $\alpha \rightarrow 0$, the α -MPE criterion approaches the WMMI criterion:

$$\mathcal{F}_{\text{WMMI}}(q) = -\frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} \mathcal{A}(c_n, c) \log q(c|x_n). \quad (\text{WMMI})$$

The α -MPE and WMMI criteria are evaluated with Log-Linear Mixture models.

7.1 Log-Linear Mixture Model

For automatic speech recognition, the classes c correspond to word sequences $w_1^N = w_1, \dots, w_N$. The notation of the corresponding HMM state sequences and features is $s_1^T = s_1, \dots, s_T$ and $x_1^T = x_1, \dots, x_T$. We choose the frame-wise phone accuracy [Gibson 08] as accuracy. As demanded by (4.16), this accuracy leads to a positive expected reward. We formulate the model-based posterior involving HMMs by

$$q(w_1^N | x_1^T) = \sum_{s_1^T: w_1^N} \frac{[q(x_1^T, s_1^T, w_1^N)]^\gamma}{Z(x_1^T | \gamma)}$$

with the posterior re-normalization $Z(x_1^T | \gamma)$ and the posterior scale γ . Given the vector representation of feature functions $f(x_1^T, s_1^T)$ and parameter set $\Lambda = \{\lambda\}$, the choice

$$q(x_1^T, s_1^T, w_1^N) = q(w_1^N) \exp(\lambda^T f(x_1^T, s_1^T))$$

results in a log-linear posterior model. In order to avoid over-fitting, an ℓ_2 regularization is used, centered around the generative maximum likelihood Gaussian mixture model. Only first-order features $f_{t,s,d}(x_1^T, s_1^T) = \delta(s, s_t) x_{t,d}$ are used in combination with zeroth-order features defined similarly. The training criteria are optimized using the gradient-based procedure using RPROP in a transducer-based framework.

7.2 Transducer-Based Framework Implementation

This section discusses the efficient calculation of the gradient of the α -MPE criterion by extending the Eisner expectation semiring [Eisner 01]. We implemented all criteria in the transducer-based framework of [Heigold & Deselaers⁺ 08, Heigold 10]. We start with an extension of the expectation semiring, which then calculates the gradient of the α -MPE and WMMI criterion.

α -scaled expectation semiring: The α -scaled expectation semiring is a multiplex semiring with weights $(p, v) \in \mathbb{R}^+ \times \mathbb{R}$, and

- $(p_1, v_1) \oplus (p_2, v_2) = (p_1 + p_2, v_1 + v_2)$,
- $(p_1, v_1) \otimes (p_2, v_2) = (p_1 p_2, p_1^\alpha v_2 + v_1 p_2^\alpha)$, and
- $\bar{1} = (1, 0), \bar{0} = (0, 0)$.

In addition, the inverse is defined to be $\text{inv}(p, v) = (p^{-1}, -vp^{-2\alpha})$. In terms of notations and definitions, we stick to [Heigold & Deselaers⁺ 08]. In particular, the expectation transducer is denoted by $R_{\mathcal{P}}[Z]$ for random variable Z and the probabilistic transducer $\mathcal{P}$. The expectation can be calculated by forward potentials α_q and backward potentials β_q at the state q (here q_0 denotes the initial state). The gradient computations of the α -MPE and WMMI objective functions base on this multiplex semiring.

7.3 Gradient of the Objective Function

Let $\mathcal{P}$ be the word lattice with the joint probabilities $p_{\Lambda}(x_1^T, s_1^T, w_1^N)$. This lattice is an acyclic transducer with probability semiring. The transducer $\mathcal{A}$ is the accuracy transducer corresponding to $\mathcal{P}$, having the same topology but different weights. Define transducer Z with the α -scaled expectation semiring and assign the weights $w_Z[a] = (w_{\mathcal{P}}[a], w_{\mathcal{P}}^\alpha[a] w_{\mathcal{A}}[a])$ to the arcs. Then, the gradient of the α -MPE criterion can be calculated by:

$$\nabla \mathcal{F}_{\alpha\text{-MPE}}(\Lambda) = \sum_{e \in \mathcal{P}} w_{\nabla \log \mathcal{P}}[e] \cdot \left(\frac{w_{E_{\mathcal{P}}[Z]}[e][v]}{\beta_{q_0}^\alpha[p]} - \frac{w_{E_{\mathcal{P}}[Z]}[e][p] \beta_{q_0}[v]}{\beta_{q_0}^{1+\alpha}[p]} \right).$$

Furthermore, if $\alpha \rightarrow 0$, the α -MPE gradient converges to the gradient of the WMMI criterion.

7.4 Experimental Setup

We test the criteria on the TIMIT phone recognition task. Table 7.1 presents the corpus statistics.

Table 7.1: *Corpus statistics.*

task	corpus	data [h]	#run. words	frames
TIMIT	Train	3.14	30 k	1 M
	Test	0.16	1,5 k	100 k

The acoustic front-end comprises 16 MFCC features. We concatenate feature vectors from nine consecutive frames, and a Linear Discriminant Analysis (LDA) is used to reduce the dimension to 33. At training time, the recognizer using the baseline GMM system using a unigram language model, acting as a weak margin, generated the word-conditioned lattices for discriminative

training. The language model scale γ during training was tuned on a hold-out set and kept fixed during training. The log-linear models were initialized with the ML model and trained with a gradient descent method using the RPROP algorithm. We chose the regularization and margin to the point where the WER started to increase rapidly. The GHMM baseline recognition system for TIMIT uses 114 mono phone states plus one silence state. We also use phoneme folding [Lee & Hon 89]. GMM models the emission probabilities with a total of about 20k densities, all sharing a single diagonal covariance matrix. The recognition uses a trigram phoneme language model.

7.5 Experimental Results

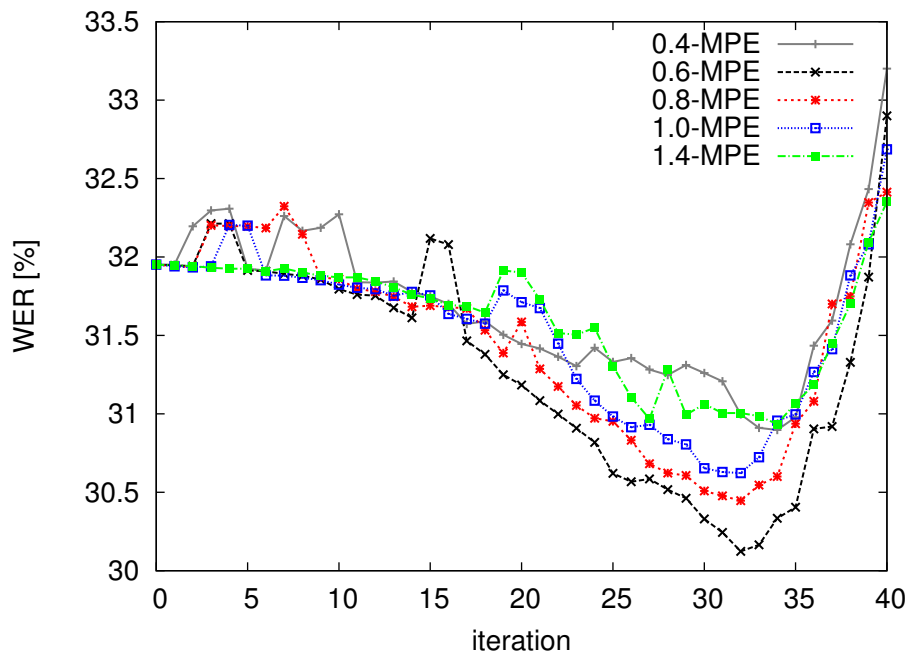


Figure 7.1: *WER progress of the α -MPE criterion for different α values on the TIMIT test set.*

Table 7.2: *WER[%] results for the ML and modified criteria on the TIMIT test set.*

criterion				
ML	MMI	MPE	WMMI	0.6-MPE
32.0	30.6	30.6	31.7	30.1

The experimental results in Table 7.2 indicate that the WMMI gives some improvement but performs worse than the MMI and MPE criteria on TIMIT. Both the MMI and MPE show a relative improvement of 5%. At the same time, the 0.6-MPE criterion performs even better than the modified MPE criterion with a relative improvement of 6%. We expected these results since MPE is identical to the 1.0-MPE criterion, which is a stable numerical approximation of the WMMI criterion. The WER progress of the α -MPE criteria in Figure 7.1 also confirms this result: The 0.6-MPE and 0.8-MPE criteria perform noticeably better than MPE. However, the 0.4-MPE criterion performs worse, probably because $\alpha = 0.4$ is closer to the numerically more unstable WMMI criterion with $\alpha = 0$.

Conclusion We introduced new error bounds on the global error mismatch. These bounds yield new discriminative training criteria based on a more general loss function similar to the Levenshtein distance in the case of ASR. Besides, the posterior-scaled MPE criterion, which is an approximation to one of the proposed criteria, was presented. This connection to global error bounds gives a better theoretical justification on the superior performance of the MPE criterion for the first time.

We discussed the implementation of the posterior-scaled MPE criterion in a transducer-based framework via a small modification of the Eisner expectation semiring. Experiments compared the posterior-scaled MPE criterion to other discriminative training criteria on an ASR task. Theoretical results were confirmed and they show that the posterior-scaled MPE criterion performs better than the state-of-the-art MPE criterion.

7.6 Joint Work

The work in this section builds on G. Heigolds transducer-based framework for sequence training of log-linear models [Heigold & Deselaers⁺ 08, Heigold 10]. Zoltan Tüske built the TIMIT baseline system. I solely built the WSJ baseline. I conducted all the experiments in this section.

8. GERMAN QUAERO EVALUATION CAMPAIGN

In this section, we present the results of our participation in the international QUAERO automatic speech recognition evaluation campaign from 2009 through 2013. The competing teams in this evaluation campaign were the University of Karlsruhe (UKA) and Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingenieur (LIMSI). Quaero is a large vocabulary task with a focus on transcribing web data. The data includes speech types like web data, European Parliament plenary sessions and Broadcast Conversation (BC), Broadcast News (BN), cooking sessions, interviews, and talk-shows. Recognition on the data is challenging because of considerable variability in the acoustic conditions, and a large portion includes spontaneous speech.

A particular challenge is that the project provided almost no in-domain training data, and the test data contains a large variety of speech types. German development and evaluation data sum up to 12 hours in 2009 and 50 hours in the remaining time of the project. Another challenge is that a training-testing-mismatch exists since only a small amount of official training data is available from the test domain.

Overall, the RWTH participated in this project for German with the best or competitive results. The novel RWTH segmenter determines the segmentation of an audio stream based on a log-linear segment model [Rybach & Gollan⁺ 09]. As described in [Hoffmeister & Plahl⁺ 07, Löff & Gollan⁺ 07], the German systems consist of several subsystems, that differ in the features or the language models used. We used a four or five pass strategy with a 4-gram decoder in the different phases of the evaluation campaign. We apply a fast Vocal Tract Length Normalization (VTLN) in the first pass, Constrained Maximum Likelihood Linear Regression (CMLLR), and Maximum Likelihood Linear Regression (MLLR) in the second pass. Depending on the system, either language model rescoring or cross-adaptation is applied. Finally, we combine all subsystems by a confusion network system combination. Contributing to the enhancements in different years of the evaluation phase are the systematic use of hierarchical neural network based posterior features, system combination, speaker adaptation, cross speaker adaptation, and the usage of additional training data.

The sections organize as follows: In Section 8.1 we describe our training data, the creation of the lexica, and the language modeling. Next, in Section 8.2, the acoustic models, speaker normalization, and adaptation techniques are presented. Section 8.3 describes the Log-Linear model segmenter, Section 8.4 and Section 8.5 describe exemplarily the system development in 2009 and 2010. The remaining years use more refined versions of the 2009 and 2010 systems. Finally, Section 8.6 describes the evaluation results of all teams throughout the project.

8.1 Language Resources

The development and evaluation data of the QUAERO evaluation campaign consist of data from three domains. While the majority of the data is from the web (WEB) data, or broadcast news, some European Parliament plenary sessions (EPPS) are also covered. The evaluation proceeded

in an open condition, where all the training data before 2008 is allowed. Furthermore, in-domain training data is provided for German.

8.1.1 Training Data

The project provided different amounts of transcribed in-domain training data throughout the years of the evaluation campaign. While in the 2008 dry-run, no data were provided in 2009, the WEB08 and EPPS09 corpus and in 2010 the QUAERO-AM-09 corpus were made available to all partners for system development for the remaining time. As shown in Table 8.1, the WEB08 corpus covers podcast data from various domains like comedy, seminar, report, and interview, whereas the EPPS08 corpus consists of speeches from the European Parliament only.

Table 8.1: *In-domain transcribed audio data for German.*

corpus	duration [h]
WEB08	14
EPPS08	5
QUAERO-AM-09 ($\supseteq$ WEB08 + EPPS08)	50

Due to the small amount of in-domain training data, we also use common training data, which is summarized in Table 8.2 and considers the cut-off date.

Table 8.2: *Additional transcribed audio data for the German system available for acoustic modeling.*

corpus	duration [h]	# segments	# running words
Verbmobil	63	36440	736,058
WDR	79	42011	625,018
Report Mainz	11	8928	100,641
Zeit	171.5	329384	2,500,866

The *Verbmobil* corpus recorded in the *Verbmobil* project consists of dialogues for travel appointments [Kanthak & Sixtus⁺ 00]. In contrast, the audio material of the *WDR* and *Report Mainz* corpora is from the BN domain [Macherey & Ney 02]. Furthermore, RWTH has downloaded read articles from the newspaper *Zeit* for which almost correct transcripts are available. A reasonable segmentation is generated by aligning the transcripts to the audio data. Segment boundaries are introduced on silence chunks not shorter than 35 seconds.

