CHAPTER 8

Epilogue

8.1 SUMMARY OF BOOK CONTENTS

This book starts by providing an introduction to discriminative learning, speech recognition, and the roles of discriminative learning in speech recognition. Then it presents some background material on basic probability distributions and on optimization techniques; both serve as mathematical requisites for the remaining book content dealing with detailed techniques for discriminative learning in speech recognition. The basic probability distributions covered in the background material include multinomial distribution and multivariate Gaussian distribution, both belonging to the more general exponential-distribution family, as well as Gaussian mixture distribution, which is outside of the exponential-distribution family. The optimization concepts and techniques covered in the background material include definitions of global and local optimums, their necessary condition, Lagrange multiplier method, gradient-based method, and, finally, growth transformation (GT) method.

The book then proceeds by a tutorial on statistical speech recognition, where the hidden Markov model (HMM) is formally introduced as a popular acoustic model for speech feature sequences, and the language model is also introduced as the prior probability of word sequences. Introduction of the HMM sets up the context in which discriminative learning, as the main subject of this book, is subsequently discussed in a great detail.

Given the concepts of acoustic modeling and language modeling, in this book, we then provide a unified account for three common objective functions for discriminative training of HMMs currently in use in speech recognition practice, including maximum mutual information (MMI), minimum classification error (MCE), and minimum phone error/minimum word error (MPE/MWE). Comparisons are made between our unified form of the objective functions with another unified form of discriminative objective function in literature and insights are offered in the comparisons.

The subsequent materials covered in this book focus on ways of carrying out discriminative parameter learning using the unified form of objective functions via the GT technique. The coverage starts with a relatively simple case where the IID assumption for data observation (or model stationarity) is made and where exponential-family distributions are assumed for the data. Next, the more complex, nonstationary model for sequentially correlated data observations is discussed.

HMMs are used in this case and its parameter estimation problem via the GT technique is presented in detail.

Some practical implementation issues of the GT technique for HMM parameter learning are then discussed, filling in some details that were not dealt with in the preceding portions so as not to unnecessarily divert the main topics. Then, finally, we present some selected experimental results in speech recognition, demonstrating the effectiveness of the techniques presented in this book in practice.

8.2 SUMMARY OF CONTRIBUTIONS

The main technical contribution of this book is to provide three aspects of unification of the common discriminative learning techniques for sequential pattern recognition, in particular, those for speech recognition.

- First, the unification is in the objective function for optimization, which has been derived rigorously to be a rational functional form of (3.2). Although the rational functional form for MMI has been known in the past, we provide the first proof that the same form applies to MCE and MPE/MWE, differing from MMI only in the constant weighting factors.
- Second, the unification is in the optimization technique. The unified, rational functional form of the objective function for MMI, MCE, and MPE/MWE enables the use of the special technique of GT/extended Baum–Welch for optimization. In the past, MCE had always been implemented by gradient descent, due to the lack of any rational functional form in its objective function. The essence of the GT technique is to optimize the specially constructed auxiliary function, for both the discrete valued HMM and for the continuous valued Gaussian HMM.
- Third, the unification is in the final parameter reestimation formulas. The formula for the HMM transition probability is shown to be (5.26). The formula for the discrete HMM's emitting probability is shown to be (5.21). The formulas for the Gaussian HMM is shown to be (5.35) and (5.36) for its mean vectors and covariance matrices, respectively.

The unifying review of discriminative learning for HMMs provided in this book is motivated by the striking success of such various techniques in recent speech recognition research and system development. Yet, there has been a conspicuous lack of common understanding of the relationships among these techniques including MMI, MCE, and MPE/MWE, despite the relatively long history of MMI (since 1987 [6]), MCE (since 1992 [24]), and MPE/MWE (since 2002 [38]).

Because of the complexity of these techniques and the lack of a unifying theoretical theme underlying them, only a very small number of speech recognition laboratories worldwide have been able to successfully implement these techniques and to achieve similarly strong performance gains for large vocabulary speech recognition. The main goal of this paper is to provide just such a unifying theoretical framework in hopes of promoting more widespread use of discriminative learning not only in speech recognition, but possibly in other types of sequential pattern recognition and signal processing problems as well. It is also hoped that given a solid theoretical foundation presented in this book, other more advanced pattern recognition concepts (e.g., discriminative margins [48]) can be more elegantly integrated with current discriminative learning techniques. The new goal then is not only to reduce empirical errors but also to enhance generalization capabilities.

In this book, we show in a step-by-step fashion that our approach leads to consistent parameter estimation results and it is scalable for large-scale pattern recognition tasks. We also analyzed the algorithmic properties of the MCE- and MPE/MWE-based learning methods under the GT parameter estimation framework for sequential pattern recognition using HMMs.

8.3 REMAINING THEORETICAL ISSUE AND FUTURE DIRECTION

The material covered in this book is probably the most comprehensive one on the topic of discriminative learning designed for sequential pattern recognition such as speech recognition. One important theoretical issue for the CDHMM concerns the convergence properties of the GT method. In Section 5.3, we discussed this issue in depth, where we outlined a proof (based on Axelrod et al.'s work [3]) that the GT update formulas for the CDHMM are valid given a sufficiently large (but bounded) constant D_i. However, in that analysis, no explicit construction of D_i was given. Therefore, it constitutes only an existence proof. The remaining issue is whether one can provide a constructive proof. In this final section, we outline one constructive proof for advanced readers, based on Jebera's work [22, 23], where reverse Jensen's inequality was used for optimization.

