ISR
INSTITUTE FOR SYSTEMS RESEARCH

TECHNICAL RESEARCH REPORT

An Information Geometric Treatment of Maximum Likelihood Criteria and Generalization in Hidden Markov Modeling

by W. Byrne

T.R. 93-50

The Institute for Systems Research is supported by the
National Science Foundation Engineering Research Center Program (NSFD CD 8803012),
the University of Maryland, Harvard University, and Industry

An Information Geometric Treatment of Maximum Likelihood Criteria and Generalization in Hidden Markov Modeling

William Byrne
Institute for Systems Research and
Department of Electrical Engineering
University of Maryland
College Park, MD 20742
bbyrne@src.umd.edu

Abstract

It is shown here that several techniques for maximum likelihood training of Hidden Markov Models are instances of the EM algorithm and have very similar descriptions when formulated as instances of the Alternating Minimization procedure. The N-Best and Segmental K-Means algorithms are derived under a minimum discrimination information criterion and are shown to result from an additional restriction placed on the minimum discrimination information formulation which yields the Baum Welch algorithm. This uniform formulation is employed in an exploration of generalization by the EM algorithm.

It has been noted that the EM algorithm can introduce artifacts as training progresses. A related phenomenon is that over-training can occur: although the performance as measured on the training set continues to improve as the algorithm progresses, performance on related data sets may eventually begin to deteriorate. This is inherent in the maximum likelihood criterion and its cause can be seen when the training problem is stated in the Alternating Minimization framework. A modification of the maximum likelihood training criterion is suggested to counter this behavior and is applied to the broader problem of maximum likelihood training of exponential models from incomplete data. It leads to a simple modification of the learning algorithms which relates generalization to learning speed. Relationships to other techniques which encourage generalization, particularly methods of incorporating prior information, are discussed.

Contents

1 Introduction

Many methods are available for training Hidden Markov Models from data under a maximum likelihood criterion [1, 2, 3, 4]. It is shown here that several techniques for maximum likelihood training are instances of the EM [5] algorithm and have very similar descriptions when formulated as instances of the Alternating Minimization [6] technique. The value of this uniform formulation is shown in an exploration of generalization by the EM algorithm.

Unless otherwise noted, the Hidden Markov Models considered here have discrete observation densities. The observation symbols will be taken from the set $\{1, \ldots, M\}$, and the hidden states from the set $\{1, \ldots, N\}$. Observable sequences will be denoted I, and assumed to be of variable, finite, random length. The hidden state sequences will be denoted S. The set of observed sequences $\mathcal{I}$ and the set of hidden sequences $\mathcal{S}$ are countable. The models will be denoted Q and form a family of models $\mathcal{Q}$, which is a subset of the general set of distributions $\mathcal{P}$ on $\mathcal{I} \times \mathcal{S}$.

2 Maximum Likelihood Training of Hidden Markov Models

In Hidden Markov Modeling it is necessary in both recognition and training to evaluate the likelihood of an observed sequence based on a model. Likelihood criteria differ in how an observed sequence is "associated" with hidden state sequences.

2.1 Maximum Likelihood Criteria

One method of associating hidden variables with observations is through the marginal distribution

$$Q(I) = \sum_{S \in \mathcal{S}} Q(I, S). \tag{1}$$

This criterion assumes that there is no particular underlying sequence that should be associated with an observed sequence and that all are equally good.

Another association relies on finding the Most Likely State Sequence (MLSS) [7, 8] for any observed sequence I:

$$S_1 : \ Q(I, S_1) = \max_{S \in \mathcal{S}} Q(I, S). \tag{2}$$

The choice of a MLSS is consistent with the use of HMMs as source models in which an underlying process controls the production of the observed sequence. The MLSS might be a good estimate of which underlying sequence produced the observed sequence.

A generalization of the MLSS approach is to find the best few hidden sequences for an observation [9, 10]. For an observed sequence I, the N most likely hidden sequences $S_1, \ldots, S_N$ are found so that

$$Q(I, S_i) \geq Q(I, S') \quad \forall S' \notin \{S_1, \ldots, S_N\}. \tag{3}$$

The likelihood of an observed sequence can be evaluated as

$$\sum_{i=1}^{N} Q(I, S_i). \tag{4}$$

The N-Best criterion acknowledges that there may be more than one underlying sequence which might be associated with an observed sequence.

Training algorithms can be derived which attempt to find models which perform well under these association criteria. The Baum-Welch (BW) algorithm [11] is based on the marginal distributions, while the Segmental K-Means (SKM) [12, 13], or Viterbi, training algorithm is based on the best sequence. The N-Best criterion has been used primarily to provide multiple hypotheses during recognition but it can also be used in training.

These training algorithms can be derived using the Alternating Minimization technique [6], which requires formulating HMM training as a problem in Minimum Discrimination Information (MDI) modeling [14, 15].

Ephraim and Rabiner [16] described the Baum-Welch algorithm, the Maximum Mutual Information criterion, and a general MDI algorithm as Minimum Discrimination Information modeling approaches which differ in the source being modeled and the statistics attributed to the source. In this work the MDI approach is extended to the likelihood criteria described above. The Segmental K-Means and N-Best criteria are used to define MDI modeling problems and it is shown that the algorithms which result exist on a continuum, with the BW and SKM at the end points and the N-Best algorithms distributed between them.

2.2 Alternating Minimization Procedure

Once the likelihood criterion is established, the Maximum Likelihood training problem is defined by the training data. The training set, T, consists of sequences of observations of a single source, e.g. several utterances of a particular word. The goal of Maximum Likelihood training is to construct a single HMM that describes the training set well. The number of hidden states and the allowable state transitions are assumed to be fixed in advance. Competitive, or discriminative, training in which sequences from different sources are used jointly to train multiple models is not treated here, although it is considered as a problem in MDI modeling in [16].

The Alternating Minimization procedure can be used to describe a search between two sets of distributions. One set, the model set $\mathcal{Q}$, is described above. The second set, $\mathcal{D}$, is called the *set of desirable distributions*. It is defined by the training data and its members are those distributions which satisfy the likelihood criterion.

The distributions in these sets will be compared using the information divergence as defined on countable sets. The divergence compares two distributions U and V according to

$$D(U \parallel V) = \sum_{x \in \mathcal{X}} U(x) \log \frac{U(x)}{V(x)}. \tag{5}$$

The desirability of a specific model, Q, is determined by its distance from the set of desirable distributions as measured by the divergence. The divergence defines the I-Projection [17] of Q onto $\mathcal{D}$ [1]

$$D(\mathcal{D} \parallel Q) = \min_{P \in \mathcal{D}} D(P \parallel Q). \tag{6}$$

Models closer to $\mathcal{D}$ are more desirable. Ideally, the goal of training would be to find the best of all models by solving

$$D(\mathcal{D} \parallel \mathcal{Q}) = \min_{Q \in \mathcal{Q}} \min_{P \in \mathcal{D}} D(P \parallel Q). \tag{7}$$

This is usually not practical, so suboptimal, or locally optimal procedures are used.

The Alternating Minimization procedure can be used to describe many such searches. An initial model Q^1 is chosen from the model family $\mathcal{Q}$. First, the I-Projection of Q^1 on $\mathcal{D}$ is computed

$$P^1 : \quad D(P^1 \parallel Q^1) = \min_{P \in \mathcal{D}} D(P \parallel Q^1). \tag{8}$$

A new model Q^2 is then found by solving

$$Q^2 : \quad D(P^1 \parallel Q^2) = \min_{Q \in \mathcal{Q}} D(P^1 \parallel Q). \tag{9}$$

It is shown in [6], that repeatedly applying this procedure produces a sequence of models which approaches the set of desired distributions

$$D(\mathcal{D} \parallel Q^{p+1}) \leq D(\mathcal{D} \parallel Q^p). \tag{10}$$

and that these steps form the EM algorithm [5]. I-projection onto the set of desired distributions corresponds to the E-step while the M-step corresponds to solving Equation 9.

[1] In all the problems considered here, the I-projection will be shown to belong to $\mathcal{D}$ and so is defined by a minimum rather than an infimum.

2.3 Desired Distributions

2.3.1 Definitions

The following definitions are used in describing the desired distributions for each algorithm and also in describing the I-Projection onto each set.

In the algorithms discussed here, the desired distributions are defined using an empirical distribution $\hat{P}$ on the observation set $\mathcal{I}$

$$\hat{P}(I) = \frac{\#_T(I)}{|T|} \tag{11}$$

where $\#_T(I)$ is the number of times I appears in the training set T.

Define the support of $\hat{P}$ to be $\mathcal{T} = \{I : I \in T\}$.

With Q^p fixed, for each training sequence I define $B_N^p(I) = \{S_1, \ldots, S_N\}$ to be the set of N-Best hidden state sequences, as measured by Q^p:

$$Q^p(I, S_i) \geq Q^p(I, S') \qquad \forall S' \notin B_N^p(I). \tag{12}$$

The notation $Q^p(I, B_N^p(I))$ denotes $\sum_{S \in B_N^p(I)} Q^p(I, S)$, the N-Best criterion defined in Equation 4.

Define the conditional distribution $Q^p(S|B_N^p(I), I)$ on $\mathcal{S}$ as

$$for\ I \in \mathcal{T} : \ Q^p(S|B_N^p(I), I) = \begin{cases} Q^p(I,S)/Q^p(I, B_N^p(I)) & S \in B_N^p(I) \\ 0 & \text{otherwise} \end{cases} . \tag{13}$$

This is a valid conditional distribution on $\mathcal{S}$ for $I \in \mathcal{T}$.