8.1.2 Lexicon Modeling

The recognition vocabulary is derived from the text data described in Table 8.3. The text data is cleaned up and normalized by a manually defined set of rules and semi-automatic methods. While in the years 2008 and 2009 the lexicon consisted of most frequent 100k words, in the remaining years either the most 300k or a sub-lexical approach was used. This is described in more detail in Section 8.1.3. For words, where no pronunciation has been available, the pronunciations are generated by the statistical grapheme-to-phoneme (g2p) conversion toolkit [Bisani & Ney 08]. The source pronunciations for German are obtained from the German LC-STAR lexicon.

8.1.3 Language Modeling

Based on the available training data, 4-gram language models (LMs) were estimated using [Stolcke 02], smoothed by the Modified Kneser-Ney method. We partitioned the LM data into blocks,

Table 8.3: Language model training data

Corpus	# running words	type
QUAERO-LM-2010	228 M	blog+news
TAZ	151M	news
German-news	155M	news
transcriptions	0.5 M	BC+BN

estimating n -gram probabilities for each block individually. Then the LMs were linearly interpolated while optimizing the perplexity on a holdout data set.

Table 8.3 gives an overview of the training material that was used. From the 2010 evaluation onwards, substantial amounts of in-domain text data downloaded from web blogs were distributed to all participants.

A wide lexical variety is characteristic for the German language, as a large number of distinct lexical forms can be generated by derivation, compounding, and inflection. For this reason, an alternative LM approach based on word decomposition into sub-lexical fragments was investigated from 2010 on to reduce OOV rates and minimize perplexity. This method was shown to have good performance in Arabic, which also is a morphologically rich language.

Words were decomposed using a statistical data-driven tool. Then a standard n -gram LM was estimated on the decomposed text. To prevent confusion between the most important German words and other less-frequent fragments, the most frequent decomposable words were excluded from decomposition as proposed in [El-Desoky Mousa & Gollan⁺ 09]. In spite of the huge vocabulary size for German (300 K), the OOV rate was still at 1.13 %.

8.2 Acoustic Modeling

The German systems is composed of several subsystems that use either MFCC or PLP as base features. For each feature type, a segment-wise mean and variance normalization is applied and fed into a sliding window of length nine. All feature vectors in the window are concatenated and projected to a 45-dimensional feature space by applying linear discriminant analysis. The feature vector is augmented with a voicedness feature and phone posterior features, estimated using a multilayer perceptron. The hierarchical neural network (HMRSTA) is trained using the phonemes of the given language based on MRSTA features [Valente & Hermansky 08, Plahl & Schlüter⁺ 10], as estimated on a phone alignment. The dimensionality of the phone posterior features is reduced using a principal component analysis.

For the sake of simplicity, we will refer to the MFCC augmented by a voicedness and MLP features as MFCC+voiced+MLPs and PLP as PLP+voiced+MLP features later on.

Acoustic models for all systems are across word triphone left-to-right hidden markov models based on Gaussian mixtures with the globally pooled diagonal covariance matrix. The system uses a 3-state HMM. A number of 4500 generalized triphone states define our HMM states. The baseline acoustic models (AMs) are trained using maximum likelihood (ML)/ *Viterbi* on the available training data. The resulting models comprise about 1M Gaussians with a globally pooled covariance matrix.

8.2.1 Speaker Normalization and Adaptation

All systems use the same approach for speaker normalization and adaptation. Vocal Tract Length Normalization (VTLN) is applied to the filterbank within the MFCC or PLP extraction both in training and testing. In recognition a fast one pass VTLN approach is used, where the

Training data		Structure
DE ₁	344 h BC+BN+read speech	H-MLP
DE ₂	Quaero	HB-MLP

Table 8.4: MLP feature training data and network topology

warping factor is estimated using a Gaussian mixture classifier, trained on the acoustic training corpora.

Speaker adaptive training (SAT) based on Constrained Maximum Likelihood Linear Regression (CMLLR) [Gales 98] is applied to compensate for speaker variation in both training and testing. For the German system, the Maximum Likelihood Linear Regression (MLLR) is applied to the means of the Gaussian mixtures during recognition.

Both CMLLR and MLLR are text-dependent and need a two-pass setup. They are carried out in a speaker-dependent manner, and no speaker identities are provided in the evaluation, as a consequence an automatic speaker labeling is performed. To provide a speaker labeling for SAT, a generalized likelihood ratio based segment clustering with a Bayesian information criterion-based stopping condition is applied to the segmented training and recognition corpus [Chen & Gopalakrishnan 98]. In the third pass, the subsystems are cross-adapted to each other.

8.2.2 MLP features

Showing significant improvements, MLP features for acoustic modeling have attracted much attention during the last years. Considerable effort has been spent on improving the input features of the neural networks as well as on optimizing the network topology itself.

The neural networks were trained by feeding the Multi-resolution RASTA (MRASTA) features as inputs and the phone-posterior probabilities, computed by a forced alignment of the acoustic training data, as desired outputs. As the last step, the MLP features were decorrelated by a PCA transformation. This also allows an additional dimensionality reduction in the case of the H-MLPs. Table 8.4 lists the different MLP feature types and the corresponding acoustic data that were used for training.

Subsystems also improved when MLP features across languages (though estimated on different phone sets) were exchanged. This property, first reported in [Stolcke & yuh Hwang⁺ 06], was made use of to increase subsystem variability.

8.2.3 Discriminative training

To sharpen acoustic models, discriminative training was applied. Lattices were computed using the current best acoustic models where only a unigram language model was used to avoid overfitting. Based on these lattices the MPE training criterion was optimized [Povey & Woodland 02b].

8.3 Segmentation

The audio segmentation of the RWTH system makes use of a log-linear classifier that decides if a time frame corresponds to a segment boundary or not. The features for the log-linear model were chosen as to cover, e.g., the variability of the acoustic signal, the speaker homogeneity.

Also, the set of log-linear features was augmented by information obtained by a one-pass recognition on the unsegmented audio data. This includes the number of words within a segment as well as the time stamps of sentence boundary tokens hypothesized by the recognizer.

Finally, an HMM-based classifier was used to detect speech/non-speech, music, and noise segments. From this information additional features were derived and included in the log-linear segment boundary classifier. Further details on the RWTH segmentation software can be found in [Rybach & Gollan⁺ 09].

We adapted the segmenter to the BC+BN task using the German Quaero acoustic training data. By filtering the audio data for non-speech parts, 9.5 h were retained for training the HMM-based classifier.

8.4 System Development in 2009

All German AMs are trained on the WEB08, and EPPS09 data mentioned in Tables 8.1 and 8.2. The German system is subdivided into two subsystems. The main difference of these two subsystems originates from the features used to train the systems. The German subsystem g1 utilizes an MFCC+voiced+MLP, whereas the subsystem g2 uses a PLP+voiced+MLP front-end. Both subsystems use the same language model, which was estimated on the text data described in Table 8.3. Table 8.5 gives the data, a number of running words, vocabulary size, perplexities of the final LMs, and OOV rates on the German dev09 and eval09 data sets.

Table 8.5: *Statistics for the dev09 and eval09 corpora for German.*

corpus domain	German					
	dev09			eval09		
domain	web	BN	EPPS	web	BN	EPPS
dur. [h]	7.4	0.0	0.38	2.95	0.46	0.42
run. wrds	68k			36k		
vocab	10k			6k		
PP	394			353		
OOV [%]	5.0			4.79		

8.4.1 Recognition Setup

For all languages and AMs, the first pass is realized by a 4-gram Viterbi decoder using a fast-VTLN normalization. In the second pass, a SAT/CMLLR recognition is applied where the statistics for adaptation are collected from the first pass output. The German system also uses MLLR. Figure 8.1 shows the framework for the German system.

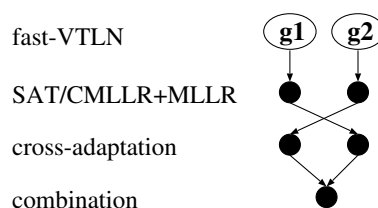


Figure 8.1: *Schematic diagram of the decoding framework for German.*

The German subsystems are cross adapted to each other in the third pass. In the last pass, the German subsystems are combined using system combination.

8.4.2 Experiments

For parameter optimization, the 2009 development sets are used. Statistics are given in Table 8.5. Table 8.6 summarizes the recognition results for the methods applied in the 2009 evaluation for German on dev09 and eval09 sets. The recognition results of the previous year’s baseline system are also given, for comparison.

The German system achieves a significant improvement compared to the baseline system. The baseline system is the second pass of system g1 without the use of MLP features and the acoustic training data from the *WEB08*, *EPPS08*, and the *Zeit* corpora. The application of MLP features and additional audio data leads to an improvement of 26.1% WER relative on dev09 and 31.1% WER relative on eval09, compared to the baseline system. However, due to the use of additional training data, the gain of the MLP features is not separable, and it seems that the improvements are too promising because of a not well-tuned baseline system. Cross-adaptation and system combination give only a small improvement, probably due to the small number of subsystems which is cross adapted and combined.

Table 8.6: *Results on the German dev09 and eval09 corpus.*

corpus	dev09		eval09	
baseline	44.0	-	40.9	-
subsystem	g1	g2	g1	g2
fast-VTLN+MLP	37.7	37.7	34.2	34.3
+SAT/CMLLR+MLLR	34.9	35.0	31.0	31.2
+cross-adaptation	34.6	34.9	30.5	30.5
+combination	33.3		30.0	

8.5 System Development in 2010

The recognition of broadcast conversation data also necessitates updated and improved pronunciation lexica, as speakers tend to pronounce words less carefully in conversational speech. This problem was addressed by introducing automatically generated pronunciation variants into the training lexicon. A large number of pronunciation variants were derived from those systematically found in the baseline lexicon by the deletion of phonemes. After computing a forced alignment using the updated lexicon, only those pronunciations were retained for future training and recognition that were seen in training at least a certain number of times. By this technique, the WER was reduced from 21.4 % by 0.4 % absolute.

8.5.1 Recognition setup

The RWTH systems rely on five subsequent recognition passes, as depicted in Figure 8.2. In an initial unadapted pass, a first transcription was obtained, which formed the basis for the second, CMLLR-adapted recognition pass.

The resulting transcriptions then were exchanged between subsystems for cross-adaptation, leading to the third full recognition pass. In the case of English and French, for cross-adaptation, only the previous recognition output of one subsystem was used, whereas, for German all subsystems were used for CMLLR-adaptation as indicated by the dashed lines.

As for the full recognition, only a pruned LM was used, third pass lattices were rescored using the full LM. System combination was applied as the last step. The lattices were converted to a confusion network (CN) by an iterative procedure. Afterward, the final transcription was obtained by CN combination as presented in [Evermann & Woodland 00a]. In the case of the

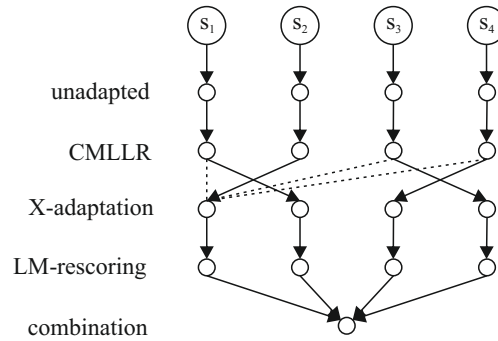


Figure 8.2: Schematic view of the RWTH recognition process

German system, a post-processing step was also added to concatenate numbers to avoid spelling errors.

8.5.2 Experiments

Table 8.7 shows detailed WER results for the individual subsystems. The acoustic training was kept fixed always using all material from Table 8.1.

Table 8.7: Results on the 2010 development and evaluation set.

	DE		
	STF	MLP	WER
s ₁	MFCC	DE ₁	20.6
s ₂	PLP	DE ₁	20.3
s ₃	MFCC	DE ₂	19.8
s ₄	PLP	DE ₂	20.3 %
CNC	17.3 %		
eval10	16.9 % (10.6 %)		

The results from the table show that the new HBN-MLP features gave a significant gain in WER compared to the H-MLP features, as can be seen, e.g., in the performance of the German s₁ and s₃ systems. Furthermore, it seems that the quality of the MLP features strongly depends on the availability of large amounts of training data. Additional conclusions can be drawn from the relative improvements, as summarized in Table 8.8, that also gives baseline results for the 2009 systems ([Nußbaum-Thom & Wiesler⁺ 10b]). Gains by MPE were quite small, which we assume to be caused by the incorporation of MLP features that already form a discriminative training approach.

For the German language, the sublexical language modeling approach yielded an improvement from 22.5 % to 21.7 % in WER after CMLLR adaptation. All systems could be improved significantly compared to the previous year's systems, which was also due to the availability of more in-domain training data distributed to all evaluation participants.

8.6 Evaluations

In this section, we present the results for the participating teams throughout all periods of the evaluation campaign. The systems architecture following up to the years 2009 and 2010 remained relatively similar, but the individual components were refined, e.g., multi-lingual MLP features

Table 8.8: Improvements in terms of WER obtained by different methods, measured on the 2010 development corpus

	DE (s_1)
baseline	26.2 %
VTLN	21.9 %
+CMLLR	20.2 %
+MPE	19.9 %
+X-adaptation	19.9 %
+LM-rescoring	19.8 %

Table 8.9: Improvements in terms of WER obtained by different partners for the German system from 2008 through 2013.

project period	year	test data [h]	WER[%]		
			RWTH	T1	T2
P2	2009	3.83	30.40	39.40	32.70
P3	2010	3.83	16.94	24.14	21.05
P4	2011	3.23	17.49	17.40	18.04
P5	2012	3.13	18.71	19.38	19.63
P6	2013	3.24	13.53	15.95	-

were trained. In Table 8.9, the evaluation results for different project periods are shown. For discreteness, the names of the competing teams were anonymized to T1 and T2. I participated in all years of the project with the best or competing results for German.