In principle, Jebara's method is applicable to maximizing any rational function, whose numerator and denominator can be mixtures of exponential models. Therefore, it is applicable to optimizing our unified discriminative criterion (3.2) for all MMI, MCE, and MPE/MWE. In the brief review below, we introduce the principle of reverse Jensen's inequality and its application to discriminative objective function optimization.

Q1

For a rational function in the form of (1.26) and (1.27), we desire to maximize the following equivalent function:

$$\log O(\Lambda) = \log G(\Lambda) - \log H(\Lambda) = \log \sum_s p(X,s|\Lambda)\ C(s) - \log \sum_s p(X,s|\Lambda) \qquad (8.1)$$

The first term in (8.1) is a log-sum function similar to log likelihood. Based on the well-known Jensen's inequality and after several steps of simplifications, we have

$$\log \sum_s p(X,s|\Lambda)\,C(s) \geq Q_G(\Lambda;\Lambda') + J$$

$$= \underbrace{\sum_s \left(\frac{p(X,s|\Lambda)C(s)}{\sum_s p(X,s|\Lambda')\,C(s)} \right) \log p(X,s|\Lambda)\,C(s)}_{Q_G(\Lambda;\Lambda')} + J \qquad (8.2)$$

where J is a constant irrelevant to Λ (although relevant to Λ'), that is, $J = \log\sum_s p(X, s|\Lambda')C(s) - Q_G(\Lambda'; \Lambda')$. This is similar to the E-step in the EM algorithm.

The left-hand side of (8.2) is a lower bound of $\log G(\Lambda)$, and makes tangential contact with $\log G(\Lambda)$ at Λ'. Therefore maximizing the auxiliary function $Q_G(\Lambda; \Lambda')$ guarantees increase of $\log G(\Lambda)$ iteratively.

However, to maximize $\log O(\Lambda)$, we need a lower bound for $\log O(\Lambda)$, which in turn requires an upper bound of $\log H(\Lambda)$. In [22], it was shown (nontrivially) that using reverse Jensen's inequality an auxiliary function $Q_H(\Lambda; \Lambda')$ can be constructed so that

$$\log \sum_s p(X,s|\Lambda) \sum_s p(X,s|\Lambda) \leq Q_H(\Lambda;\Lambda') + \tilde{J} = \underbrace{\sum_s (-w_s) \log p(Y_s,s|\Lambda)}_{Q_H(\Lambda;\Lambda')} + \tilde{J} \qquad (8.3)$$

where $\tilde{J}$ is a Λ-irrelevant constant that makes $Q_H(\Lambda; \Lambda') + \tilde{J}$ tangential contact with $\log H(\Lambda)$ at Λ', and w_s and Y_s are positive weights and modified observations, respectively. Reverse Jensen's inequality was derived by exploiting the convexity of the cumulant generating function of exponential family in [22] and will not be elaborated further here.

Given (8.2) and (8.3), one can construct the auxiliary function for $\log O(\Lambda)$ as:

$$Q(\Lambda;\Lambda') = Q_G(\Lambda;\Lambda') - Q_H(\Lambda;\Lambda') \qquad (8.4)$$

where $Q(\Lambda; \Lambda')$ is a lower bound of $\log O(\Lambda)$ and makes tangential contact with $\log O(\Lambda)$ at Λ'. Therefore, optimizing $\log O(\Lambda)$ can be achieved by iteratively optimizing $Q(\Lambda; \Lambda')$, which takes the same step as the M-step in the conventional EM algorithm for an HMM (i.e., with a closed-form solution in the M-step).

Note for our unifying discriminative objective function (3.2), the summand of $G(\Lambda)$ may take a negative value for MPE; that is, for some paths s that have many insertion errors, the corresponding $C(s)$ may be negative. In this case, we can add extra dummy training tokens to the training set,

while these dummy tokens can only be recognized as correct references. Appending these dummy tokens to s can effectively increase its raw accuracy count to be positive. Moreover, because the dummy token will not compete with any other tokens in the training set, it will not affect the training performance.

Applications of reverse Jensen's inequality for discriminative training have been an interesting research area recently. In [1], the method for MMI optimization was explored and was compared with the conventional GT method using empirical setting of the constant D_i. After approximation and simplification, the author showed similar forms of model estimation formulas to those derived from the empirical GT method, but with a larger D_i, and slower convergence. Further investigation of the method based on reverse Jensen's inequality for discriminative training is warranted, and this constitutes one fruitful theoretical research direction for full maturity of learning algorithms for discriminative training in the statistical recognizer design.

• • • •

Author Query Form

(Queries are to be answered by the Author)

He – Chapter 8

The following queries have arisen during the typesetting of your manuscript. Please answer these queries.

Query Marker	Query	Reply
Q1	"nominator" was changed to "numerator" – Please check if ok.	YES

Thank you very much.