For each observation $I \in \mathcal{T}$ and each conditional probability $P(S|I)$ of $P \in \mathcal{D}$, define

$$P^{-1}(\cdot|I) = \{S \in \mathcal{S} : P(S|I) > 0\}. \tag{14}$$

This random set is the support of the conditional distribution and is a valid event, i.e. $P^{-1}(\cdot|I) \in \sigma(\mathcal{S})$. For a given $P \in \mathcal{D}$, the notation $Q(I, P^{-1}(\cdot|I))$ denotes $\sum_{S \in P^{-1}(\cdot|I)} Q(I, S)$.

2.3.2 The Baum Welch Algorithm

The set of desirable distributions which defines the Baum Welch algorithm is

$$\mathcal{D}^{BW} = \{P \in \mathcal{P} : \sum_S P(I, S) = \hat{P}(I)\}. \tag{15}$$

Because $\hat{P}(I) = 0 \ \forall I \notin \mathcal{T}$, it is sufficient to insist that the above be satisfied only for $I \in \mathcal{T}$:

$$\mathcal{D}^{BW} = \{P \in \mathcal{P} : P(I) = \hat{P}(I) \ \forall I \in \mathcal{T}\}. \tag{16}$$

It is shown elsewhere (e.g. Appendix A) that the I-projection of Q^p onto $\mathcal{D}^{BW}$ is

$$P^p(I, S) = \hat{P}(I)\, Q^p(S|I). \tag{17}$$

The I-Projection satisfies $P^p(I) = \hat{P}(I)$ and also has the property that $P^p(S|I) = Q^p(S|I)$. The updated models are improved by the Baum Welch algorithm in that

$$\sum_{I \in \mathcal{T}} \hat{P}(I) \ \log Q^p(I) \leq \sum_{I \in \mathcal{T}} \hat{P}(I) \ \log Q^{p+1}(I). \tag{18}$$

2.3.3 The Segmental K-Means Algorithm

The set of desirable distributions associated with the SKM algorithm is a subset of $\mathcal{D}^{BW}$. It can be defined as

$$\mathcal{D}^{SKM} = \{P \in \mathcal{P} : P(I) = \hat{P}(I) \quad \text{and } |P^{-1}(\cdot|I)| = 1 \ \forall I \in \mathcal{T}\}. \tag{19}$$

Because $|P^{-1}(\cdot|I)| = 1$, every P in $\mathcal{D}^{SKM}$ associates each training sequence with a single hidden sequence. For a fixed P and I, call this sequence S_I and note that $P(S_I|I) = 1$. To see that this set of desirable distributions leads to the SKM algorithm, consider $D(P \,||\, Q^p)$ for $P \in \mathcal{D}^{SKM}$:

$$\begin{aligned} D(P \,||\, Q^p) &= \sum_{I,S} P(I,S) \log \frac{P(I,S)}{Q^p(I,S)} && (20) \\ &= \sum_I \sum_S P(I)P(S|I) \ \log \hat{P}(I)P(S|I) - \sum_I \sum_S P(I)P(S|I) \ \log Q^p(I,S) \\ &= \sum_{I \in \mathcal{T}} \hat{P}(I) \log \hat{P}(I) - \sum_{I \in \mathcal{T}} \hat{P}(I) \log Q^p(I, S_I). && (21) \end{aligned}$$

To minimize this expression, P^p should be chosen so that each observed sequence in the training set is associated with its MLSS.

The projection from Q^p to $\mathcal{D}^{SKM}$ is therefore found by performing a maximum likelihood alignment of each training sequence with a hidden state sequence under Q^p:

$$P^p(I,S) = \begin{cases} \hat{P}(I) & S = \text{argmax}_S Q^P(I,S), \ I \in \mathcal{T} \\ 0 & \text{otherwise} \end{cases}. \tag{22}$$

For $I \in \mathcal{T}$, $P^p(S|I) = 1_{B_1^p(I)}(S)$. The improvement guaranteed by the SKM algorithm is

$$\sum_{I \in \mathcal{T}} \hat{P}(I) \ \log \max_S Q^p(I,S) \leq \sum_{I \in \mathcal{T}} \hat{P}(I) \ \log \max_S Q^{p+1}(I,S). \tag{23}$$

Although the Forward-Backward algorithm used to solve Equation 17 is very different from the Viterbi search usually used to find the MLSS and Equation 22, the Baum Welch and SKM algorithms are seen to be instances of the Alternating Minimization technique which are distinguished by constraints on the set of desirable distributions. This is touched upon in [7], where it is shown that under certain conditions, these two techniques asymptotically yield identical results; the additional restriction which defines $\mathcal{D}^{SKM}$ is shown to be unimportant under some conditions.

2.3.4 The N-Best Algorithm

The N-Best criterion associates an observed sequence with its N Most Likely State Sequences. The likelihood that follows from this association is $Q^p(I, B_N^p(I))$. To obtain an algorithm based on this criterion, define the set of desirable distributions as

$$\mathcal{D}^N = \{P \in \mathcal{P} : P(I) = \hat{P}(I) \quad \text{and } |P^{-1}(\cdot|I)| \leq N \ \ \forall I \in \mathcal{T}\}. \tag{24}$$

The I-Projection of Q^p onto $\mathcal{D}^N$ is

$$P^p(I,S) = \hat{P}(I) \ Q^p(S|B_N^p(I), I). \tag{25}$$

To show this, it is sufficient first to find a lower bound on $D(\mathcal{D}^N \,||\, Q^p)$, and then to show that $D(P^p \,||\, Q^p)$ achieves this bound.

Consider $D(P \parallel Q^p)$ for $P \in \mathcal{D}^N$:

$$\begin{aligned} D(P \parallel Q^p) &= \sum_{I,S} P(I,S) \log \frac{P(I,S)}{Q^p(I,S)} && (26) \\ &= \sum_{I} \sum_{S \in P^{-1}(\cdot|I)} P(I,S) \log \frac{P(I,S)}{Q^p(I,S)} && (27) \\ &\geq \sum_{I} P(I) \log \frac{P(I)}{\sum_{S \in P^{-1}(\cdot|I)} Q^p(I,S)} && (28) \\ &= \sum_{I} \hat{P}(I) \log \frac{\hat{P}(I)}{Q^p(I, P^{-1}(\cdot|I))} && (29) \end{aligned}$$

where the inequality follows by applying the log-sum inequality [18] [2] to the inner sum. Equality holds if

$$\frac{P(I,S)}{P(I,P^{-1}(\cdot|I))} = \frac{Q^p(I,S)}{Q^p(I,P^{-1}(\cdot|I))} \qquad (30)$$

or equivalently, using $P(I,P^{-1}(\cdot|I)) = P(I)$,

$$P(S|I) = \frac{Q^p(I,S)}{Q^p(I,P^{-1}(\cdot|I))}. \qquad (31)$$

By the definition of the $B_N^p(I)$ (Equation 12), for $P \in \mathcal{D}^N$

$$Q^p(I,P^{-1}(\cdot|I)) \quad \leq \quad Q^p(I,B_N^p(I)) \quad \forall I \in \mathcal{T}. \qquad (32)$$

This and Equation 29 establishes a lower bound for $D(\mathcal{D}^N \parallel Q^p)$:

$$D(P \parallel Q^p) \geq \sum_{I} \hat{P}(I) \log \frac{\hat{P}(I)}{Q^p(I, B_N^p(I))} \quad P \in \mathcal{D}^N \qquad (33)$$

where equality holds if Equation 31 is satisfied and $P^{-1}(\cdot|I) = B_N^p(I)$.

P^p as defined above belongs to $\mathcal{D}^N$ and satisfies both the conditions for achieving the lower bound. For $I \in \mathcal{T}$, $P^p(S|I) = Q^p(S|B_N^p(I), I)$.

Projection of Q^p onto $\mathcal{D}^N$ requires finding the N-Best sequences for each training sequence and determining $P^p(S|I)$ in proportion to their likelihood under Q^p. The algorithm guarantees improvement in that

$$\sum_{I \in \mathcal{T}} \hat{P}(I) \; \log Q^p(I, B_N^p(I)) \leq \sum_{I \in \mathcal{T}} \hat{P}(I) \; \log Q^{p+1}(I, B_N^{p+1}(I)) \;. \qquad (34)$$

2.4 A Uniform Formulation of the Algorithms

This description of the Baum Welch, Segmental K-Means, and N-Best maximum likelihood training algorithms can be summarized as searches over distributions which belong to a general set of desirable distributions

$$\mathcal{D} = \{P \in \mathcal{P} : \sum_{S} P(I,S) = \hat{P}(I) \quad \forall I \in \mathcal{T}\}. \qquad (35)$$

This is $\mathcal{D}^{BW}$, the set of desired distributions associated with the BW algorithm. The other desirable distributions are

$$\begin{aligned} D^{SKM} &= \mathcal{D} \cap \{P : |P^{-1}(\cdot|I)| = 1\} && (36) \\ D^N &= \mathcal{D} \cap \{P : |P^{-1}(\cdot|I)| \leq N\}. && (37) \end{aligned}$$

Table 1: HMM Maximum Likelihood Criteria and Training Algorithms

Algorithm	Likelihood Assigned to Training Items	Restrictions on $P \in \mathcal{D}$ to Form $\mathcal{D}^N$	I-Projection of Q^p onto $\mathcal{D}^n$ $P^p(I,S)$
Baum Welch	$\log Q(I)$	–	$Q^p(S\|I)\ \hat{P}(I)$
N-Best	$\overset{\max}{S_1,\ldots S_N} \sum_{i=1}^{N} \log Q(I,S_i)$	$\|P^{-1}(\cdot\|I)\| \leq N$	$Q^p(S\|B_N^p(I),I)\ \hat{P}(I)$
SKM	$\overset{\max}{S} \log Q(I,S)$	$\|P^{-1}(\cdot\|I)\| = 1$	$\hat{P}(I)\ \ S = \arg\max_{S'} Q(I,S')$

The algorithms and their properties are given in Table 1.