8.7 Joint Work

In the QUAERO campaign from 2008 through 2013 the language models were trained by Amr Mousa, the RWTH segmenter was implemented and trained by David Rybach, the MLP features were trained by Christian Plahl and Zoltan Tüske, the system combination framework was provided by Björn Hoffmeister. My individual contributions to the German system are the training of the acoustic model, the experimental choice of system architecture, the tuning of the overall architecture and subsystems, the decoding of all subsystems during training and in the evaluation phase.

8.8 Acknowledgements

This work was partly realized under the Quaero Programme, funded by OSEO, French State agency for innovation.

9. SCIENTIFIC ACHIEVEMENTS

In this section, we discuss the achievement of the scientific goals formulated in Section 3. The goal of this thesis was to investigate novel relations between the evaluation measure, decision rule and training criteria, i.e., to find novel error bounds and derive training criteria from the error bounds. Our interest was to find bounds and criteria not only for decision problems based on the 0-1 loss but also for those based on more general losses — like the *Levenshtein* loss. The case of the *Levenshtein* loss is especially relevant to automatic speech recognition where a mismatch exists between performance measure and decision rule. In the case of the existence of new criteria, we evaluate these for automatic speech recognition for frame-wise and discriminative sequence training on real-life data.

Error Bounds The *Kullback-Leibler* divergence has been established as an upper bound on the variational distance between the model-based and true distribution in multiple publications [Cover & Thomas 06, p.369], [Vapnik 98, p.30], [Fedotov & Harremos⁺ 03, Ney 03, Guntuboyina 11]. A proof in [Ney 03] shows that the variational distance upper bounds the classification error difference between the model-based and true decision rule. Therefore the *Kullback-Leibler* divergence also become bounds on this classification error difference. So far, such bounds were restricted only to the *Kullback-Leibler* divergence or the squared error bounds (as in [Ney 03]) and to decision problems based on the 0-1 loss.

In our work, we found novel classification error bounds on the *f-Divergence* in Section 4.1. Also, in Section 4.1, two novel proofs of these classification error bounds were presented. Different from the previous bounds in [Cover & Thomas 06, p.369], [Vapnik 98, p.30], [Fedotov & Harremos⁺ 03, Ney 03, Guntuboyina 11], our bounds are tight with the classification error difference. Another novelty is: For the first time, we reported in this thesis error bounds for decision problems based on more general losses — including the *Levenshtein* loss. In Section 4.1.4, we extended the classification error bounds for decision problems based on the 0-1 loss to more general losses. For deriving training criteria, we also managed to find proper explicit bounds on the squared error difference for both decision problems based on the 0-1 loss and more general losses in Section 4.1.4 and Section 4.2.5, respectively.

Error-Bound-Based Training Criteria The author in [Ney 03] derived the cross-entropy criterion from the *Kullback-Leibler* error bound for decision problems based on the 0-1 loss. To this point, a scheme for criteria derived from error bounds different from the *Kullback-Leibler* divergence did not exist. However, in this work, we presented a scheme in Section 5.1, which derived novel discriminative training criteria from explicit classification error bounds based on the *f-Divergence*. This scheme introduces a novel family of discriminative training criteria based on the *f-Divergence*.

We also answered the open question of the existence of a similar derivation scheme for decision problems based on a more general loss — like the *Levenshtein* loss. In this case, the training criterion is error-based, like the MPE criterion in automatic speech recognition. In Section 5.2,

we introduced such a derivation scheme to derive error-based criteria from the explicit error bounds in Section 4.2.5. The decision problems addressed in this derivation scheme also cover the case of the *Levenshtein* loss. Therefore, these training criteria are relevant to the mismatch in automatic speech recognition.

It is common knowledge that the non-parametric solution of the model trained with cross-entropy using unlimited training data converges to the posterior of the true distribution. This convergence means under ideal conditions; the model learns a posterior such that the model-based decision rule is identical with the *Bayes* decision rule based on the 0-1 loss.

As promised in Section 5.2, we analyzed the convergence of the non-parametric solution of the model trained with our novel training criteria in case of infinite training data. We found such an analysis for the power-approximation training criterion. We showed that if the criterion uses the 0-1 loss, then the non-parametric solution of the model converges to a monotonic transformation of the posterior of the true distribution. Here, similar to the case of the cross-entropy criterion, the model-based decision rule is identical to the *Bayes* decision rule. We also showed that if the criterion uses a more general loss, then the non-parametric solution of the model converges to a monotonic transformation of the true reward. This result also means that in this case of a more general loss, the model learns a posterior that results in an error-optimal decision rule. Such a result is satisfying because it also covers the case of automatic speech recognition and addresses the mismatch between performance measure and maximum probability decision rule. In practice, it answers the question of what a model trained with MPE criterion tries to achieve: The MPE criterion learns a model posterior which intends to minimize the posterior expected *Levenshtein* loss when applied in the maximum probability decision rule. So from a theoretical standpoint, in automatic speech recognition, the mismatch between the *Levenshtein* loss used in the performance measure — the word error rate — and the 0-1 loss used in the maximum probability decision rule can be resolved by using an error-based training criterion — like MPE.

Application to Frame-Wise Neural Network Training for Automatic Speech Recognition According to the goals set out in Section 5.1, we evaluated the novel training criteria in Section 6 in practice for frame-wise discriminative training of neural network acoustic models. We trained neural networks based on our novel training criteria and evaluated those in automatic speech recognition experiments on real-life data. This evaluation resulted in a better or competitive performance compared to the cross-entropy criterion.

Section 6.1 combined the cross-entropy criterion with our novel *Lin* and power-approximation criteria. In the IARPA BABEL project, we trained acoustic feed-forward neural networks for the low-resourced Vietnamese and Bengali languages. For both languages, a combination of our novel criteria with cross-entropy outperformed the standalone cross-entropy criterion in terms of word error rate on all test sets.

In Section 6.2, two novel approaches were evaluated based on the power-approximation criterion. The first approach minimized the power-approximation criterion over the hyper-parameter of the criterion and, therefore, also minimized over the respective power-approximation divergence bounds. The second approach dynamically changed the hyper-parameter of the criterion according to a *Gaussian* distribution in order to avoid getting stuck in local optima. Both approaches were evaluated as standalone as well as in combination with the cross-entropy criterion on the well-known WSJ0 corpus to train acoustic feed-forward and recurrent neural network models. All novel approaches achieved a better or competitive result compared to the cross-entropy criterion. However, the best results, of 9-16% relative word error rate improvement, were achieved in combination with the cross-entropy criterion. To the best of our knowledge our approach achieves the best results reported on the WSJ0 test sets so far: The respective experiments resulted in a word error rate of 2.6 on the development and 2.0 on the evaluation set, respectively.

Application to Sequence Training for Automatic Speech Recognition According to the goals set out in Section 5.1, we evaluated the novel criteria from Section 5.2 in practice for discriminative training of acoustic log-linear models. To this end, we implemented the novel α -MPE and weighted MMI in the RWTH discriminative transducer-based training framework. We trained acoustic log-linear models based on our novel criteria and evaluated the models in automatic speech recognition experiments on the TIMIT data set. Our novel criteria achieved a 6% relative word error rate improvement over the state-of-the-art MPE criterion.

German QUAERO Evaluation Campaign We presented our successful participation in the German QUAERO evaluation campaign over the period of six years.

10. APPENDIX

This section extends abbreviated topics of previous sections.

10.1 On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition

In [Ney 03], the *Kullback-Leibler* divergence was shown to be an upper bound on the classification error difference. Also, in the same work, a framework is presented on how to derive empirical training criteria from error bounds. Like in our framework first, the notion of the *Bayes* decision rule and error for the case of the 0-1 loss is developed, which then accordingly, is extended to the model-based case. In [Ney 03], local error bounds are derived starting from the local error difference by deriving fundamental inequalities between the local *Bayes* and model-based classification error.

$$\begin{aligned}
\Delta_{0-1}^{pr}(x) &= (1 - pr(c_{0-1}^q(x)|x)) - (1 - pr(c_{0-1}^{pr}(x)|x)) \\
&= pr(c_{0-1}^{pr}(x)|x) - pr(c_{0-1}^q(x)|x) \\
&\leq pr(c_{0-1}^{pr}(x)|x) - pr(c_{0-1}^q(x)|x) + q(c_{0-1}^q(x)|x) - q(c_{0-1}^{pr}(x)|x) \\
&= q(c_{0-1}^q(x)|x) - pr(c_{0-1}^q(x)|x) + pr(c_{0-1}^{pr}(x)|x) - q(c_{0-1}^{pr}(x)|x) \\
&\leq |q(c_{0-1}^{pr}(x)|x) - pr(c_{0-1}^{pr}(x)|x)| + |pr(c_{0-1}^q(x)|x) - q(c_{0-1}^q(x)|x)|.
\end{aligned}$$

Subsequently, those bounds are widened using relations to the $L_{\inf}$ and L_2 norm.

$$\Delta_{0-1}^{pr}(x) \leq \sum_{c \in \mathcal{C}} |q(c|x) - pr(c|x)| \quad (10.1)$$

$$\Delta_{0-1}^{pr}(x) \leq 2 \max_{c \in \mathcal{C}} \{|q(c|x) - pr(c|x)|\} \leq 2 \sqrt{\sum_{c \in \mathcal{C}} (q(c|x) - pr(c|x))^2}. \quad (10.2)$$

The global error bound is derived by computing the expectation from these local error bounds denoted as $g(x)$ based on the true distribution of the observations $pr(x)$.

$$\Delta_{0-1}^{pr} \leq \int pr(x) g(x) \, dx.$$

The expectation inequality for $E(X^2) \geq E^2(X)$ for random variables X leads to the Squared-Error bound using $g(x) = 2\sqrt{\sum_{c \in \mathcal{C}} (q(c|x) - pr(c|x))^2}$:

$$[\Delta_{0-1}^{pr}]^2 \leq 4 \int pr(x) \sum_{c \in \mathcal{C}} (q(c|x) - pr(c|x))^2 \, dx.$$

Using the *Pinsker* inequality on the previously derived inequality

$$\frac{1}{2} \sum_{c \in \mathcal{C}} |q(c|x) - pr(c|x)| \leq - \sum_{c \in \mathcal{C}} pr(c|x) \log \frac{q(c|x)}{pr(c|x)}.$$

leads to the *Kullback-Leibler* bound

$$[\Delta_{0-1}^{pr}]^2 \leq -2 \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \frac{q(c|x)}{pr(c|x)} dx.$$

These error bounds are tight if the model-distribution equals the *true distribution*.

This substitution results in case of the Squared-Error bound in the squared-Error (SE) criterion,

$$\mathcal{F}_{SE}(\Lambda) = \frac{1}{N} \sum_{n=1}^N \sum_{c \in \mathcal{C}} (q(c|x_n) - \delta(c, c_n))^2$$

and in case of the *Kullback-Leibler* bound in the cross-entropy or MMI criterion

$$\mathcal{F}_{MMI}(\Lambda) = -\frac{1}{N} \sum_{n=1}^N \log q(c_n|x_n).$$

10.2 Statistical Learning Theory [Vapnik 98, p.30]

In machine learning, the Bretagnolle-Huber bound is known in the context of density estimation. The central value to be bounded in this reviewed work is the variational mismatch between the joint true and model distribution.

$$V := \int \sum_{c \in \mathcal{C}} |pr(x, c) - q(x, c)| dx. \quad (10.3)$$

A simple relation exists to the error difference: The relation to the global error difference can be established by integrating over (10.1).

$$\begin{aligned} \Delta_{0-1}^{pr} &= \int [pr(x, c_{0-1}(x)) - q(x, c_{0-1}^q(x))] dx \\ &\leq \int \sum_{c \in \mathcal{C}} |pr(x, c) - q(x, c)| dx \\ &= V. \end{aligned}$$

The following relation to the global error mismatch, the variational mismatch, and the *Kullback-Leibler* distance can be established.

$$[\Delta_{0-1}^{pr}]^2 \leq V^2 \leq 4(1 - \exp(-D_{KL}(pr||q))).$$

The proof for this inequality is summarized in the following:

$$\begin{aligned}
 -D_{\text{KL}}(pr||q) &= \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \log \frac{q(c|x)}{pr(c|x)} dx \\
 &= \int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \left(\log \min \left\{ \frac{q(c|x)}{pr(c|x)}, 1 \right\} + \log \max \left\{ \frac{q(c|x)}{pr(c|x)}, 1 \right\} \right) dx \\
 &\leq \log \left(\int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \min \left\{ \frac{q(c|x)}{pr(c|x)}, 1 \right\} dx \right) \\
 &\quad + \log \left(\int pr(x) \sum_{c \in \mathcal{C}} pr(c|x) \max \left\{ \frac{q(c|x)}{pr(c|x)}, 1 \right\} dx \right) \quad \text{Jensen's inequality} \\
 &= \log \left(\int pr(x) \sum_{c \in \mathcal{C}} \min \{q(c|x), pr(c|x)\} dx \right) \\
 &\quad + \log \left(\int pr(x) \sum_{c \in \mathcal{C}} \max \{q(c|x), pr(c|x)\} dx \right). \tag{10.4}
 \end{aligned}$$

Now, notice that we have the following relationships.

$$\begin{aligned}
 \min(a, b) &= \frac{a+b}{2} - \frac{|a-b|}{2}, \text{ and} \\
 \max(a, b) &= \frac{a+b}{2} + \frac{|a-b|}{2}.
 \end{aligned}$$

Thus, we can continue (10.4) as follows.