These algorithms are similar in that the desired distributions satisfy $P(I) = \hat{P}(I)$. They differ in a restriction on the size of the support of the desired *a posteriori* distributions. This restriction is the requirement that all desired distributions satisfy $|P^{-1}(\cdot|I)| \leq N$ for a fixed N. For $N = 1$, the SKM algorithm results. For $N = \infty$, i.e. for $|P^{-1}(\cdot|I)|$ unrestricted, the BW algorithm is obtained. The positive integers can therefore be used to specify the likelihood criteria and determine the training algorithms.

The added restriction on $\mathcal{D}$ weakens with increasing N so that

$$\mathcal{D}^{SKM} \subset \mathcal{D}^N \subset \mathcal{D}^{BW}. \tag{38}$$

This implies that

$$D(\mathcal{D}^{SKM} \parallel \mathcal{Q}) \geq D(\mathcal{D}^N \parallel \mathcal{Q}) \geq D(\mathcal{D}^{BW} \parallel \mathcal{Q}) \quad . \tag{39}$$

This describes the relationship among globally optimum solutions to the training problem as defined in Equation 7. The locally optimum solutions found by the Alternating Minimization procedures may not obey this relationship.

The model set of HMMs and the desirable distributions are defined very differently. The HMMs are defined in a "bottom up" manner as $Q(I,S) = Q(S)Q(I|S)$, that is, as generative models in which S produces I. Typically both $Q(S)$ and $Q(I|S)$ can be freely varied so long as $Q \in \mathcal{Q}$. Conversely, the desired distributions are defined in a "top down" manner in terms of $P(I)$ and $P(S|I)$ so that $P(I,S) = P(I)P(S|I)$. Once the training data is available, $P(I)$ is fixed as $\hat{P}(I)$ and only the term $P(S|I)$ varies.

2.5 Restatement of the Likelihood Criteria

Rather than use the parameterization $N = 1, 2, \ldots$, to describe the algorithms, an *association parameter* $a \in [0,1]$ can be used to define the desired distributions:

$$\mathcal{D}^a = \mathcal{D} \cap \{P \in \mathcal{P} : |P^{-1}(\cdot|I)| \leq \frac{1}{a} \quad \forall I \in \mathcal{T}\}. \tag{40}$$

For $a = 0$, $\mathcal{D}^a = \mathcal{D}^{BW}$ and for $a = 1$, $\mathcal{D}^a = \mathcal{D}^{SKM}$. If N is the largest integer so that $N \leq \frac{1}{a}$ then $\mathcal{D}^a = \mathcal{D}^N$ for that value of N.

The association parameter determines how many "best" sequences are considered in scoring a observed sequence: for $a = 1$, there is one best sequence; for $a = 0$, none are considered best; for intermediate values of a, a collection of best candidates is considered. Using a avoids using infinity as an index.

[2] For $a_i \geq 0$ and $b_i \geq 0$, if $\sum_i a_i = a$ and $\sum_i b_i = b$, then $\sum_i a_i \log \frac{a_i}{b_i} \geq a \log \frac{a}{b}$ with equality iff $\frac{a_i}{a} = \frac{b_i}{b}$.

For convenience in the subsequent discussion of maximum likelihood training it will be assumed that the association parameter is fixed. Rather than redefine everything in terms of a, the notation based on N will be kept. The likelihood criterion is denoted

$$L_N^p(I) = \log Q^p(I, B_N^p(I)) \quad \forall I \in \mathcal{T}. \tag{41}$$

For $N = 1$, this denotes the MLSS criterion $L_N^p(I) = \log \max_S Q^p(I,S)$; for $N = \infty$, it denotes the BW criterion $L_N^p(I) = \log Q^p(I)$; for intermediate values of N, the corresponding N-Best criterion is meant.

The general likelihood criterion is then to find models which improve according to

$$\sum_{I \in \mathcal{T}} \hat{P}(I) \;\; L_N^p(I) \leq \sum_{I \in \mathcal{T}} \hat{P}(I) \;\; L_N^{p+1}(I). \tag{42}$$

3 Generalization in Maximum Likelihood Training

In many applications the observations which form the training set T are only a small subset of the possible values which the observed variable can assume. In the following arguments it is assumed that there is significant, if unknown, data not included in the training set which should be acknowledged. This data is referred to as $\mathcal{T}^c = \{I \in \mathcal{I} : I \notin T\}$ to denote the values not found in the training set ($\mathcal{T}^c \cup \mathcal{T} = \mathcal{I}$). Although training algorithms can only exploit data available in T, the resulting model is intended to describe accurately related data not found in the training set. If a training procedure produces a model with this property, it is said to *generalize* well from the training data.

One of the characteristics of the desired distributions is that their "visible support" is restricted to the training set. For $P \in \mathcal{D}$, it follows from Equations 11 and 35 that $P(\mathcal{T}) = 1$. This implies that

$$P \in \mathcal{D}: \quad P(I,S) = 0 \quad \forall I \in \mathcal{T}^c, \forall S \tag{43}$$

because for $I \in \mathcal{T}^c$, $P(I,S) \leq P(I) \leq P(\mathcal{T}^c) = 0$. This is pointed out in [19] in the estimation of point-process intensities.

If a training algorithm were to succeed in finding an exact, optimum solution, that is, to find a model Q^p such that $D(\mathcal{D} \parallel Q^p) = 0$, this model would necessarily belong to $\mathcal{D}$. Although the model would be optimum according to the criterion which defines the algorithm, it would be unable to generalize about data not in the training set. In a sense, the ability of such an algorithm to produce models which generalize well requires that the algorithm not achieve its training goal.

While it is unlikely that the algorithm will ever find anything but suboptimal solutions which do not belong to $\mathcal{D}$ (there are exceptions [20]), it is possible for the resulting models to be overtrained in the following sense. Suppose that an additional set of data V, called the validation set, is also available during training. The validation data defines a family of distributions $\mathcal{V}$ in the same way as the training data defines $\mathcal{D}$

$$\mathcal{V} = \{P \in \mathcal{P} : P(I) = \frac{\#_V(I)}{|V|}\}. \tag{44}$$

Ideally, while the training algorithm yields improving performance on the training set according to Equation 10, the performance on the validation set also improves

$$D(\mathcal{V} \parallel Q^{p+1}) \leq D(\mathcal{V} \parallel Q^p). \tag{45}$$

If it happens that at some iteration this relationship is violated the model is said to be overtrained, which is evidence of poor generalization. Methods which explicitly reserve some of the training set to perform such tests are called cross-validation [21], or holdout, methods and a procedure which halts the EM algorithm based on the test described above is described in [22].

3.1 A Modified Set of Desired Distributions

A possible way to avoid overtraining motivated by these considerations is to modify the general set of desired distributions (Equation 35) by introducing a *confidence parameter c*

$$\mathcal{D}_c = \{P \in \mathcal{P} : \sum_{S \in \mathcal{S}} P(I,S) = c \ \hat{P}(I) \quad \forall I \in \mathcal{T}\} \quad 0 \leq c \leq 1. \tag{46}$$

In this set of distributions the training set is given probability $P(\mathcal{T}) = c$, while all other possible observations are given probability $P(\mathcal{T}^c) = 1 - c$. This is prompted by the intuitive motivation for cross-validation, which is that training should not proceed as if there were no possibilities other than those included in the training data.

In typical applications of the EM algorithm the likelihood criterion completely specifies the desired distribution over the observed variable and the E-Step is used only to estimate the desired behavior of the hidden variables. Under the likelihood criterion presented here however, the likelihood criterion is incompletely specified outside the training set. No assumptions are made about the correct likelihood of the individual elements of $\mathcal{T}^c$, other than that the linear constraint $P(\mathcal{T}^c) = 1 - c$ is satisfied. As as result, the E-Step also estimates at each iteration the likelihood criterion where it is unspecified. Performing the I-Projection under the incomplete linear constraint leads to an estimate of the unspecified likelihood criterion according to the minimum discrimination information principle [14].

Another approach to generalization from data is exemplified by [23]. The data set is analyzed to attempt to match the power of the model to the data with the goal of finding the smallest model which accurately describes the data. This parsimony principle is widely used when considering problems in generalization. For many problems, though, the size of model is fixed beforehand and only the model parameters can be varied. The size can be determined either by the physics of the problem, as in [22, 24], or, as in speech recognition, by a grammar which describes combinations of words and phonemes represented by the model states. In these cases the appropriate solution to overtraining is to obtain more data because a smaller model cannot be used. The modified set of desirable distributions proposed here can be considered as a non-presumptive attempt to augment or extend the training set to a size appropriate for the model being trained.

The tendency of the EM algorithm to overfit training data is addressed in [25]. Overfitting is reduced by restricting the model set to those models whose parameters satisfy smoothness constraints. The smoothness criterion is considered prior knowledge, however, and unlike the approach proposed here, the constraints are applied to the set of models, not to the empirical distributions formed by the data set, i.e. the constraints are applied to $\mathcal{Q}$, not $\mathcal{D}$. In general these smoothness constraints make the EM algorithm harder to implement. More is required than simply smoothing the model produced by the M-step: the algorithm must be modified so that the M-step is a constrained optimization which searches for models among those with appropriate smoothness. Techniques which relax the requirement that smoothing and maximization be performed simultaneously have also been found to reduce overfitting [26, 27]. It will be shown that applying the confidence constraint to the desired distributions as suggested here adds little complexity to the learning algorithms.

3.2 The Small Data Set Assumptions

The derivations which follow make use of the *small data set* assumption

$$Q(\mathcal{T}) \approx 0 \quad \forall Q \in \mathcal{Q}. \tag{47}$$

This is clearly satisfied in most speech recognition examples where special techniques must be used to avoid numerical underflow computing the probabilities involved [1].

A slightly stronger assumption is also invoked

$$\sum_{I \in \mathcal{T}, S} Q(I,S)\, g(I,S) \approx 0 \quad \forall Q \in \mathcal{Q} \tag{48}$$

for some statistics g. Typically the g are indicator functions which specify that I and S jointly satisfy some property when I belongs to $\mathcal{T}$. When this is so

$$\sum_{I,S} Q(I,S)\, g(I,S) \leq \sum_{I,S} Q(I,S)\, 1_{\mathcal{T}}(I) \;= Q(\mathcal{T}) \tag{49}$$

so this assumption is no stronger than the first assumption.

4 Imposing the Confidence Constraint

The Maximum Likelihood training algorithms presented earlier can be modified to include this constraint on the desired distributions (see Equation 40):

$$\mathcal{D}_c^N \;\; = \;\; \mathcal{D}_c \cap \{P \in \mathcal{P} : |P^{-1}(\cdot|I)| \leq N \quad \forall I \in \mathcal{T}\}. \tag{50}$$

The initial step in each training algorithm requires finding the I-Projection P^p of the current model Q^p onto $\mathcal{D}_c^N$ by solving Equation 6. In Appendix A it is shown that I-Projection has the following form

$$P^p(I,S) = \begin{cases} c\ \hat{P}(I)\ \ Q^p(S|B_N^p(I), I) & S \in B_N^p(I),\ I \in \mathcal{T} \\ 0 & S \notin B_N^p(I),\ I \in \mathcal{T} \\ (1-c)\ \frac{Q^p(I,S)}{Q^p(\mathcal{T}^c)} & I \in \mathcal{T}^c \end{cases} . \tag{51}$$

The divergence from Q^p to $\mathcal{D}_c^N$ is

$$D(P^p \,||\, Q^p) \;\; = \;\; c \sum_{\mathcal{T}} \hat{P}(I) \log \frac{\hat{P}(I)}{Q^p(I, B_N^p(I))} + (1-c) \log \frac{1-c}{Q^p(\mathcal{T}^c)} + c \log c. \tag{52}$$

The improvement under the divergence stated in Equation 10 still holds, however the resulting improvement in likelihood scoring of the training set is slightly more complicated. Combining Equations 52 and 10 yields

$$c \sum_{\mathcal{T}} \hat{P}(I)\ \ L_N^p(I) + (1-c) \log Q^p(\mathcal{T}^c) \leq c \sum_{\mathcal{T}} \hat{P}(I)\ \ L_N^{p+1}(I) + (1-c) \log Q^{p+1}(\mathcal{T}^c). \tag{53}$$

From this it appears that the modified algorithm attempts to maximize the likelihood criterion, as in Equation 42, but is penalized if the model support becomes concentrated on the training data, i.e. if $Q^{p+1}(\mathcal{T}^c)$ decreases.

These two objectives may be incompatible. Consider the case of the Baum Welch algorithm for which $L_N^p(I) = \log Q^p(I)$ and suppose that

$$\log Q^{p+1}(I) > \log Q^p(I) \;\; \forall I \in \mathcal{T} \tag{54}$$

which is desirable because the marginal likelihood of each training item is increased. However, it implies that

$$\log Q^{p+1}(\mathcal{T}^c) < \log Q^p(\mathcal{T}^c) \tag{55}$$

and that Equation 53 may be violated.

4.1 Continuous Hidden and Observed Variables

The above presentation is complicated in the case when the model observation distributions are continuous. A derivation of the EM algorithm under the maximum likelihood criterion will be presented first. Then an application of the alternating minimization procedure will be presented for continuous models with a discounted likelihood criterion.

4.1.1 A Review of a Derivation of the EM Algorithm

The following derivation of the EM algorithm is presented in [28] and for HMMs in [29]. It is presented here for models defined on continous product spaces. Consider the models Q which are defined on the space $(\mathbf{S},\mathcal{S}) \times (\mathbf{I},\mathcal{I})$ in terms of densities f

$$Q(A,B) = \int_A \int_B f_{S,I} \quad A \in \mathcal{S}, \; B \in \mathcal{I}. \tag{56}$$

The training goal under the marginal likelihood criterion is to modify a model Q^p with density f^p to obtain a new model Q^{p+1} so that the likelihood of the training data T is improved according to

$$\prod_{I' \in T} f_I^{p+1}(I') \geq \prod_{i \in T} f_I^p(I') \tag{57}$$

or equivalently,

$$\sum_{I' \in T} \log\, f_I^{p+1}(I') \geq \sum_{I' \in T} \log\, f_I^p(I'). \tag{58}$$

Following [28], the relationship

$$\log\, f_I = \log\, f_{I,S} - \log\, f_{S|I} \tag{59}$$

and the properties of conditional expectation are used as follows:

$$\begin{aligned} \sum_{I' \in T} \log\, f_I^{p+1}(I') &= \sum_{I' \in T} E_{Q^p}[\log\, f_I^{p+1}(I)|I' = I] &(60)\\ &= \sum_{I' \in T} E_{Q^p}[\log\, f_I^{p+1}(I,S)|I' = I] - \sum_{I' \in T} E_{Q^p}[\log\, f_{S|I}^{p+1}(S)|I' = I]. &(61) \end{aligned}$$

By Jensen's Inequality, the elements of the second summation obey

$$E_{Q^p}[\log\, f_{S|I}^{p+1}(S)|I' = I] \leq E_{Q^p}[\log\, f_{S|I}^{p}(S)|I' = I] \tag{62}$$

so that

$$\sum_{I' \in T} \log\, f_I^{p+1}(I') \geq \sum_{I' \in T} E_{Q^p}[\log\, f_I^{p+1}(I,S)|I' = I] - \sum_{I' \in T} E_{Q^p}[\log\, f_{S|I}^{p}(S)|I' = I]. \tag{63}$$

Under the EM algorithm, Q^{p+1} is chosen so that

$$f^{p+1} = \underset{f}{\arg\max} \sum_{I' \in T} E_{Q^p}[\log f_{I,S}(I,S)|I' = I]. \tag{64}$$

To show that this produces an improved model under the likelihood criterion, note that this choice of f^{p+1} implies

$$\sum_{I' \in T} E_{Q^p}[\log f_{I,S}^{p+1}(I,S)|I' = I] \geq \sum_{I' \in T} E_{Q^p}[\log f_S^p(I,S)|I' = I]. \tag{65}$$

Applying this to Equation 63 yields

$$\begin{aligned} \sum_{I' \in T} \log f_I^{p+1}(I') &\geq \sum_{I' \in T} E_{Q^p}[\log f_{I,S}^p(I,S)|I' = I] - \sum_{I' \in T} E_{Q^p}[\log f_{S|I}^p(S)|I' = I] &(66)\\ &= \sum_{I' \in T} \log f_I^p(I') &(67) \end{aligned}$$

and improvement under the likelihood criterion is obtained.

4.1.2 Imposing the Confidence Constraint on Continuous Models

Formulating the EM algorithm for models with continuous observations within the alternating minimization framework is problematic. This is so because, as defined above, the desired distributions have the property that $P(\mathcal{T}) > 0$, while the models typically are smooth so that $Q(\mathcal{T}) = 0$. P is therefore not continuous with respect to Q so that $D(P||Q)$ is not finite. This is addressed in [6], Page 229, and the correct approach is described, which requires transforming the models so that the divergence is well defined. The observations are assumed to be real valued and$D$-dimensional, i.e. $\mathbf{S} = \Re^D$.

To present the desired results for the discounted data criterion, a different approach is taken here, which is technically incorrect but direct. For each training example $I_i \in \mathcal{T}$, define a set $A_i \in \Re^D$ so that $I_i \in A_i$ and $\lambda(A_i) = \epsilon$, where λ is Lebesque measure on $\Re^D$. The value of ϵ is chosen small enough so that the A_i are disjoint. The sets A_i define ϵ-neighborhoods around the training examples and are used to define a "smooth" empirical distribution on $\Re^D$

$$\hat{P}(A) = \sum_i a_i\, \lambda(A \cap A_i). \tag{68}$$

where $a_i = \frac{\#_T(I_i)}{\epsilon\,|T|}$. For each $I_i \in T$, $\hat{P}(A_i) = \frac{\#_T(I_i)}{|T|}$, which agrees with the empirical distribution for discrete observation models.

The discounted desirable distributions are defined as

$$\mathcal{D}_c = \{P : \int_A \int dP = c\,\hat{P}(A) \;\; \forall A \in \mathcal{I}\,\}. \tag{69}$$

This is equivalent to

$$\mathcal{D}_c = \{P : \int \int 1_{A_i} dP = c\,a_i \;\; i = 1, \ldots, |\mathcal{T}|\,\}. \tag{70}$$

The first step of the alternating minimization algorithm requires finding $\min_{P \in \mathcal{D}_c} D(P||Q^p)$ for a given model Q^p. The models Q are assumed to be smooth, with densities $f_{S,I}, f_I, f_{S|I}$.