$$\begin{aligned}
 &\log \left(\int pr(x) \sum_{c \in \mathcal{C}} \min \{q(c|x), pr(c|x)\} dx \right) \\
 &+ \log \left(\int pr(x) \sum_{c \in \mathcal{C}} \max \{q(c|x), pr(c|x)\} dx \right) \\
 &= \log \left(\int pr(x) \sum_{c \in \mathcal{C}} \left(\frac{q(c|x) + pr(c|x)}{2} - \frac{|q(c|x) - pr(c|x)|}{2} \right) dx \right) \\
 &+ \log \left(\int pr(x) \sum_{c \in \mathcal{C}} \left(\frac{q(c|x) + pr(c|x)}{2} + \frac{|q(c|x) - pr(c|x)|}{2} \right) dx \right) \tag{10.5}
 \end{aligned}$$

$$\begin{aligned}
 &= \log \left(1 - \int pr(x) \sum_{c \in \mathcal{C}} \frac{|q(c|x) - pr(c|x)|}{2} dx \right) \\
 &+ \log \left(1 + \int pr(x) \sum_{c \in \mathcal{C}} \frac{|q(c|x) - pr(c|x)|}{2} dx \right) \tag{10.6}
 \end{aligned}$$

$$\begin{aligned}
 &= \log \left(1 - \frac{1}{2}V \right) + \log \left(1 + \frac{1}{2}V \right) \\
 &= \log \left(\left(1 - \frac{1}{2}V \right) \left(1 + \frac{1}{2}V \right) \right) \\
 &= \log \left(1 - \frac{1}{4}V^2 \right). \tag{10.7}
 \end{aligned}$$

The equation in (10.6) is derived from (10.5) using the relation:

$$\int pr(x) \sum_{c \in \mathcal{C}} \frac{pr(c|x) + q(c|x)}{2} dx = 1.$$

Reformulating (10.7) results in the desired bound on the variational distance between the true and model distribution – the Bretagnolle-Huber bound:

$$\begin{aligned} -D_{\text{KL}}(pr||q) &\leq \log \left(1 - \left(\frac{V}{2} \right)^2 \right) \\ \Leftrightarrow \exp(-D_{\text{KL}}(pr||q)) &\leq 1 - \frac{1}{4}V^2 \\ \Leftrightarrow V^2 &\leq 4(1 - \exp(-D_{\text{KL}}(pr||q))). \end{aligned}$$

As pointed out in (10.1), the local difference is bounded by the sum over all classes of the absolute difference between the true and model-based posterior.

$$\Delta_{0-1}^{pr}(x) \leq \sum_{c \in \mathcal{C}} |q(c|x) - pr(c|x)|.$$

Therefore the Bretagnolle-Huber bound is also a bound on the error difference:

$$[\Delta_{0-1}^{pr}]^2 \leq V^2 \leq 4(1 - \exp(-D_{\text{KL}}(pr||p))).$$

10.3 Automatic Speech Recognition is NP-Complete

In complexity theory, according to [Hromkovic 01], an optimization problem is NP-complete if the corresponding decision problem is NP-complete as well. The decision problem for strings $c \in \Sigma^*$ based on the finite alphabet Σ for the non-trivial loss related Bayes decision rule can be formulated by:

BAYES:

Input Consider constants $I, m \in \mathbb{N}$ and strings $c_1, \dots, c_I \in \Sigma^*$ such that:

- $0 < pr(c_i|x) \leq 1$ for $i = 1, \dots, I$,
- $pr(c|x) = 0$ for $c \notin \{c_1, \dots, c_I\}$, and
- $\sum_{c \in \Sigma^*} pr(c|x) = \sum_{i=1}^I pr(c_i|x) = 1$.

Question Does a string $c \in \Sigma^*$ exist such that $E_{\mathcal{L}}(c|x) \leq m$?

First, BAYES is shown to be in NP. Consider a string c having a positive answer to the BAYES question:

$$\begin{aligned} m &\geq E_{\mathcal{L}}(c|x) = \sum_{i=1}^I pr(c_i|x) \mathcal{L}(c_i, c) \\ &\geq pr(c_{0-1}(x)|x) \mathcal{L}(c_{0-1}(x), c) \\ &\geq \frac{1}{I} \mathcal{L}(c_{0-1}(x), c) \\ &\Leftrightarrow \mathcal{L}(c_{0-1}(x), c) \leq \frac{m}{pr(c_{0-1}(x)|x)} \leq Im. \end{aligned}$$

Second, the strings c can only answer the question positively if $\mathcal{L}(c_{0-1}(x), c) \leq Im$. A non-deterministic polynomial calculation is formulated by guessing non-deterministic $j \leq Im$, generating non-deterministic string $d \in \Sigma^j$ and checking $E_{\mathcal{L}}(d|x) \leq m$. The computation of $E_{\mathcal{L}}(d|x)$ can be performed in polynomial time. In [de la Higuera & Casacuberta 00] MEDIAN-STRING is shown to be NP-complete:

MEDIAN-STRING:

Input Consider constants $I, m \in \mathbb{N}$ and strings $c_1, \dots, c_I \in \Sigma^*$.

Question Does $c \in \Sigma^*$ exist such that $\sum_{i=1}^I \mathcal{L}(c_i, c) \leq m$?

Finally, BAYES is shown to be NP-hard by defining a reduction to MEDIAN-STRING. Consider an input instance $(c_i)_{i=1}^I$ and m of MEDIAN-STRING. The reduction from MEDIAN-STRING to BAYES is defined by the posteriors $pr(c_i|x) := \frac{1}{I}, i = 1, \dots, I$ and $\tilde{m} := \frac{m}{I}$. Then $(c_i, pr(c_i|x))_{i=1}^I$, and $\tilde{m}$ are an input instance of BAYES. Furthermore, the following equivalence can be deduced:

$$\begin{aligned} \sum_{i=1}^I \mathcal{L}(c_i, c) &\leq m \\ \Leftrightarrow E_{\mathcal{L}}(c|x) &= \sum_{i=1}^I pr(c_i|x) \mathcal{L}(c_i, c) \\ &= \frac{1}{I} \sum_{i=1}^I \mathcal{L}(c_i, c) \leq \frac{m}{I} = \tilde{m}. \end{aligned}$$

This reduction is computable in polynomial time. Therefore, the Bayes optimization problem is NP-complete because BAYES is NP-complete.

10.4 Validation of f -Divergence Properties

Figure 10.1 and Figure 10.2 show the functions $f(u)$ corresponding to the power approximation and *Lin f -Divergences* in comparison to the *Kullback-Leibler f -Divergence*. It can be seen that both the *Lin* and power-approximation f -Divergence converge to the *Kullback-Leibler f -Divergence* for $\alpha = 0$.

Table 10.1 shows the *Kullback-Leibler* fulfills the above-specified properties.

Table 10.1: Values of $f(u), g'(u), f'(u), f''(u)$ and $f'''(u)$ for the function corresponding to the *Kullback-Leibler*, *Power-Approximation* and *Lin f -Divergence*.

$f(u)$	decomposition $g(1/u)$	$f'(u)$	$g'(q(c_n x_n))$	convex $f''(u) > 0$	monotonic $f'''(u)$
$-u \log\left(\frac{1}{u}\right)$	$-\log\left(\frac{1}{u}\right)$	$\log(u) + 1$	$-\frac{1}{q(c_n x_n)}$	$\frac{1}{u}$	$-\frac{1}{u^2}$
$u \frac{(1 - \frac{1}{u^\alpha})}{\alpha}$	$\frac{(1 - \frac{1}{u^\alpha})}{\alpha}$	$(1 - \frac{1}{u^\alpha})$	$-\frac{1}{q^{1-\alpha}(c_n x_n)}$	$\frac{1-\alpha}{u^{1+\alpha}}$	$-\frac{1-\alpha}{u^{2+\alpha}}$
$-u \log\left(\frac{1}{1+\alpha} \left(\alpha + \frac{1}{u}\right)\right)$	$-\log\left(\frac{1}{1+\alpha} \left(\alpha + \frac{1}{u}\right)\right)$	$-\log\left(\frac{1+\alpha u}{u(1+\alpha)}\right) + \frac{1}{1+\alpha u}$	$-\frac{1}{\alpha + q(c_n x_n)}$	$\frac{1}{(1+\alpha u)^2 u}$	$-\frac{1+2\alpha u}{(1+\alpha u)^3 u^2}$

Figure 10.3 and Figure 10.4 show the second derivatives $f''(u)$ of the functions corresponding to the power approximation and *Lin f -Divergences* in comparison to the *Kullback-Leibler f -Divergence* with $\alpha = 0$. It can be seen that all these functions are convex.

Figure 10.5 and Figure 10.6 show the third derivatives $f'''(u)$ of the functions corresponding to the power approximation and *Lin f -Divergences* in comparison to the *Kullback-Leibler f -Divergence*. It can be seen that all these functions are monotonic increasing in the interval $[1, 2]$.

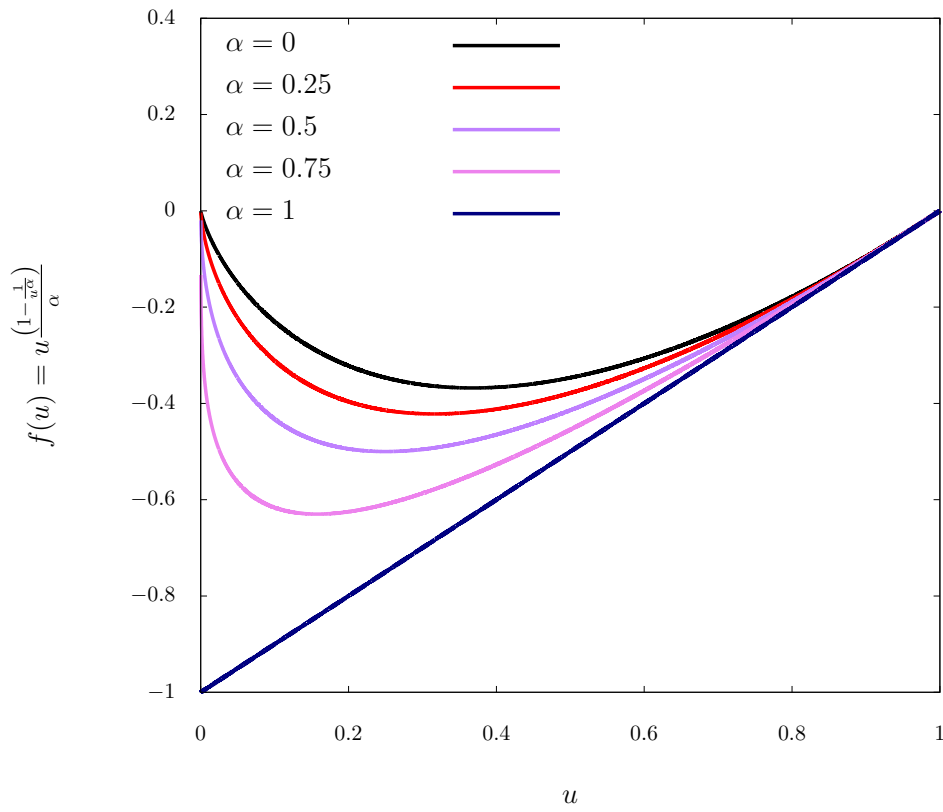


Figure 10.1: *Graph of the functions $f(u)$ corresponding to the power approximation f -Divergence. The power approximation f -Divergence converges to the Kullback-Leibler divergence for $\alpha = 0$.*

Figure 10.7 and Figure 10.8 show the function corresponding to the criteria for the posterior $q(c_n|x_n)$ for the *Kullback-Leibler*, *Lin*, and power approximation criterion.

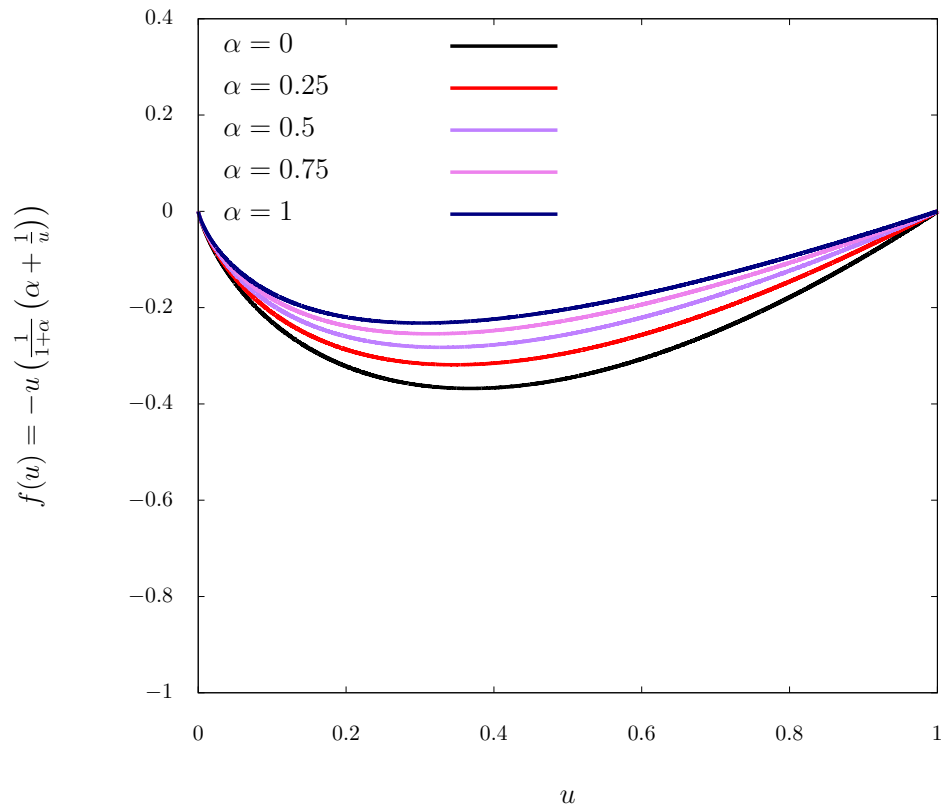


Figure 10.2: Graph of the functions $f(u)$ corresponding to the Lin f -Divergence. The Lin f -Divergence converges to the Kullback-Leibler for $\alpha = 0$.