The I-Projection of Q^p onto a set of distributions specified through the linear constraints of Equation 70 is well known and has the form

$$\frac{\partial P}{\partial Q} = b\, \exp \sum_i t_i\, 1_{A_i} \tag{71}$$

where t_i is chosen so that the linear constraints are satisified and b ensures that the I-projection is normalized. Satisfying the linear constraints yields

$$\begin{aligned} c\,a_i &= \int\int 1_{A_i}\, b\, e^{t_i}\, f_{S,I} && (72)\\ c\,a_i &= b\,e^{t_i} \int_{A_i} \int f_{S,I} && (73)\\ b\,e^{t_i} &= \frac{c\,a_i}{Q(A_i)}. && (74) \end{aligned}$$

Define $A = \cup A_i$, and the following yields the value of b

$$\begin{aligned} 1 &= \int\int \frac{\partial P}{\partial Q} f_{S,I} && (75)\\ 1 &= \sum_i \int_{A_i} \int \frac{c\,a_i}{Q(A_i)} f_{S,I} + \int_{A^c} \int b\, q_{S,I} && (76)\\ 1 &= c + b\, Q(A^c) && (77) \end{aligned}$$

so that $b = \frac{1-c}{Q(A^c)}$.

Combining these results yields the density of P^p, the I-Projection of Q^p onto $\mathcal{D}_c$

$$\frac{\partial P^p}{\partial Q^p} = c \sum_i \frac{a_i}{Q(A_i)} 1_{A_i} + (1-c) \frac{1-c}{Q^p(A^c)} 1_{A^c}. \tag{78}$$

The density of P^p with respect to Lebesque measure can be written as

$$h^p = c \sum_i \frac{a_i}{Q(A_i)} 1_{A_i} f^p_{S,I} + \frac{1-c}{Q^p(A^c)} 1_{A^c} f^p_{S,I}. \tag{79}$$

where f^p denotes the density of Q^p.

The next step requires finding Q^{p+1} as $\min_{Q \in \mathcal{Q}} D(P^p || Q)$. This requires solving

$$\begin{aligned}
\min_{Q \in \mathcal{Q}} D(P^p||Q) &= \min_f \int \int h^p \log \frac{h^p}{f} && (80)\\
&= \max_f \int \int h^p \log f && (81)\\
&= \max_f [\int \int \sum_i \frac{c\, a_i}{Q(A_i)} 1_{A_i} \log f + \int \int \frac{1-c}{Q^p(A^c)} 1_{A^c} \log f\,] && (82)\\
&= \max_f [\sum_i \frac{c\, a_i}{Q^p(A_i)} \int_{A_i} \int f^p \log f + \frac{1-c}{Q^p(A^c)} \int_{A^c} \int f^p \log f\,]. && (83)
\end{aligned}$$

For small enough ϵ and smooth densities f, the following approximation is valid

$$\begin{aligned}
\int \int_{A_i} f^p(I,S) \log f(I,S)\, dI\, dS &\approx \int f^p(I_i,S) \log f(I_i,S)\, dS \int_{A_i} dI && (84)\\
&= \epsilon \int f^p(I_i,S) \log f^p(I_i,S)\, dS. && (85)
\end{aligned}$$

Similarly, $Q^p(A_i) \approx \epsilon\, f^p(I_i)$ and $Q^p(A^c) \approx 1$. Substituting these approximations into Equation 83 yields

$$\begin{aligned}
\min_{Q \in \mathcal{Q}} D(P^p||Q) &\approx \max_f [\sum_i \frac{c\, a_i\, \epsilon}{\epsilon f^p(I_i)} \int f^p(I_i,S) \log f(I_i,S) + (1-c) \int \int f^p \log f] && (86)\\
&= \max_f [\, c \sum_i a_i \int f^p(S|I_i) \log f(I_i,S) + (1-c) \int \int f^p \log f\,] && (87)\\
&= \max_f [\, c \frac{1}{|T|} \sum_{I' \in T} E_{Q^p}[\, \log f(S,I) | I' = I] + (1-c)\, E_{Q^p} \log f(S,I)\,]. && (88)
\end{aligned}$$

By comparison to Equation 64, it appears that that the discounted likelihood criterion modifies the usual maximum likelihood reestimation by introducing an additive penalty based on the cross-entropy between the current model and the new model. Because the second term is maximized when f^{p+1} equals f^p, smaller values of c which emphasize the second term will prevent the reestimated model from differing too much from the current model, effectively slowing the EM algorithm.

The addition of this penalty term falls within the framework of [25], where the EM algorithm is applied to cases where the maximum likelihood criterion is augmented with an arbitrary, additive penalty term. In that work though, the penalty term is fixed, i.e. does not vary with each iteration, and describes general desired properties of the reestimated models. It will be shown that the penalty presented here which is due to the discounted data criterion introduces little additional computation to the EM algorithm.

Except where otherwise noted, discrete observation models will be the focus of the remainder of this work.

5 Model Reestimates Under the Modified Likelihood Criterion

The next step in the training procedure requires finding a new model Q^{p+1} from the desired distribution P^p by solving Equation 9. While the actual solution requires knowledge of the family $\mathcal{Q}$, general characteristics of the solution can be described. Consider the exponential form of the model distribution

$$
\begin{aligned}
Q(I,S) &= Q(I|S)\ Q(S) && (89)\\
&= \frac{1}{z_w}\exp\{\sum_{j,k} w_{j,k}\ \ g_{j,k}(I,S) + \sum_k w_k\ \ g_k(S)\} && (90)\\
&= \frac{1}{z_w}\exp\{w\cdot g(I,S)\} && (91)\\
z_w &= \sum_{I,S}\exp\{w\cdot g(I,S)\}. && (92)
\end{aligned}
$$

A distribution on countable spaces can be put in this form, for example, by choosing the possibly redundant representation $w_k = \log Q(S)$, $w_{j,k} = \log Q(I|S)$, where j and k index the sets $\mathcal{I}$ and $\mathcal{S}$, and the g are indicator functions. The parameters w are called the natural parameters of the distribution.

Associated with the natural parameters are the moments, or expectation parameters, of the function g

$$
\begin{aligned}
q_k &= \sum_S g_k(S)\ Q(S) && (93)\\
q_{j,k} &= \sum_{I,S} g_{j,k}(I,S)\ Q(I,S) && (94)
\end{aligned}
$$

The moments of Q^{p+1} and P^p are denoted as $q^{p+1} = E_{Q^{p+1}}g$ and $p^p = E_{P^p}g$, respectively.

The model Q^{p+1} which achieves $\min_Q D(P^p \parallel Q)$ can be shown to be the model whose moments are identical to those of the desired distribution P^p

$$Q^{p+1}:\ q^{p+1} = p^p. \qquad (95)$$

This is shown in Appendix B for HMMs.

The EM procedure can be summarized in terms of these moments: from the current model Q^p the desired moments p^p must first be found; the model Q^{p+1} is then found so that it has the same moments.

The computation of p^p proceeds by substituting Equation 51 into 94

$$
\begin{aligned}
p^p &= \sum_{I,S} g(I,S)\ P^p(I,S) && (96)\\
&= c\sum_{I\in\mathcal{T},S} g(I,S)\ Q^p(S|B_N^p(I),I)\ \hat{P}(I) + \frac{1-c}{Q^p(\mathcal{T}^c)}\sum_{I\in\mathcal{T}^c,S} g(I,S)\ Q^p(I,S)\\
&= c\sum_{I\in\mathcal{T}} \hat{P}(I)\ E_{Q^p}[g(I,S)|B_N^p(I),I] + \frac{1-c}{Q^p(\mathcal{T}^c)}[\sum_{I,S} g(I,S)\ Q^p(I,S) - \sum_{I\in\mathcal{T},S} g(I,S)\ Q^p(I,S)]\\
&= c\sum_{I\in\mathcal{T}} \hat{P}(I)\ E_{Q^p}[g(I,S)|B_N^p(I),I] + \frac{1-c}{Q^p(\mathcal{T}^c)}[q^p - \sum_{I\in\mathcal{T},S} g(I,S)\ Q^p(I,S)]. && (97)
\end{aligned}
$$

The next step assumes Equation 95 is satisfied at the previous iteration

$$p^p = c\sum_{I\in\mathcal{T}} \hat{P}(I) E_{Q^p}[g(I,S)|B_N^p(I),I] + \frac{1-c}{Q^p(\mathcal{T}^c)}[p^{p-1} - \sum_{I\in\mathcal{T},S} g(I,S)\ Q^p(I,S)]. \qquad (98)$$

In the usual, $c = 1$, form of the training algorithms the reestimate of p^p which would be found from Q^p is denoted $\tilde{p}^p$. Its value is found by setting $c = 1$ in Equation 98

$$\tilde{p}^p = \sum_{I \in \mathcal{T}} \hat{P}(I) E_{Q^p}[g(I,S)|B^p_N(I), I]. \tag{99}$$

To compare the discounted update rule to the usual rule, Equation 98 can be written as

$$p^p = c\,\tilde{p}^p + \frac{1-c}{1-Q^p(\mathcal{T})}\,[\,p^{p-1} - \sum_{I \in \mathcal{T},S} g(I,S)\;\; Q^p(I,S)\,]. \tag{100}$$

Under the small data set assumptions described earlier, these equations become

$$p^p = c\,\tilde{p}^p + (1-c)\,p^{p-1}. \tag{101}$$

The modification to the maximum likelihood criterion effectively slows the training algorithm by low-pass filtering the reestimates of the moments which determine the model parameters.

The addition of the confidence parameter adds little complexity to the EM algorithm. The M-step is unaffected. The E-step is performed as usual and the resulting moments are mixed with the moments produced by the previous E-step; the added complexity is in storing and mixing moments.

It is worth stressing that the *moment* estimates are being filtered, not the *parameter* estimates. This distinction is important because parameters and moments live in separate spaces and the simple filtering in the moment space described by Equation 101 can lead to more complicated updating behavior in the parameter space. The term "filtering" also stresses the difference between this technique and smoothing.

5.1 EM as Gradient Descent in Moment Space

Consider the marginal likelihood criterion

$$Q(I) = \sum_S Q(I,S) \tag{102}$$

and its associated, discounted desired distributions

$$\mathcal{D}_c = \{P : \sum_S P(I,S) = c\,\hat{P}(I)\}. \tag{103}$$

In Appendix D it is shown that

$$\frac{\delta}{\delta w} D(\mathcal{D}_c \,||\, Q) = -c\,(\tilde{p} - q) + (1-c)\,\frac{1}{Q(\mathcal{T}^c)}(\,Q(\mathcal{T})\,q - \sum_{\mathcal{T},S} g(I,S)\,Q(I,S)\,) \tag{104}$$

which under the small data set assumption becomes

$$\frac{\delta}{\delta w} D(\mathcal{D}_c \,||\, Q) = -c\,(\tilde{p} - q)\;. \tag{105}$$

Under exact updating, Equation 95, the moment low-pass filtering of Equation 101 can be written as

$$\begin{aligned} p^p &= c\,\tilde{p}^p + (1-c)\,p^{p-1} && (106)\\ &= p^{p-1} + c\,(\tilde{p}^p - p^{p-1}) && (107)\\ &= p^{p-1} - c\,\frac{\delta}{\delta w} D(\mathcal{D}_c \,||\, Q)|_{Q^p}. && (108) \end{aligned}$$

Define the moment covariance matrix Σ as

$$\Sigma_{i,j} = \frac{\delta p_i}{\delta w_j}. \tag{109}$$

That this is the moment covariance matrix can be derived as follows

$$\begin{aligned}
\frac{\delta p_i}{\delta w_j} &= \frac{\delta}{\delta w_j} \sum_{I,S} g_i \frac{1}{z_w} \exp\{w \cdot g\} && (110)\\
&= \sum_{I,S} [g_i\, g_j \frac{1}{z_w} \exp\{w \cdot g\} - g_i \exp\{w \cdot g\}\, \frac{1}{z_w^2} \frac{\delta}{\delta w_j} z_w] && (111)\\
&= E_Q\, g_i\, g_j - [\frac{1}{z_w} \frac{\delta}{\delta w_j} z_w] \sum_{I,S} g_i\, \frac{1}{z_w} \exp\{w \cdot g\} && (112)\\
&= E_Q\, g_i\, g_j - [\frac{\delta}{\delta w_j} \log z_w]\, p_i && (113)\\
&= E_Q\, g_i\, g_j - p_i\, p_j. && (114)
\end{aligned}$$

The last step uses the relationship

$$\begin{aligned}
\frac{\delta}{\delta w_j} \log z_w &= \frac{1}{z_w} \sum_{I,S} \frac{\delta}{\delta w_j} \exp\{g \cdot w\} && (115)\\
&= \sum_{I,S} g_j \exp\{g \cdot w\} && (116)\\
&= p_j. && (117)
\end{aligned}$$

Σ is the Fisher Information matrix. It also forms a metric tensor which relates the expectation coordinate system and the natural parameter coordinate system by [30, 31]

$$\frac{\delta}{\delta w_j} = \sum_i \frac{\delta p_i}{\delta w_j} \frac{\delta}{\delta p_i}. \tag{118}$$

Applying this relationship to Equation 104 yields

$$\nabla_w D(\mathcal{D}_c \parallel Q) = \Sigma\, \nabla_p D(\mathcal{D}_c \parallel Q). \tag{119}$$

Using Σ^p to denote the moment covariances determined under Q^p, the EM algorithm can be described as a gradient descent in the expectation coordinate system with a step size of 1 in a direction defined by the moment covariance matrix Σ^p

$$\tilde{p}^p \;=\; p^{p-1} \;-\; \Sigma^p\, \nabla_p D(\mathcal{D} \parallel Q)|_{Q^p}\,. \tag{120}$$

Under the small data set assumption, the discounted, $c < 1$, algorithm can be described as an update in the same direction, but with a reduced step size

$$p^p \;=\; p^{p-1} \;-\; c\,\Sigma^p\, \nabla_p D(\mathcal{D}_c \parallel Q)|_{Q^p}\,. \tag{121}$$

5.2 Reestimates of Hidden Markov Models

The model family Q considered here first will be that of left-to-right, variable duration Hidden Markov Models (VDHMM) [32].

$$I = (I_1, \ldots, I_L) \quad I_t \in \{1, \ldots, M\} \tag{122}$$

$$S = (S_1, \ldots, S_L) \quad S_t \in \{1, \ldots, N\} \quad S_t \leq S_{t+1}. \tag{123}$$

The variables S_t describe the state occupancy of the system at time t and I_t is an observation generated while the system is in that state. The model distribution $Q(I, S) = Q(I|S)\, Q(S)$ is determined through component distributions b_n and d_n, where the b_n are state-dependent observation distributions and the d_n are state-duration densities

$$Q(I|S) \;=\; \prod_{t=1}^{L} b_{S_t}(I_t) \tag{124}$$

$$Q(S) \;=\; \prod_{n=1}^{N} d_n(\tau_n), \quad \tau_n = \sum_{t=1}^{L} \delta_n(S_t). \tag{125}$$

The variable τ_n describes the duration of state n in sequence S.

In Appendix B, it is shown that Q^{p+1} is found from Q^p according to

$$d_n^{p+1}(\tau) \;=\; c\; \tilde{d}_n^{p+1}(\tau) + \frac{1-c}{1-Q^p(\mathcal{T})}\; [d_n^p(\tau) - Q^p(\mathcal{T}, \tau_n = \tau)] \tag{126}$$

$$b_n^{p+1}(i) \;=\; c\, \tilde{b}_n^{p+1}(i) + \frac{1-c}{1-Q^p(\mathcal{T})}\; [b_n^p(i) - \frac{1}{\bar{\tau}_n^p} \sum_{I\in\mathcal{T}} \sum_t Q^p(I, S : I_t = i, S_t = n)] \tag{127}$$

where $\tilde{b}$ and $\tilde{d}$ are the reestimates which would have been found from Q^p by the usual, $c = 1$, training algorithm. These are found using the *scaled* version of the forward-backward algorithm [3]; the correction quantities inside the brackets can be found using the algorithm in its unscaled form.

Under the small data set assumptions the reestimates of the component model distributions become

$$d^{p+1} \;=\; c\, \tilde{d}^{p+1} + (1-c)\, d^p \tag{128}$$

$$b^{p+1} \;=\; c\, \tilde{b}^{p+1} + (1-c)\, b^p. \tag{129}$$

The effect is to retard overtraining by slowing the rate at which any element of the component distributions approaches zero.

In this example, the HMM component distributions are filtered directly. This is not inconsistent with the earlier argument which led to filtering of the moments of the distribution. Consider the exponential form of $P(S)$

$$P(S) \;=\; \prod_{n=1}^{N} d_n(\tau_n) = \exp\{\sum_{n=1}^{N} \sum_{\tau} \log d_n(\tau)\;\; \delta_\tau(\tau_n)\} \tag{130}$$

$$\;=\; \exp\{\sum_n \sum_\tau w_{n,\tau} g_{n,\tau}(S)\} \tag{131}$$

where $w_{n,\tau} = \log d_n(\tau)$ and $g_{n,\tau}(s) = \delta_\tau(\tau_n)$. The moments of the distribution, $\bar{g}_{n,\tau} = E g_{n,\tau}(S)$, satisfy $\bar{g}_{n,\tau} = d_n(\tau)$. So when the component distributions of the model are unconstrained, filtering the model moments is equivalent to filtering the component distributions. However, this need not be so when the component distributions of the model are constrained to particular parameterizations.

5.2.1 Homogeneous State Transition Probabilities

Often the hidden process is defined by homogeneous state transition probabilities a so that

$$Pr(S_{t+1} = n'|S_t) = a_{n,n'} \quad S_t = n \tag{132}$$

where arbitrary state transitions are allowed. The distribution of the hidden sequence is then

$$Q(S) = \prod_{n,n'=1}^{N} a_{n,n'}^{\#_{n,n'}(S)} \tag{133}$$

where $\#_{n,n'}(S)$ is the number of transitions from state n to state n' in S.

The observation distributions b^{p+1} which define $Q^{p+1}(I|S)$ are found as above in Equation 126. However $Q^{p+1}(S)$ must be reestimated in terms of a^{p+1}. In Appendix B.1 it is shown that

$$a^{p+1}_{n,n'} = \frac{\#^{p+1}_{n,n'}}{\#^{p+1}_{n}} \tag{134}$$

where

$$\begin{aligned} \#^{p+1}_{n,n'} &= \sum_{S} P^p(S)\,\#_{n,n'}(S) && (135)\\ \#^{p+1}_{n} &= \sum_{n'} \#^{p+1}_{n,n'}. && (136) \end{aligned}$$

The moments $\#^{p+1}$ are found in the same manner as those in Equation 100 and under the small data set assumption they become (see Appendix B.1)

$$\#^{p+1}_{n,n'} = c\,\tilde{\#}^{p+1}_{n,n'} + (1-c)\,\#^{p}_{n,n'} \tag{137}$$

where $\tilde{\#}^{p+1}_{n,n'}$ is the moment reestimate under the usual, $c = 1.0$, algorithm

$$\tilde{\#}^{p+1}_{n,n'} = \sum_{I\in\mathcal{T}} \hat{P}(I) E_{Q^p}[\#_{n,n'}(S)|B^p_N(I), I]. \tag{138}$$

The parameter reestimates are therefore

$$a^{p+1}_{n,n'} = \frac{c\,\tilde{\#}^{p+1}_{n,n'} + (1-c)\,\#^{p}_{n,n'}}{c\,\tilde{\#}^{p+1}_{n} + (1-c)\,\#^{p}_{n}}. \tag{139}$$

The simple linear filtering of the moments in Equation 137 leads to a more complicated updating of the distribution parameters. However the increase in complexity is negligible: computing the filtered moments is simple and the task of computing parameters from the moments is unchanged.