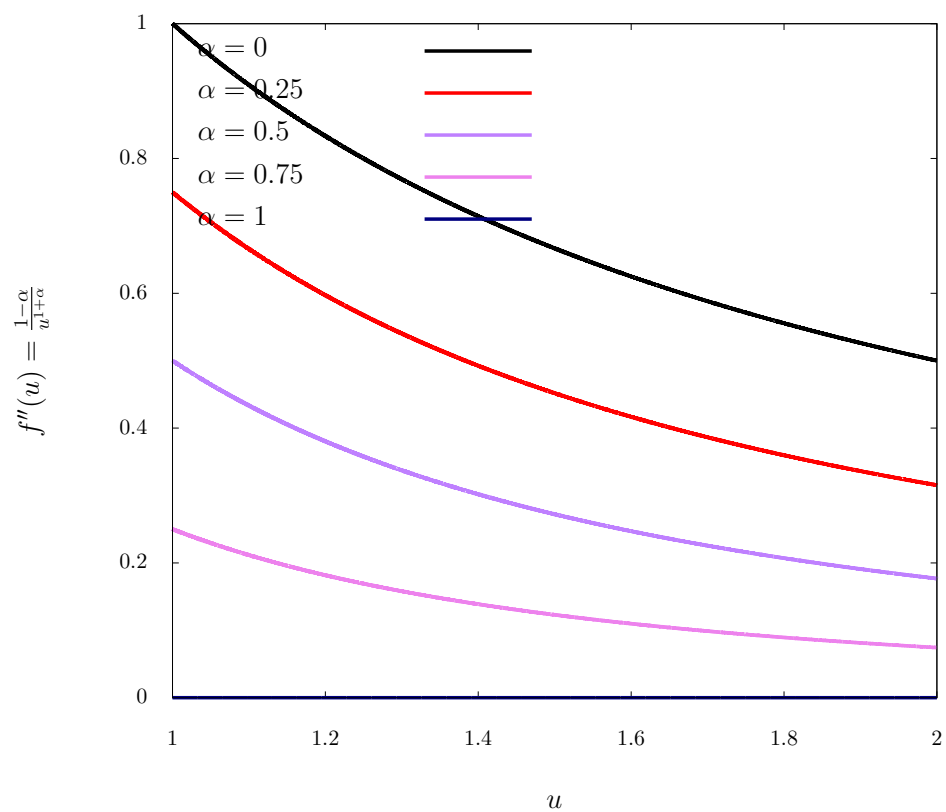


Figure 10.3: *Second derivative of the functions $f''(u)$ corresponding to the power approximation f -Divergence. The power approximation f -Divergence converges to the Kullback-Leibler for $\alpha = 0$.*

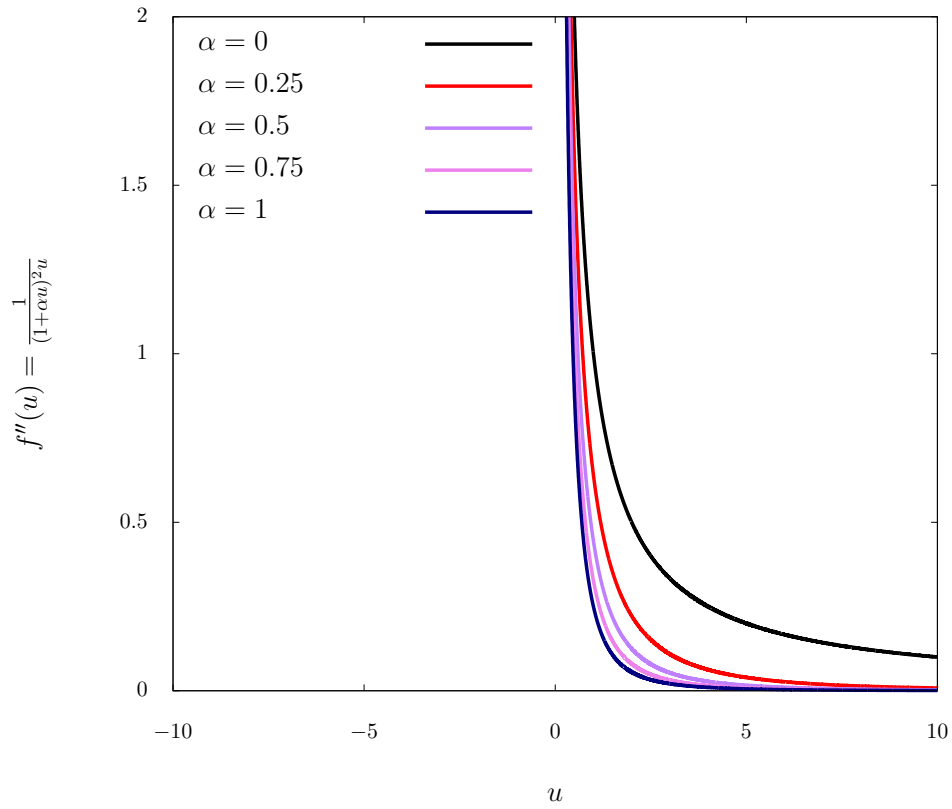


Figure 10.4: Second derivative of the functions $f''(u)$ corresponding to the Lin f -Divergence. The Lin f -Divergence converges to the Kullback-Leibler for $\alpha = 0$.

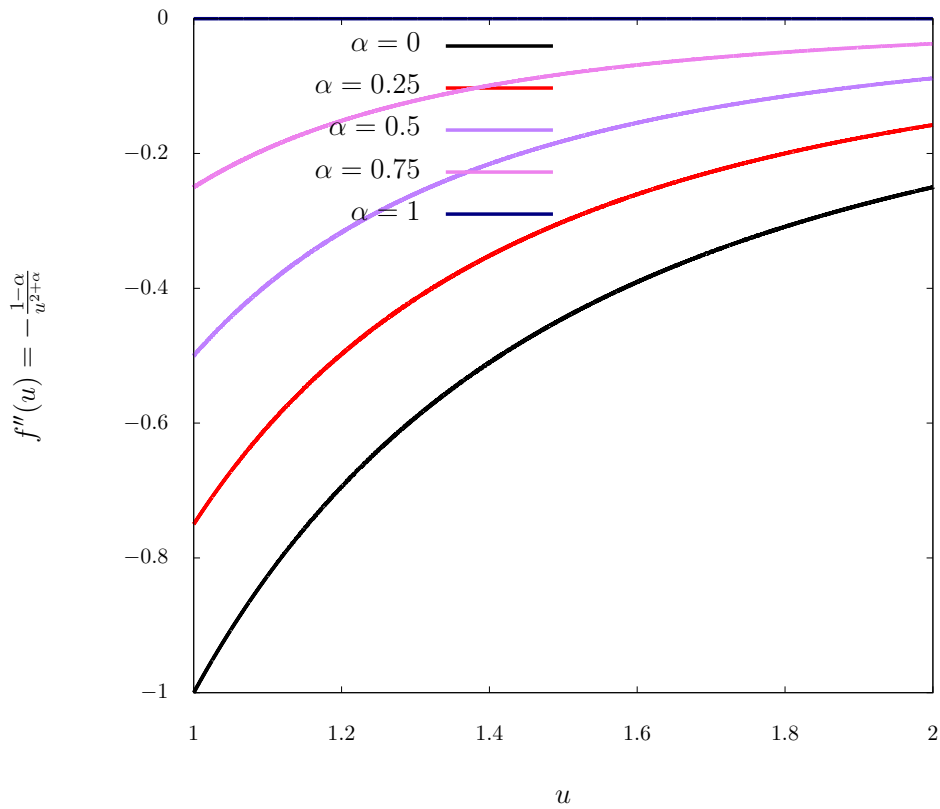


Figure 10.5: *Third derivative of the functions $f'''(u)$ corresponding to the power approximation f -Divergence. The power approximation f -Divergence converges to the Kullback-Leibler for $\alpha = 0$.*

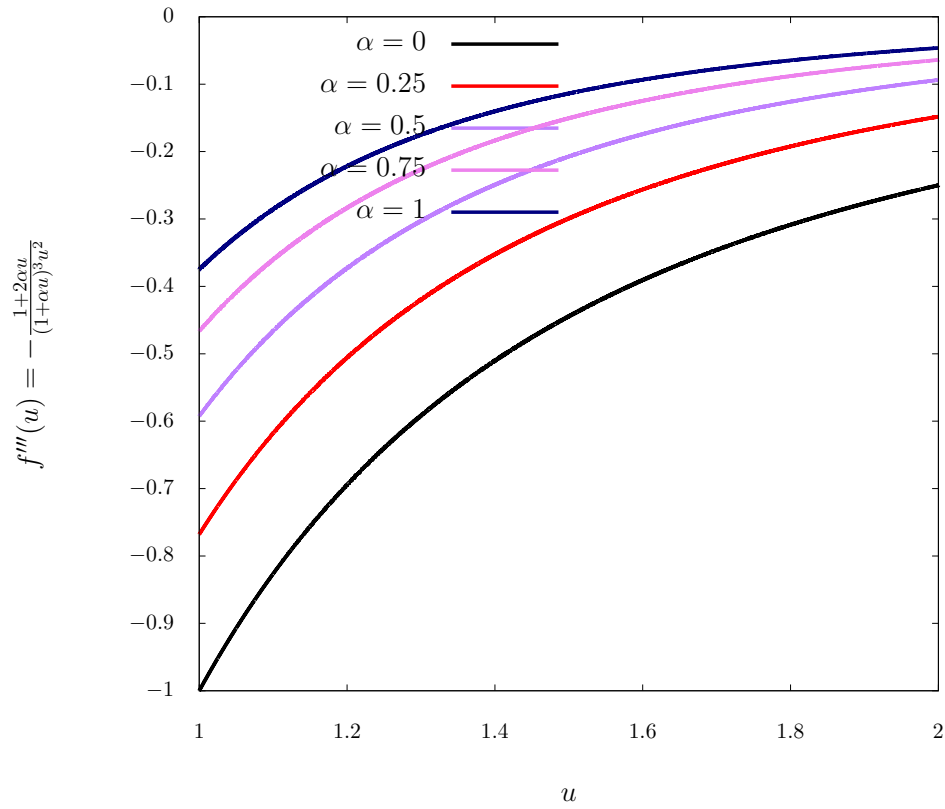


Figure 10.6: Third derivative of the functions $f'''(u)$ corresponding to the Lin f -Divergence. The Lin f -Divergence converges to the Kullback-Leibler for $\alpha = 0$.

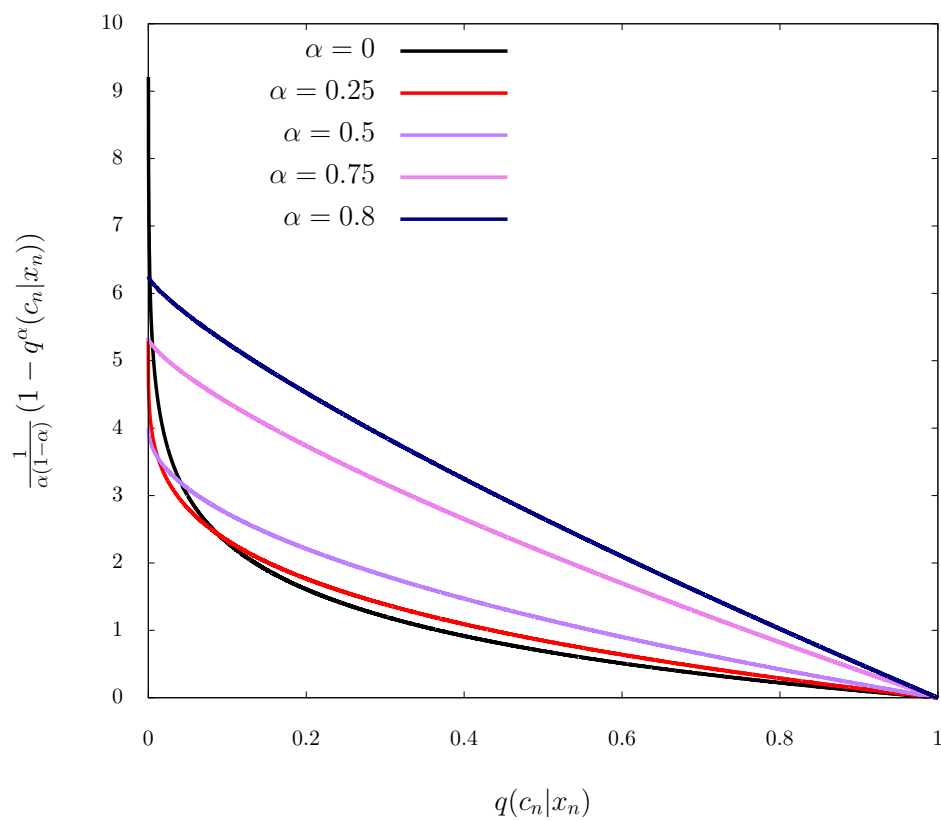


Figure 10.7: *Functions corresponding to the criterion for the power approximation and Kullback-Leibler criterion.*

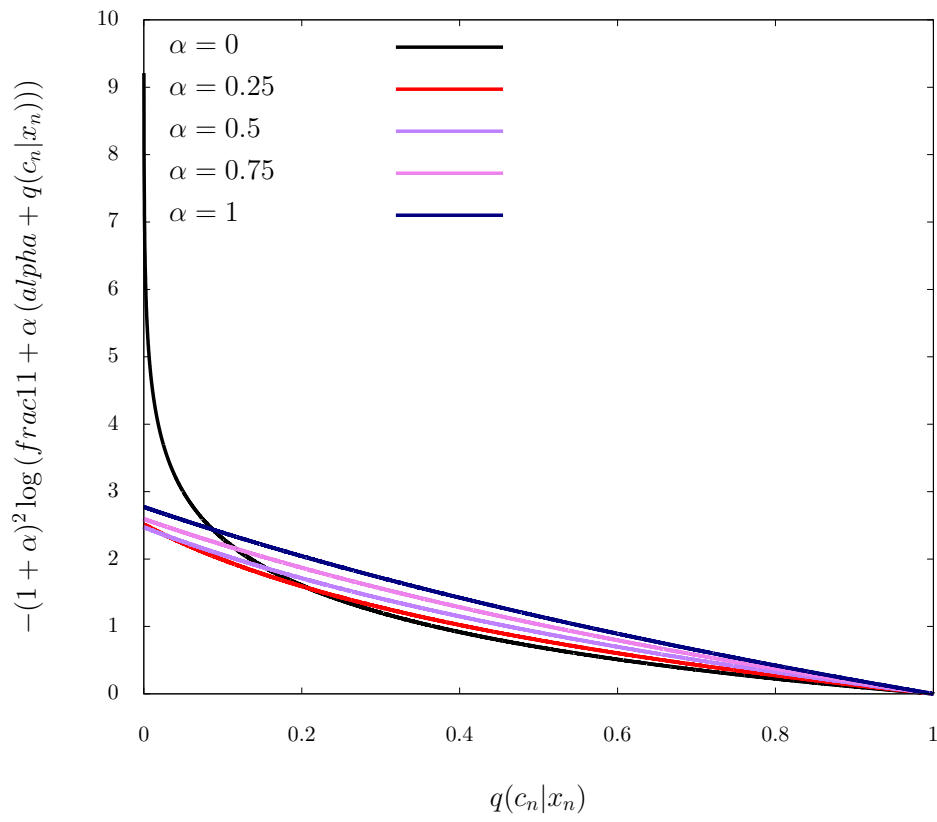


Figure 10.8: Functions corresponding to the criterion for the Lin and Kullback-Leibler criterion.