To see the effect of this filtering, note that the usual $c = 1.0$ parameter update is

$$\tilde{a}^{p+1}_{n,n'} = \frac{\tilde{\#}^{p+1}_{n,n'}}{\tilde{\#}^{p+1}_{n}} \tag{140}$$

and that from Equation 134

$$a^{p}_{n,n'} = \frac{\#^{p}_{n,n'}}{\#^{p}_{n}} \tag{141}$$

so that

$$a^{p+1}_{n,n'} = \frac{c\,\tilde{a}^{p+1}_{n,n'} + (1-c)\,(\#^p_n/\tilde{\#}^{p+1}_n)\,a^p_{n,n'}}{c + (1-c)\,(\#^p_n/\tilde{\#}^{p+1}_n)}. \tag{142}$$

Suppose that $\tilde{\#}^{p+1}_n \gg \#^p_n$. This implies that

$$a^{p+1}_{n,n'} \approx \tilde{a}^{p+1}_{n,n'} \tag{143}$$

and that no correction is applied to the parameter reestimate. Alternatively, if $\tilde{\#}^{p+1}_n \ll \#^p_n$ then

$$a^{p+1}_{n,n'} \approx a^{p}_{n,n'} \tag{144}$$

and the algorithm is effectively prevented from modifying the parameter.

This shows that the modified algorithm favors relatively large values of the moment reestimate $\tilde{\#}^{p+1}_n$ and that if it produces such values the algorithm is allowed to procede uncorrected. Since this term is the expected state duration given the observations, the algorithm is encouraged to produce models in which all states are expected to appear frequently. Conversely, the algorithm is discouraged from producing models in which states may be expected to appear infrequently and which rely on a few states to model the observations.

5.2.2 Continuous Observation Densities

To generalize this techniqe to HMMs with continuous observation densities, the results from the previous discussion of continuous variables will be applied. The derivation will follow [29] (see also [33, 34]), although only the single mixture case will be presented for models with homogeneous transition probabilities and training under the marginal likelihood criterion. Applications to observation densities formed by mixtures of Gaussians and other likelihood criterion follow the procedure presented here.

Suppose that the input observation is taken from $\Re^D$, and the state dependent observation densities are Gaussians with the form

$$b_n(I) = \frac{1}{(2\pi)^{D/2}|\Sigma_n|^{\frac{1}{2}}} \exp[-\frac{1}{2}(I-\mu_n)'\Sigma_n^{-1}(I-\mu_n)] \tag{145}$$

with mean μ_n, and covariance Σ_n. The joint likelihood of an observed sequence I and a hidden sequence S is

$$f(S,I) = \prod_{t=1}^{T} b_{S_t}(I_t) \prod_{n',n=1}^{N} a_{n,n'}^{\#_{n,n'}(S)} \tag{146}$$

where f is parameterized by $\{a, \mu, \Sigma\}$.

From Equation 88, the parameters $\{a^{p+1}, \mu^{p+1}, \Sigma^{p+1}\}$ are found from the model Q^p parameters by solving

$$\max_f \left[\frac{c}{|T|} \sum_{I' \in T} E_{Q^p}[\log f(S,I)|I' = I] + (1-c)E_{Q^p} \log q(S,I)\right]. \tag{147}$$

It is shown in Appendix E that the parameter updates take the following form

$$\mu_n^{p+1} = \frac{c\,\tilde{\mu}_n^{p+1} + (1-c)\,(\#_n^p/\tilde{\#}_n^{p+1})\,\mu_n^p}{c + (1-c)\,(\#_n^p/\tilde{\#}_n^{p+1})} \tag{148}$$

$$\begin{aligned}\Sigma_n^{p+1} = {} & \frac{c\,\tilde{\Sigma}_n^{p+1} + (1-c)\,(\#_n^p/\tilde{\#}_n^{p+1})\Sigma_n^p}{c + (1-c)\,(\#_n^p/\tilde{\#}_n^{p+1})} + \\ & \frac{c\,(2\,\tilde{\#}_n^{p+1} + 1)}{c + (1-c)\,(\#_n^p/\tilde{\#}_n^{p+1})}\,(\tilde{\mu}_n^{p+1} - \mu_n^{p+1})'(\tilde{\mu}_n^{p+1} - \mu_n^{p+1}) + \\ & \frac{1-c}{c\,(\tilde{\#}_n^{p+1}/\#_n^p) + (1-c)}\,(\mu_n^p - \mu_n^{p+1})'(\mu_n^p - \mu_n^{p+1})\end{aligned} \tag{149}$$

As in the reestimation of the state transition parameters, the relative values of the current expected state durations and the reestimated durations control the progression of the algorithm.

6 Relationship to Other Algorithms

6.1 Moment Decay and Weight Decay

This technique bears some resemblance to estimates of model parameters in which the training criterion is augmented by a function which constrains the model parameters. An example of this is in training artificial neural networks from labeled data sets. A feed-forward neural network classifier can be described as a function f_w parameterized by a weight vector w. A labeled data set is a set of observations $T = \{(I, O)\}$ where I is a feature or data vector and O is the correct, or target network output value when I is the network input. A typical training goal is to find network weights which minimize the empirical error

$$C_w(T) = \sum_{(I,O)\in T} \| f_w(I) - O \|_2 . \tag{150}$$

To encourage generalization from the training data, a bias term can be added to the cost [35]

$$C'_w = C_w(T) + B_w. \tag{151}$$

This additional term is non-empirical, in that it is not explicitly a function of the training data. If B_w is chosen to be quadratic, the gradient descent parameter update rule with step size α is [35, ?]

$$\begin{aligned} w^{p+1} &= w^p - \alpha \nabla_w C'_w|_{w^p} \qquad (152)\\ &= (1-2\alpha)w^p - \alpha \nabla_w C_w|_{w^p}. \qquad (153) \end{aligned}$$

For small α, this is called *weight decay*, because irrelevant weights, i.e. weights for which $\nabla_w C_w|_{w^p} = 0$, decay to zero. This is a "shrinkage" technique, and its relationship to ridge regression is discussed in [36]; this penalty can also be incorporated in the EM algorithm [25].

The parameter update can be rewritten as

$$\begin{aligned} w^{p+1} &= \alpha\,(w^{p+1} - \nabla_w C_w(T)|_{w^p}) + (1-\alpha)\,w^p \qquad (154)\\ &= \alpha\,\tilde{w}^{p+1} + (1-\alpha)\,w^p \qquad (155) \end{aligned}$$

which resembles the low-pass version of the algorithm in Equation 101. This is a gradient algorithm, so the mixture above arises from the gradient step size rather than from a balance of constraints.

When the models have an exponential form as in Equation 89, the moments and natural parameters are *dual* [31]. So within the framework of exponential models, and allowing for the differences between the exact updating of the Alternating Minimization algorithm and the incremental nature of gradient search, the technique proposed here is the dual of weight decay and might be termed *moment decay*.

6.2 Complete Data

The modification to the maximum likelihood criterion above is introduced in HMM training, however it can also be used when there are no hidden variables and the estimation is based on completely observable data. Suppose a variable takes values in the set $\mathcal{X} = \{1, \ldots, M\}$ and that m values are not observed in the training set, i.e. $|\{x \in \mathcal{X} : \#_T(x) = 0\}| = m$. The set of desired distributions is

$$\mathcal{D}_c = \{P : P(x) = c\,\frac{\#_T(x)}{|T|} \quad x \in \mathcal{T}\}. \tag{156}$$

If the models are unconstrained, then Q can be chosen so that $\min_Q D(\mathcal{D}_c \parallel Q) = 0$. For $c = 1$, $\mathcal{D}_c$ contains one member, the maximum likelihood solution

$$Q^{ML}(x) = \frac{\#_T(x)}{|T|}. \tag{157}$$

For $c < 1$, choosing the maximum entropy solution of $\min_{P \in \mathcal{D}_c} \min_Q D(P \parallel Q)$ yields

$$Q^c(x) = \begin{cases} c\,\frac{\#_T(x)}{|T|} & x \in \mathcal{T} \\ (1-c)\,\frac{1}{m} & x \in \mathcal{T}^c \end{cases}. \tag{158}$$

The value of c is arbitrary, however picking $c = \frac{|T|}{|T|+m}$ leads to

$$Q^c(x) = \begin{cases} \frac{\#_T(x)}{|T|+m} & x \in \mathcal{T} \\ \frac{1}{|T|+m} & x \in \mathcal{T}^c \end{cases} \tag{159}$$

which is equivalent to adding 1 to the bins left empty by the training data. This is a variation of the Add-One technique which, although widely used, has undesirable properties as an estimator [37]. It also shows that, in this formulation, c should approach 1 as the number of unobserved symbols (m) approaches 0. Cross-validation of density estimates based on divergence criteria are discussed in [38].