LIST OF FIGURES

1.1	The Standard Approach to Automatic Speech Recognition	2
1.2	6-state hidden Markov model in Bakis topology for the triphone seh_v in the word “seven”. The HMM segments are denoted by $\langle 1 \rangle$, $\langle 2 \rangle$, and $\langle 3 \rangle$	5
4.1	<i>Chi-Squared Bound.</i>	28
4.2	<i>Vajda Bound for $f_V(u) = u - 1$.</i>	29
4.3	Example of the distribution $\phi(y x)$ for mean $\mu(x) = 1, 2, 4$	33
6.1	TER[%] progress vs. the linear combination of the 0.5-PA criterion with CE on the Bengali Babel test set.	54
6.2	TER[%] progress vs. factor α of the α -PA criterion measured on the Bengali development set.	54
6.3	<i>Function $\frac{1-q^\alpha(c_n x_n)}{\alpha(1-\alpha)}$ corresponding to one sample of the α-PA criterion which converges to the cross-entropy criterion for $\alpha \rightarrow 0$.</i>	57
6.4	<i>The WER[%] as a function of the parameter α of the α-PA criterion trained using DNN models.</i>	60
6.5	<i>The WER[%] as a function of the constraint β for the CE+MIN-SAMP-PA criterion for BGRU models with dropout 0.1 on the WSJ0 test corpora.</i>	61
6.6	<i>The WER[%] as a function of the variance σ^2 for the CE+RAND-SAMP-PA criterion with $\mu = 10^{-6}$ for BGRU models with dropout 0.1 on the WSJ0 test corpora.</i>	61
7.1	WER progress of the α -MPE criterion for different α values on the TIMIT test set.	65
8.1	Schematic diagram of the decoding framework for German.	71
8.2	Schematic view of the RWTH recognition process	73
10.1	<i>Graph of the functions $f(u)$ corresponding to the power approximation f-Divergence. The power approximation f-Divergence converges to the Kullback-Leibler divergence for $\alpha = 0$.</i>	84
10.2	<i>Graph of the functions $f(u)$ corresponding to the Lin f-Divergence. The Lin f-Divergence converges to the Kullback-Leibler for $\alpha = 0$.</i>	85
10.3	<i>Second derivative of the functions $f''(u)$ corresponding to the power approximation f-Divergence. The power approximation f-Divergence converges to the Kullback-Leibler for $\alpha = 0$.</i>	86
10.4	<i>Second derivative of the functions $f''(u)$ corresponding to the Lin f-Divergence. The Lin f-Divergence converges to the Kullback-Leibler for $\alpha = 0$.</i>	87

10.5	Third derivative of the functions $f'''(u)$ corresponding to the power approximation f -Divergence. The power approximation f -Divergence converges to the Kullback-Leibler for $\alpha = 0$	88
10.6	Third derivative of the functions $f'''(u)$ corresponding to the Lin f -Divergence. The Lin f -Divergence converges to the Kullback-Leibler for $\alpha = 0$	89
10.7	Functions corresponding to the criterion for the power approximation and Kullback-Leibler criterion.	90
10.8	Functions corresponding to the criterion for the Lin and Kullback-Leibler criterion.	91

LIST OF TABLES

6.1	<i>TER[%] results for frame-wise training using the CE criterion in combination with the 0.5-PA and Lin (+1-LIN) criteria evaluated on the Vietnamese (V) and Bengali (B) test sets, followed by HF sMBR training.</i>	55
6.2	Corpus statistics.	58
6.3	The WER[%] as a function of different training criteria working as a regularization scheme for DNN, BGRU and BGRU(0.1) (BGRU with dropout rate 0.1) model on the WSJ0 test corpora.	59
7.1	<i>Corpus statistics.</i>	64
7.2	<i>WER[%] results for the ML and modified criteria on the TIMIT test set.</i>	65
8.1	<i>In-domain transcribed audio data for German.</i>	68
8.2	<i>Additional transcribed audio data for the German system available for acoustic modeling.</i>	68
8.3	Language model training data	69
8.4	MLP feature training data and network topology	70
8.5	<i>Statistics for the dev09 and eval09 corpora for German.</i>	71
8.6	<i>Results on the German dev09 and eval09 corpus.</i>	72
8.7	Results on the 2010 development and evaluation set.	73
8.8	Improvements in terms of WER obtained by different methods, measured on the 2010 development corpus	74
8.9	Improvements in terms of WER obtained by different partners for the German system from 2008 through 2013.	74
10.1	<i>Values of $f(u)$, $g'(u)$, $f'(u)$, $f''(u)$ and $f'''(u)$ for the function corresponding to the Kullback-Leibler, Power-Approximation and Lin f-Divergence.</i>	83

BIBLIOGRAPHY

- [Alleva & Huang⁺ 96] P. Alleva, X.D. Huang, M.Y. Hwang: Improvements on the Pronunciation Prefix Tree Search Organization. Vol. 1, pp. 133–136, Atlanta, GA, USA, May 1996.
- [Bahl & Baker⁺ 76] L.R. Bahl, J.K. Baker, P.S. Cohen, N.R. Dixon, F. Jelinek, R.L. Mercer, H.F. Silverman: Preliminary results on the performance of a system for the automatic recognition of continuous speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '76, Philadelphia, Pennsylvania, USA, April 12-14, 1976*, pp. 425–429, 1976.
- [Bahl & Brown⁺ 86] L. Bahl, P. Brown, P. de Souza, R. Mercer: Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition. pp. 49–52, Tokyo, Japan, May 1986.
- [Bahl & Jelinek⁺ 83] L.R. Bahl, F. Jelinek, R.L. Mercer: A Maximum Likelihood Approach to Continuous Speech Recognition. *PAT*, Vol. 5, pp. 179–190, Mar 1983.
- [Bahl & Padmanabhan⁺ 96] L.R. Bahl, M. Padmanabhan, D. Nahamoo, P.S. Gopalakrishnan: Discriminative Training of Gaussian Mixture Models for Large Vocabulary Speech Recognition Systems. pp. 613–616, Atlanta, GA, USA, May 1996.
- [Baker 75] J.K. Baker: Stochastic Modeling for Automatic Speech Understanding. In D.R. Reddy, editor, *Speech Recognition*, pp. 512–542. Academic Press, New York, NY, USA, 1975.
- [Bakis 76] R. Bakis: Continuous Speech Word Recognition via Centisecond Acoustic States. In *ASA Meeting*, Washington, DC, USA, April 1976.
- [Baum 72] L.E. Baum: An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. In O. Shisha, editor, *Inequalities*, Vol. 3, pp. 1–8. Academic Press, New York, NY, 1972.
- [Bellman 57] R.E. Bellman: Dynamic Programming. Princeton University Press, Princeton, NJ, USA, 1957.
- [Berger 85] J. Berger: *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer-Verlag, New York, revised and expanded 2nd edition edition, 1985.
- [Beulen & Ortmanns⁺ 99] K. Beulen, S. Ortmanns, C. Elting: Dynamic Programming Search Techniques for Across-Word Modeling in Speech Recognition. pp. 609–612, Phoenix, AZ, March 1999.
- [Bisani & Ney 08] M. Bisani, H. Ney: Joint-Sequence Models for Grapheme-to-Phoneme Conversion. *Speech Commun.*, Vol. 50, No. 5, pp. 434–451, May 2008.

- [Bocchieri 93] E. Bocchieri: Vector Quantization for the Efficient Computation of Continuous Density Likelihoods. pp. 692–695, Minneapolis, MN, April 1993.
- [Bourlard & Morgan 93] H.A. Bourlard, N. Morgan: *Connectionist Speech Recognition: A Hybrid Approach*. Kluwer Academic Publishers, USA, 1993.
- [Bracewell 99] R. Bracewell: *The Sifting Property*. McGraw-Hill, New York, 3rd edition, 1999.
- [Bridle 90] J.S. Bridle: Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. In F.F. Soulié, J. Héroult, editors, *Neurocomputing*, pp. 227–236, Berlin, Heidelberg, 1990. Springer.
- [Cardin & Normandin⁺ 93] R. Cardin, Y. Normandin, E. Millien: Inter-Word Coarticulation Modeling and MMIE Training for Improved Connected Digit Recognition. pp. 243–246, Minneapolis, MN, USA, April 1993.
- [Chen & Gopalakrishnan 98] S.S. Chen, P.S. Gopalakrishnan: Clustering via the Bayesian information criterion with applications in speech recognition. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, Vol. 2, pp. 645–648, 1998.
- [Cho & Van Merriënboer⁺ 14] K. Cho, B. Van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, Oct. 2014. Association for Computational Linguistics.
- [Chow 90] Y.L. Chow: Maximum Mutual Information Estimation of HMM Parameters for Continuous Speech Recognition using the N-best Algorithm. pp. 701–704, Albuquerque, NM, USA, April 1990.
- [Chung & Gülçehre⁺ 14] J. Chung, Ç. Gülçehre, K. Cho, Y. Bengio: Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *CoRR*, Vol. abs/1412.3555, 2014.
- [Cover & Thomas 06] T.M. Cover, J.A. Thomas: *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.
- [Csiszár & Shields 04] I. Csiszár, P.C. Shields: Information Theory and Statistics: A Tutorial. Vol. 1, No. 4, 2004.
- [Cui & Cui⁺ 13] J. Cui, X. Cui, B. Ramabhadran, J. Kim, B. Kingsbury, J. Mamou, L. Mangu, M. Picheny, T.N. Sainath, A. Sethy: Developing Speech Recognition Systems for Corpus Indexing under the IARPA Babel Program. In *ICASSP*, 2013.
- [Cui & Kingsbury⁺ 15] J. Cui, B. Kingsbury, B. Ramabhadran, A. Sethy, K. Audhkhasi, X. Cui, E. Kislal, L. Mangu, M. Nußbaum-Thom, M. Picheny, Z. Tüske, P. Golik, R. Schlüter, H. Ney, M.J.F. Gales, K.M. Knill, A. Ragni, H. Wang, P.C. Woodland: Multilingual representations for low resource speech recognition and keyword search. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, pp. 259–266, 2015.
- [Dahl & Ranzato⁺ 10] G.E. Dahl, M. Ranzato, A. rahman Mohamed, G.E. Hinton: Phone Recognition with the Mean-Covariance Restricted Boltzmann Machine. In *NIPS*, pp. 469–477. Curran Associates, Inc., 2010.

- [Davis & Mermelstein 80] S. Davis, P. Mermelstein: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. Vol. ASSP-28, No. 4, pp. 357 – 366, Aug. 1980.
- [de la Higuera & Casacuberta 00] C. de la Higuera, F. Casacuberta: Topology of strings: Median string is NP-complete. *Theoretical Computer Science*, Vol. 230, No. 1-2, pp. 39–48, 2000.
- [Dempster & Laird⁺ 77] A. Dempster, N. Laird, D. Rubin: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, Vol. 39, No. B, pp. 1 – 38, 1977.
- [Deng & Acero⁺ 12] L. Deng, A. Acero, G. Dahl, D. Yu: Context-Dependent Pre-trained Deep Neural Networks for Large Vocabulary Speech Recognition. In *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 20, pp. 30–42, January 2012. IEEE SPS 2013 best paper award.
- [Doumpiotis & Byrne 04] V. Doumpiotis, W. Byrne: Pinched Lattice Minimum Bayes Risk Discriminative Training for Large Vocabulary Continuous Speech Recognition. In *Proc. of the International Conference on Spoken Language Processing*, (4 pages), 2004.
- [Duda & Hart⁺ 00] R.O. Duda, P.E. Hart, D.G. Stork: *Pattern Classification*. John Wiley and Sons, 2nd edition, 2000.
- [Duda & Hart⁺ 01] R.O. Duda, P.E. Hart, D.G. Stork: *Pattern Classification*. John Wiley & Sons, New York, NY, USA, 2001.
- [Eisner 01] J. Eisner: Expectation semirings: Flexible EM for finite-state transducers. In *International Workshop on Finite-State Methods and Natural Language Processing (FSMNLP)*, Helsinki, Finland, Aug. 2001.
- [El-Desoky Mousa & Gollan⁺ 09] A. El-Desoky Mousa, C. Gollan, D. Rybach, R. Schlüter, H. Ney: Investigating the Use of Morphological Decomposition and Diacritization for Improving Arabic LVCSR. In *Interspeech*, pp. 2679–2682, Brighton, UK, Sept. 2009.
- [Evermann & Woodland 00a] G. Evermann, P. Woodland: Posterior Probability Decoding, Confidence Estimation And System Combination, 2000.
- [Evermann & Woodland 00b] G. Evermann, P.C. Woodland: Large vocabulary decoding and confidence estimation using word posterior probabilities. In *IEEE International Conference on Acoustics, Speech, and Signal Processing. ICASSP 2000, 5-9 June, 2000, Hilton Hotel and Convention Center, Istanbul, Turkey*, pp. 1655–1658, 2000.
- [Fedotov & Harremos⁺ 03] A. Fedotov, P. Harremos, F. Topsoe: Refinements of Pinsker’s inequality. *Information Theory, IEEE Transactions on*, Vol. 49, pp. 1491 – 1498, 07 2003.
- [Fisher 36] R.A. Fisher: The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, Vol. 7, No. 179-188, 1936.
- [Fritsch 97] J. Fritsch: ACID/HNN: A Framework for Hierarchical Connectionist Acoustic Modeling. pp. 164–171, Santa Barbara, CA, USA, Dec. 1997.
- [Gales 98] M.J.F. Gales: *Computer Speech and Language*, Vol. 12, No. 2, pp. 75–98, 1998.
- [Generet & Ney⁺ 95] M. Generet, H. Ney, F. Wessel: Extensions to Absolute Discounting for Language Modeling. Vol. 2, pp. 1245–1248, Madrid, Spain, Sept. 1995.

- [Gibson 08] M. Gibson: *Minimum Bayes risk acoustic model estimation and adaptation*. Ph.D. thesis, University of Sheffield, Sheffield, UK, 2008.
- [Goel & Byrne⁺ 98] V. Goel, W. Byrne, S. Khudanpur: LVCSR rescoring with modified loss functions: a decision theoretic perspective. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98, Seattle, Washington, USA, May 12-15, 1998*, pp. 425–428, 1998.
- [Graves & Jaitly⁺ 13] A. Graves, N. Jaitly, A. Mohamed: Hybrid Speech Recognition with Deep Bidirectional LSTM. In *ASRU*, pp. 273–278. IEEE, 2013.
- [Guntuboyina 11] A. Guntuboyina: Lower Bounds for the Minimax Risk using f-Divergences and Applications. In *IEEE Transactions on Information Theory*, 2011.
- [Hüb-Umbach & Ney 94] R. Hüb-Umbach, H. Ney: Improvements in Beam Search for 10000-Word Continuous-Speech Recognition. Vol. 2, No. 2, pp. 353–356, April 1994.
- [Heigold 10] G. Heigold: *A Log-Linear Discriminative Modeling Framework for Speech Recognition*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, June 2010.
- [Heigold & Deselaers⁺ 08] G. Heigold, T. Deselaers, R. Schlüter, H. Ney: Modified MMI/MPE: A Direct Evaluation of the Margin in Speech Recognition. In *Proceedings of the 24th International Conference on Machine Learning (ICML 2008)*, pp. 384–391, Helsinki, Finland, July 2008.
- [Heigold & Ney⁺ 12] G. Heigold, H. Ney, R. Schlüter, S. Wiesler: Discriminative Training for Automatic Speech Recognition: Modeling, Criteria, Optimization, Implementation, and Performance. *IEEE Signal Process. Mag.*, Vol. 29, No. 6, pp. 58–69, 2012.
- [Heigold & Wiesler⁺ 10] G. Heigold, S. Wiesler, M. Nußbaum-Thom, P. Lehnen, R. Schlüter, H. Ney: Discriminative HMMS, log-linear models, and CRFS: What is the difference? In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2010, 14-19 March 2010, Sheraton Dallas Hotel, Dallas, Texas, USA*, pp. 5546–5549, 2010.
- [Hermansky 90] H. Hermansky: Perceptual Linear Predictive (PLP) Analysis of Speech. Vol. 87, No. 4, pp. 1738 – 1752, June 1990.
- [Hoffmeister & Klein⁺ 06] B. Hoffmeister, T. Klein, R. Schlüter, H. Ney: Frame based system combination and a comparison with weighted ROVER and CNC. In *INTERSPEECH 2006 - ICSLP, Ninth International Conference on Spoken Language Processing, Pittsburgh, PA, USA, September 17-21, 2006*, 2006.
- [Hoffmeister & Plahl⁺ 07] B. Hoffmeister, C. Plahl, P. Fritz, G. Heigold, J. Löff, R. Schlüter, H. Ney: Development of the 2007 RWTH Mandarin LVCSR System. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 455–460, Kyoto, Japan, Dec. 2007.
- [Hon & Lee 91] H.W. Hon, K.F. Lee: Recent Progress in Robust Vocabulary-Independent Speech Recognition. In *DARPA Speech and Natural Language Processing Workshop*, pp. 258–263, Pacific Grove, Feb. 1991.
- [Hromkovic 01] J. Hromkovic: *Algorithmics for hard problems - introduction to combinatorial optimization, randomization, approximation, and heuristics*. Springer, 2001.
- [Huang & Jack 89] X.D. Huang, M.A. Jack: Semi-Continuous Hidden Markov Models for Speech Signals. Vol. 3, No. 3, pp. 329–252, 1989.

- [Jelinek 69] F. Jelinek: A Fast Sequential Decoding Algorithm Using a Stack. *IBM Journal of Research and Development*, Vol. 13, pp. 675–685, Nov. 1969.
- [Józefowicz & Zaremba⁺ 15] R. Józefowicz, W. Zaremba, I. Sutskever: An Empirical Exploration of Recurrent Network Architectures. In *ICML*, Vol. 37 of *JMLR Proceedings*, pp. 2342–2350. JMLR.org, 2015.
- [Kaiser & Horvat⁺ 00] J. Kaiser, B. Horvat, Z. Kacic: A novel loss function for the overall risk criterion based discriminative training of HMM models. In *INTERSPEECH*, pp. 887–890. ISCA, 2000.
- [Kaiser & Horvat⁺ 02] J. Kaiser, B. Horvat, Z. Kacic: Overall risk criterion estimation of hidden Markov model parameters. *Speech Communication*, Vol. 38, pp. 383–398, 2002.
- [Kanthak & Schütz⁺ 00] S. Kanthak, K. Schütz, H. Ney: Using SIMD Instructions for Fast Likelihood Calculation in LVCSR. pp. 1531–1534, Istanbul, Turkey, June 2000.
- [Kanthak & Sixtus⁺ 00] S. Kanthak, A. Sixtus, S. Molau, R. Schlüter, H. Ney: *Fast Search for Large Vocabulary Speech Recognition*, chapter "From Speech Input to Augmented Word Lattices", pp. 63–78. Springer Verlag, Berlin, Heidelberg, New York, July 2000.
- [Kapadia & Valtchev⁺ 93] S. Kapadia, V. Valtchev, S.J. Young: MMI Training for Continuous Phoneme Recognition on the TIMIT Database. pp. 491–494, Minneapolis, MN, April 1993.
- [Katz 87] S.M. Katz: Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. Vol. 35, pp. 400–401, March 1987.
- [Kingsbury & Sainath⁺ 12] B. Kingsbury, T.N. Sainath, H. Soltau: Scalable Minimum Bayes Risk Training of Deep Neural Network Acoustic Models Using Distributed Hessian-free Optimization. In *INTERSPEECH*, 2012.
- [Kneser & Ney 95] R. Kneser, H. Ney: Improved Backing-Off for m-Gram Language Modeling. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. I, pp. 181–184, Detroit, Michigan, May 1995.
- [Lamel & Courcinous⁺ 11] L. Lamel, S. Courcinous, J. Despres, J. Gauvain, Y. Josse, K. Kilgour, F. Kraft, V.B. Le, H. Ney, M. Nußbaum-Thom, I. Oparin, T. Schlippe, R. Schlüter, T. Schultz, T. Fraga-Silva, S. Stüker, M. Sundermeyer, B. Vieru, N.T. Vu, A. Waibel, C. Woehrling: Speech recognition for machine translation in Quaero. In *2011 International Workshop on Spoken Language Translation, IWSLT 2011, San Francisco, CA, USA, December 8-9, 2011*, pp. 121–128, 2011.
- [Lee & Hon 89] K. Lee, H. Hon: Speaker-independent phone recognition using hidden Markov models. *IEEE Trans. Acoustics, Speech, and Signal Processing*, Vol. 37, No. 11, pp. 1641–1648, 1989.
- [Levinson & Rabiner⁺ 83] S.E. Levinson, L.R. Rabiner, M.M. Sondhi: An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell System Technical Journal*, Vol. 62, No. 4, pp. 1035–1074, April 1983.
- [Liese & Vajda 06] F. Liese, I. Vajda: On Divergences and Informations in Statistics and Information Theory. Vol. 52, No. 10, pp. 4394–4412, 2006.

- [Lööf & Gollan⁺ 07] J. Lööf, C. Gollan, S. Hahn, G. Heigold, B. Hoffmeister, C. Plahl, D. Rybach, R. Schlüter, H. Ney: The RWTH 2007 TC-STAR evaluation system for european English and Spanish. In *INTERSPEECH 2007, 8th Annual Conference of the International Speech Communication Association, Antwerp, Belgium, August 27-31, 2007*, pp. 2145–2148, 2007.
- [Lowerre 76] B. Lowerre: *A Comparative Performance Analysis of Speech Understanding Systems*. Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 1976.
- [Macherey 10] W. Macherey: *Discriminative training and acoustic modeling for automatic speech recognition*. Ph.D. thesis, RWTH Aachen University, 2010.
- [Macherey & Haferkamp⁺ 05] W. Macherey, L. Haferkamp, R. Schlüter, H. Ney: Investigations on Error Minimizing Training Criteria for Discriminative Training in Automatic Speech Recognition. In *Interspeech*, pp. 2133–2136, Lisboa, Portugal, Sept. 2005.
- [Macherey & Ney 02] W. Macherey, H. Ney: Towards Automatic Corpus Preparation for a German Broadcast News Transcription System. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 733–736, Orlando, FL, USA, May 2002.
- [Mangu & Brill⁺ 00] L. Mangu, E. Brill, A. Stolcke: Finding consensus in speech recognition: word error minimization and other applications of confusion networks. *Comput. Speech Lang.*, Vol. 14, No. 4, pp. 373–400, 2000.
- [Mohamed & Dahl⁺ 09] A. Mohamed, G. Dahl, G. Hinton: Deep Belief Networks for phone recognition. *NIPS 22 workshop on deep learning for speech recognition*, Vol., 2009.
- [Molau 03] S. Molau: *Normalization in the Acoustic Feature Space for Improved Speech Recognition*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, 2003.
- [Ney 84] H. Ney: The Use of a One-Stage Dynamic Programming Algorithm for Connected Word Recognition. Vol. 32, No. 2, pp. 263–271, April 1984.
- [Ney 90] H. Ney: Acoustic Modeling of Phoneme Units for Continuous Speech Recognition. In L. Torres, E. Masgrau, M.A. Lagunas, editors, *Signal Processing V: Theories and Applications, Fifth European Signal Processing Conference*, pp. 65–72. Elsevier Science Publishers B. V., Barcelona, Spain, 1990.
- [Ney 03] 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)*, pp. 636–645, Puerto de Andratx, Spain, June 2003.
- [Ney 05] H. Ney: Speech Reconition. Lecture Notes, 2005.
- [Ney & Aubert 94] H. Ney, X. Aubert: A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition. Vol. 3, pp. 1355–1358, Yokohama, Japan, Sept. 1994.
- [Ney & Essen⁺ 94] H. Ney, U. Essen, R. Kneser: On Structuring Probabilistic Dependencies in Language Modeling. Vol. 2, No. 8, pp. 1–38, 1994.
- [Ney & Häb-Umbach⁺ 92] H. Ney, R. Häb-Umbach, B.H. Tran, M. Oerder: Improvements in Beam Search for 10000-Word Continuous Speech Recognition. Vol. 1, pp. 9–12, San Francisco, CA, March 1992.

- [Ney & Martin⁺ 97] H. Ney, S.C. Martin, F. Wessel: Statistical Language Modeling using Leaving-One-Out. In S. Young, G. Bloothoof, editors, *Corpus Based Methods in Language and Speech Processing*, pp. 1–26. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.
- [Ney & Mergel⁺ 87] H. Ney, D. Mergel, A. Noll, A. Paeseler: A Data-Driven Organization of the Dynamic Programming Beam Search for Continuous Speech Recognition. pp. 833–836, Dallas, TX, USA, April 1987.
- [Normandin & Morgera 91] Y. Normandin, S. Morgera: An Improved MMIE Training Algorithm for Speaker-Independent, Small Vocabulary, Continuous Speech Recognition. pp. 537–540, Toronto, Canada, May 1991.
- [Nußbaum-Thom & Beck⁺ 13] M. Nußbaum-Thom, E. Beck, T. Alkhouli, R. Schlüter, H. Ney: Relative error bounds for statistical classifiers based on the f-divergence. In *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, pp. 2197–2201, 2013.
- [Nußbaum-Thom & Cui⁺ 14] M. Nußbaum-Thom, X. Cui, R. Schlüter, V. Goel, H. Ney: A family of discriminative training criteria based on the f-divergence for deep neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2014, Florence, Italy, May 4-9, 2014*, pp. 5612–5616, 2014.
- [Nußbaum-Thom & Cui⁺ 16a] M. Nußbaum-Thom, J. Cui, B. Ramabhadran, V. Goel: Acoustic Modeling Using Bidirectional Gated Recurrent Convolutional Units. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pp. 390–394, 2016.
- [Nußbaum-Thom & Cui⁺ 16b] M. Nußbaum-Thom, J. Cui, B. Ramabhadran, V. Goel: Acoustic Modeling Using Bidirectional Gated Recurrent Convolutional Units. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pp. 390–394, 2016.
- [Nußbaum-Thom & Schlüter⁺ 17] M. Nußbaum-Thom, R. Schlüter, V. Goel, H. Ney: Noisy objective functions based on the f-divergence. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*, pp. 2327–2331, 2017.
- [Nußbaum-Thom & Tüske⁺ 12] M. Nußbaum-Thom, Z. Tüske, G. Heigold, R. Schlüter, H. Ney: Posterior-Scaled MPE: Novel Discriminative Training Criteria. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pp. 2614–2617, 2012.
- [Nußbaum-Thom & Wiesler⁺ 10a] M. Nußbaum-Thom, S. Wiesler, M. Sundermeyer, C. Plahl, S. Hahn, R. Schlüter, H. Ney: The RWTH 2009 quaero ASR evaluation system for English and German. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 1517–1520, 2010.
- [Nußbaum-Thom & Wiesler⁺ 10b] M. Nußbaum-Thom, S. Wiesler, M. Sundermeyer, C. Plahl, S. Hahn, R. Schlüter, H. Ney: The RWTH 2009 Quaero ASR Evaluation System for English and German. In *Interspeech*, pp. 1517–1520, Makuhari, Japan, Sept. 2010.

- [Odell & Valtchev⁺ 94] J.J. Odell, V. Valtchev, P.C. Woodland, S.J. Young: A One-Pass Decoder Design for Large Vocabulary Recognition. In *ARPA Spoken Language Technology Workshop*, pp. 405–410, Plainsboro, NJ, March 1994.
- [Ortmanns 98] S. Ortmanns: *Effiziente Suchverfahren zur Erkennung kontinuierlich gesprochener Sprache*. Ph.D. thesis, RWTH Aachen, Aachen, Germany, Nov. 1998.
- [Ortmanns & Ney 95] S. Ortmanns, H. Ney: An Experimental Study of the Search Space for 20000-Word Speech Recognition. Vol. 2, pp. 901–904, Madrid, Spain, Sept. 1995.
- [Ortmanns & Ney⁺ 96] S. Ortmanns, H. Ney, A. Eiden: Language-Model Look-Ahead for Large Vocabulary Speech Recognition. Vol. 4, pp. 2095–2098, Philadelphia, PA, Oct. 1996.
- [Ortmanns & Ney⁺ 97a] S. Ortmanns, H. Ney, X. Aubert: A Word Graph Algorithm for Large Vocabulary Continuous Speech Recognition. Vol. 11, No. 1, pp. 43–72, Jan. 1997.
- [Ortmanns & Ney⁺ 97b] S. Ortmanns, H. Ney, T. Firzlaß: Fast Likelihood Computation Methods for Continuous Mixture Densities in Large Vocabulary Speech Recognition. Vol. 1, pp. 139–142, Rhodes, Greece, Sept. 1997.
- [Österreicher 02] F. Österreicher: 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.
- [Paul 91] D.B. Paul: Algorithms for an Optimal A* Search and Linearizing the Search in the Stack Decoder. Vol. 1, pp. 693–696, Toronto, Canada, May 1991.
- [Peitz & Mansour⁺ 12] S. Peitz, S. Mansour, M. Freitag, M. Feng, M. Huck, J. Wuebker, M. Nuhn, M. Nußbaum-Thom, H. Ney: The RWTH Aachen speech recognition and machine translation system for IWSLT 2012. In *2012 International Workshop on Spoken Language Translation, IWSLT 2012, Hong Kong, December 6-7, 2012*, pp. 69–76, 2012.
- [Pitz 05] M. Pitz: *Investigations on Linear Transformations for Speaker Adaptation and Normalization*. Ph.D. thesis, RWTH Aachen University, 2005.
- [Plahl & Schlüter⁺ 10] C. Plahl, R. Schlüter, H. Ney: Hierarchical Bottle Neck Features for LVCSR. In *Interspeech*, pp. 1197–1200, Makuhari, Japan, Sept. 2010.
- [Povey 04] D. Povey: *Discriminative Training for Large Vocabulary Speech Recognition*. Ph.D. thesis, Cambridge, England, 2004.
- [Povey & Kingsbury 07] D. Povey, B. Kingsbury: Evaluation of proposed Modifications to MPE for Large Scale Discriminative Training. Honolulu, HI, USA, April 2007.
- [Povey & Woodland 99] D. Povey, P. Woodland: Frame Discrimination Training Of HMMs For Large Vocabulary Speech Recognition. In *Proc. ICASSP99*, pp. 333–336, 1999.
- [Povey & Woodland 02a] D. Povey, P.C. Woodland: Minimum Phone Error and I-smoothing for Improved Discriminative Training. Vol. 1, pp. 105 – 108, Orlando, FL, May 2002.
- [Povey & Woodland 02b] D. Povey, P.C. Woodland: Minimum Phone Error and I-smoothing for improved discriminative training. In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1, pp. I–105–I–108, May 2002.
- [Rabiner 89] L.R. Rabiner: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, pp. 257–286, Feb. 1989.

- [Rabiner & Juang 86] L. Rabiner, B.H. Juang: An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, Vol. 3, No. 1, pp. 4–16, 1986.
- [Rabiner & Schafer 78] L.R. Rabiner, R.W. Schafer: *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ, 1978.
- [Ramasubramansian & Paliwal 92] V. Ramasubramansian, K.K. Paliwal: Fast k -dimensional Tree Algorithms for Nearest Neighbor Search with Application to Vector Quantization Encoding. Vol. 40, No. 3, pp. 518–528, March 1992.
- [Rumelhart & Hinton⁺ 86] D.E. Rumelhart, G.E. Hinton, R.J. Williams: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1. chapter Learning Internal Representations by Error Propagation, pp. 318–362. MIT Press, Cambridge, MA, USA, 1986.
- [Rybach & Gollan⁺ 09] D. Rybach, C. Gollan, R. Schlüter, H. Ney: Audio Segmentation for Speech Recognition using Segment Features. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 4197–4200, Taipei, Taiwan, April 2009.
- [Sak & Senior⁺ 14] H. Sak, A.W. Senior, F. Beaufays: Long Short-term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling. In *INTERSPEECH*, pp. 338–342. ISCA, 2014.
- [Sakoe 79] H. Sakoe: Two-Level DP-Matching - A Dynamic Programming-Based Pattern Matching Algorithm for Connected Word Recognition. Vol. 27, pp. 588–595, Dec. 1979.
- [Schlüter 00] R. Schlüter: *Investigations on Discriminative Training Criteria*. Ph.D. thesis, RWTH Aachen University, Aachen, Germany, Sept. 2000.
- [Schlüter & Nußbaum-Thom⁺ 10] R. Schlüter, M. Nußbaum-Thom, H. Ney: On the relation of Bayes risk, word error, and word posteriors in ASR. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 230–233, 2010.
- [Schlüter & Nußbaum-Thom⁺ 11] R. Schlüter, M. Nußbaum-Thom, H. Ney: On the Relationship Between Bayes Risk and Word Error Rate in ASR. *IEEE Trans. Audio, Speech & Language Processing*, Vol. 19, No. 5, pp. 1103–1112, 2011.
- [Schlüter & Nußbaum-Thom⁺ 12] R. Schlüter, M. Nußbaum-Thom, H. Ney: Does the Cost Function Matter in Bayes Decision Rule? *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 34, No. 2, pp. 292–301, 2012.
- [Schlüter & Nußbaum-Thom⁺ 13] R. Schlüter, M. Nußbaum-Thom, E. Beck, T. Alkhoul, H. Ney: Novel tight classification error bounds under mismatch conditions based on f-Divergence. In *2013 IEEE Information Theory Workshop, ITW 2013, Sevilla, Spain, September 9-13, 2013*, pp. 1–5, 2013.
- [Schwartz & Austin 91] R. Schwartz, S. Austin: A Comparison of Several Approximate Algorithms for Finding Multiple (N -Best) Sentence Hypotheses. Vol. 1, pp. 701–704, Toronto, Canada, May 1991.
- [Schwartz & Chow 90] R. Schwartz, Y.L. Chow: The N -Best Algorithm: An Efficient and Exact Procedure for Finding the N most likely Sentence Hypotheses. pp. 81–84, Albuquerque, NM, April 1990.

- [Seide & Li⁺ 11] F. Seide, G. Li, D. Yu: Conversational Speech Transcription Using Context-Dependent Deep Neural Networks. In *Interspeech 2011*, pp. 437–440. International Speech Communication Association, August 2011.
- [Shaik & Tüske⁺ 14] M.A.B. Shaik, Z. Tüske, M.A. Tahir, M. Nußbaum-Thom, R. Schlüter, H. Ney: RWTH LVCSR systems for quaero and EU-bridge: German, Polish, Spanish and Portuguese. In *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pp. 973–977, 2014.
- [Shaik & Tüske⁺ 15] M.A.B. Shaik, Z. Tüske, M.A. Tahir, M. Nußbaum-Thom, R. Schlüter, H. Ney: Improvements in RWTH LVCSR evaluation systems for Polish, Portuguese, English, Urdu, and Arabic. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pp. 3154–3158, 2015.
- [Sixtus 03] A. Sixtus: *Across-Word Phoneme Models for Large Vocabulary Continuous Speech Recognition*. Ph.D. thesis, RWTH Aachen, Jan. 2003.
- [Srivastava & Hinton⁺ 14] N. Srivastava, G.E. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov: Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1929–1958, 2014.
- [Steinbiss 91] V. Steinbiss: A search organization for large-vocabulary recognition based on n-best decoding. 01 1991.
- [Steinbiss & Ney⁺ 93] V. Steinbiss, H. Ney, R. Häb-Umbach, B. Tran, U. Essen, R. Kneser, M. Oerder, H. Meier, X. Aubert, C. Dugast, D. Geller: The Philips Research System for Large-Vocabulary Continuous-Speech Recognition. pp. 2125–2128, Berlin, Germany, Sept. 1993.
- [Stolcke 02] A. Stolcke: SRILM – An extensible language modeling toolkit. In *IN PROCEEDINGS OF THE 7TH INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING (ICSLP 2002)*, pp. 901–904, 2002.
- [Stolcke & König⁺ 97] A. Stolcke, Y. König, M. Weintraub: Explicit word error minimization in n-best list rescoring. In *Fifth European Conference on Speech Communication and Technology, EUROSPEECH 1997, Rhodes, Greece, September 22-25, 1997*, 1997.
- [Stolcke & yuh Hwang⁺ 06] A. Stolcke, M. yuh Hwang, X. Lei, N. Morgan, D. Vergyri: Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons. In *In Proceedings of ICASSP*, pp. 321–324, 2006.
- [Sundermeyer & Nußbaum-Thom⁺ 11] M. Sundermeyer, M. Nußbaum-Thom, S. Wiesler, C. Plahl, A.E. Mousa, S. Hahn, D. Nolden, R. Schlüter, H. Ney: The RWTH 2010 Quaero ASR evaluation system for English, French, and German. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2011, May 22-27, 2011, Prague Congress Center, Prague, Czech Republic*, pp. 2212–2215, 2011.
- [Tahir & Nußbaum-Thom⁺ 12] M.A. Tahir, M. Nußbaum-Thom, R. Schlüter, H. Ney: Simultaneous Discriminative Training and Mixture Splitting of HMMs for Speech Recognition. In *INTERSPEECH 2012, 13th Annual Conference of the International Speech Communication Association, Portland, Oregon, USA, September 9-13, 2012*, pp. 571–574, 2012.
- [Valente & Hermansky 08] F. Valente, H. Hermansky: Hierarchical and parallel processing of modulation spectrum for ASR applications. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4165–4168, March 2008.

- [Vapnik 98] V. Vapnik: *Statistical learning theory*. Wiley, 1998.
- [Vintsyuk 71] T.K. Vintsyuk: Elementwise Recognition of Continuous Speech Composed of Words from a Specified Dictionary. *Kibernetika*, Vol. 7, pp. 133–143, March 1971.
- [Viterbi 67] A. Viterbi: Error Bounds for Convolutional Codes and an Asymptotically Optimal Decoding Algorithm. *IEEE Transactions on Information Theory*, Vol. 13, pp. 260–269, 1967.
- [Wessel & Schlüter⁺ 01] F. Wessel, R. Schlüter, H. Ney: Explicit word error minimization using word hypothesis posterior probabilities. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2001, 7-11 May, 2001, Salt Palace Convention Center, Salt Lake City, Utah, USA, Proceedings*, pp. 33–36. IEEE, 2001.
- [Wiesler 16] S.B. Wiesler: *Optimization of Discriminative Models for Speech and Handwriting Recognition*. Ph.D. thesis, RWTH Aachen University, Computer Science Department, RWTH Aachen University, Aachen, Germany, Dec. 2016.
- [Wiesler & Heigold⁺ 10] S. Wiesler, G. Heigold, M. Nußbaum-Thom, R. Schlüter, H. Ney: A discriminative splitting criterion for phonetic decision trees. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pp. 54–57, 2010.
- [Woodland & Povey 00] P.C. Woodland, D. Povey: Large scale discriminative training for speech recognition. In *Automatic Speech Recognition (ASR)*, pp. 7 – 16, Paris, France, Sept. 2000.
- [Woodland & Povey 02] P.C. Woodland, D. Povey: Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition. Vol. 16, No. 1, pp. 25–48, 2002.
- [Young 92] S.J. Young: The General Use of Tying in Phoneme Based HMM Recognizers. Vol. 1, pp. 569–572, San Francisco, CA, March 1992.
- [Zheng & Stolcke 05] J. Zheng, A. Stolcke: Improved discriminative training using phone lattices, 2005.