6.3 Prior Information

A Bayesian approach to including prior information in HMM training algorithms is presented in [39] as a way to avoid overtraining. Suppose the prior information consists of knowledge about the moments. A set of moments p^0 is known or suspected independently of the training data. These moments determine a model Q^0 such that $E_{Q^0} g(I,S) = q^0$. It is suggested that this knowledge be included in the moment reestimation as ([39],Equation 7)

$$p^p = \tilde{p}^p + p^0. \tag{160}$$

Because the p^0 bound the reestimates from zero, this is termed a regularization technique. The Add-One technique as applied in HMM modeling can also be described in this way.

This regularization can be derived within an Alternating Minimization framework, if the prior information leads to a model Q^0 which obeys the small data assumptions. By this assertion,

$$\sum_{I,S} Q^0(I,S)\, 1_{\mathcal{T}^c}(I)\, g(I,S) \approx q^0. \tag{161}$$

Define a set of distributions whose moments are constrained on the *complement* of the training set:

$$\mathcal{M}_c = \{P : \sum_{I,S} P(I,S)\, 1_{\mathcal{T}^c}(I)\, g(I,S) = (1-c)\, q^0\}. \tag{162}$$

The set of desired distributions which leads to Equation 160 is

$$\mathcal{D}_c^N \cap \mathcal{M}_c. \tag{163}$$

The moment constraints are enforced outside the training set to avoid incompatible constraints.

If the confidence value c is 0, $p^1 = q^0$ and $Q^2 = Q^0$, so that training algorithm ignores the training data and produces the model consistent with the prior information. If $c = 1$, the prior information is ignored and the algorithm proceeds with complete confidence in the training data.

It is shown in Appendix C that the I-Projection of Q^p onto $\mathcal{D}_c^N \cap \mathcal{M}_c$ is

$$P^p(I,S) = \begin{cases} c\, \hat{P}(I)\, Q^p(S|B_N^p(I), I) & I \in \mathcal{T} \\ (1-c)\, Q^0(I,S) & I \in \mathcal{T}^c \end{cases}. \tag{164}$$

This is found by projecting onto $\mathcal{D}_c^N \cap \mathcal{M}_c$ the projection of Q^p onto $\mathcal{D}_c^N$.

Under the small data set assumption, the resulting moment update equation is

$$p^p = c\, \tilde{p}^p + (1-c)\, p^0. \tag{165}$$

This derivation of the Bayesian regularization technique leads to an interesting interpretation of moment decay. The set of desirable distributions $\mathcal{D}_c^N$ does not provide enough information to carry out the model reestimation. Including the prior information $\mathcal{M}_c$ provides enough information to perform the reestimation, but if it is not supplied, it is computed from the current model. Moment decay is equivalent to guessing prior information based on the current model. In a sense, the results of each reestimation, as well as the initial model, are treated as prior information at the next reestimation. The value of c can be used to control how quickly the prior information is discarded.

It is possible to include incomplete prior knowledge, i.e. some of the $[q_j^0, q_{j,k}^0]$ may be unspecified so that Q^0 is not uniquely determined. The moment estimates for which priors are given are updated using Equation 165 while those for which no prior values are given are updated using Equation 101.

6.4 Sequential Algorithms

In the Alternating Minimization procedure the model parameters are reestimated so that Equation 95 holds exactly

$$w^{p+1} : q^{p+1} = p^p . \tag{166}$$

In Maximum Likelihood HMM training, the Baum-Welch algorithm yields in one step a model which satisfies this relationship. In training other models, such as the Boltzmann Machine [40], no such one-step, exact algorithm is available. Typically sequential, gradient descent algorithms are used to find the model with the correct moments. In some cases it is possible to implement the gradient search so that Equation 95 holds, for example Boltzmann Machine training can be formulated using Iterative Proportional Fitting to meet this constraint [41]. In other cases, though, the model parameters are modified in the direction of decreasing $D(\mathcal{D}_c \parallel Q)$, and the minimization is not solved exactly. This leads to a stochastic approximation to the Alternating Minimization algorithm (e.g. [42], Eq. 4) which uses sequential rather than exact updating.

The model parameters are updated as

$$w^{p+1} = w^p - \alpha \, \nabla_w D(\mathcal{D}_c \parallel Q)|_{Q^p} \tag{167}$$

for a small step size α. In Appendix D it is shown for the marginal likelihood scoring that

$$\nabla_w D(\mathcal{D}_c \parallel Q)|_{Q^p} = -c\,(\tilde{p}^p - q^p) + (1-c)\,\frac{1}{Q^p(\mathcal{T}^c)}(\, Q^p(\mathcal{T})\, q^p - \sum_{\mathcal{T},\mathcal{S}} g(I,S)\, Q^p(I,S)\,) \tag{168}$$

which under the Small Data Set Assumptions becomes

$$\begin{aligned} \nabla_w D(\mathcal{D}_c \parallel Q)|_{Q^p} &= -c\,(\tilde{p}^p - q^p) & (169) \\ &= c\, \nabla_w D(\mathcal{D}_{c=1} \parallel Q)|_{Q^p} & (170) \end{aligned}$$

The parameter update is then

$$w^{p+1} = w^p - c\,\alpha\, \nabla_w D(\mathcal{D}_{c=1} \parallel Q)|_{Q^p} \tag{171}$$

The confidence parameter leads directly to slower learning in these sequential algorithms by reducing the step size.

6.5 Clamping in Boltzmann Machine Learning

Boltzmann Machines [40] are artificial neural networks of binary valued, stochastic units which can be made to learn in an approximation of the EM algorithm. Learning proceeds in two phases. The network visible units are *clamped* according to an environmental distribution $\hat{P}$ and the hidden units operate freely. In the free-running phase, all units are allowed to operate freely. Statistics are accumulated while the network is operating in each mode, and the network weights are modified so that the free-running behavior of the network better matches the clamped behavior.

The steady-state distribution of the clamped network corresponds to the I-Projection of the current model onto $\mathcal{D}$ [41]

$$\bar{Q}(I,S) = Q(S|I)\, \hat{P}(I). \tag{172}$$

Under the discounted data criterion and the small data set assumption, the steady-state distribution according to the I-Projection onto $\mathcal{D}_c$ is

$$\bar{Q}(I,S) = Q(S|I)\, [c\, \hat{P}(I) + (1-c)\, Q(I)]. \tag{173}$$

The discounted data criterion effectively leads to a modified environmental distribution. While the statistics are accumulated, the network is clamped as usual to $\hat{P}$ for a portion c of the accumulation. For the remaining portion of the accumulation, data is collected with the network free-running. Moment decay can be implemented in the Boltzmann Machine distributed computational architecture with only a minor change in the usual learning rule. It is not necessary to store the moments from the previous iteration because they are found again during the free-running portion of the moment accumulation.

7 Examples of HMM Training and Generalization

7.1 VDHMM

The first examples of training and validation are presented for a simple VDHMM and a small training set. The training and validation sets each consist of 10 separate instances of a spoken digit. Acoustic features are obtained from a cochlear model and vector quantized using a codebook of size 32 so that each word is represented by a sequence of codeword indices [43]. A single model with $N = 5$ hidden states is trained for different values of c and its performance on the validation set is evaluated at each iteration. This case is somewhat artificial in that the model is much more powerful than required by the training set. However this is exactly the situation in which models do not generalize well.

7.1.1 Baum Welch Algorithm

Figure 1 shows the performance of the Baum Welch algorithm for $c = 1$ and $c = 0.5$. Each instance of the algorithm starts with the same initial model in which the duration and observation distributions are uniform. The trained models show improved performance at each iteration as measured on the training set and specified by Equation 10. However, the generalization ability of the usual ($c = 1$) algorithm (shown as -o-) is fairly poor. In training each word, the performance on the validation set fails to improve at the third iteration. The generalization shown when $c = 0.5$ is improved. Performance on the validation set continues to improve until at least the seventh iteration and achieves a higher score overall. For both values of c, the final training scores are nearly identical.

Some explanation of this improvement in generalization is given by Figure 2 which shows the sequence of reestimates of the duration distribution d_1, obtained in training the word "nine". When $c = 1.0$ the algorithm quickly, by iteration 3, concentrates the distribution about isolated durations. By contrast when $c = 0.5$, terms in the distribution decay to zero much more slowly. At the third iteration, it also concentrates on individual durations, however, the distribution is not so depressed around these values. Although it obscures the details of the distributions, this is most visible in the logarithmic plot shown for iteration 4 where it is clear that the distribution found when $c = 0.5$ supports a much broader range of durations than when $c = 1.0$.

In both training examples shown here, the likelihood score of the training set improves at each iteration. This occurs despite the modified likelihood gain of Equation 53, which suggests that some exchange between the likelihood score and the likelihood of T^c might occur.

7.1.2 Out-of-Class Rejection

Although it is not ensured by the maximum likelihood criterion, it is additionally desirable that a model trained to maximize the likelihood of data from one class also yield a low score when used to evaluate data from other classes. If this is so, a classifier based on a maximum likelihood rule will work reliably.

The data presented in Figure 1 is presented again in Figure 3. Also presented (plotted as -*-) are the results of testing the model trained on utterances of "nine" using the validation set of "zero", and the results of testing the model trained on utterances of "zero" using the validation set of "nine". The rejection of out-of-class data is better when $c = 1.0$ than when $c = 0.5$, although it is sufficient for correct recognition when $c = 0.5$.

7.2 Triphone Modeling with Gaussian Observation Densities

The acoustic properties of a phoneme depend upon the context in which it occurs. One method of modeling this variability is to create context-dependent models so that different phoneme models can be used in different contexts. In *triphone* modeling, the context information is the identity of the preceeding and following phonemes.

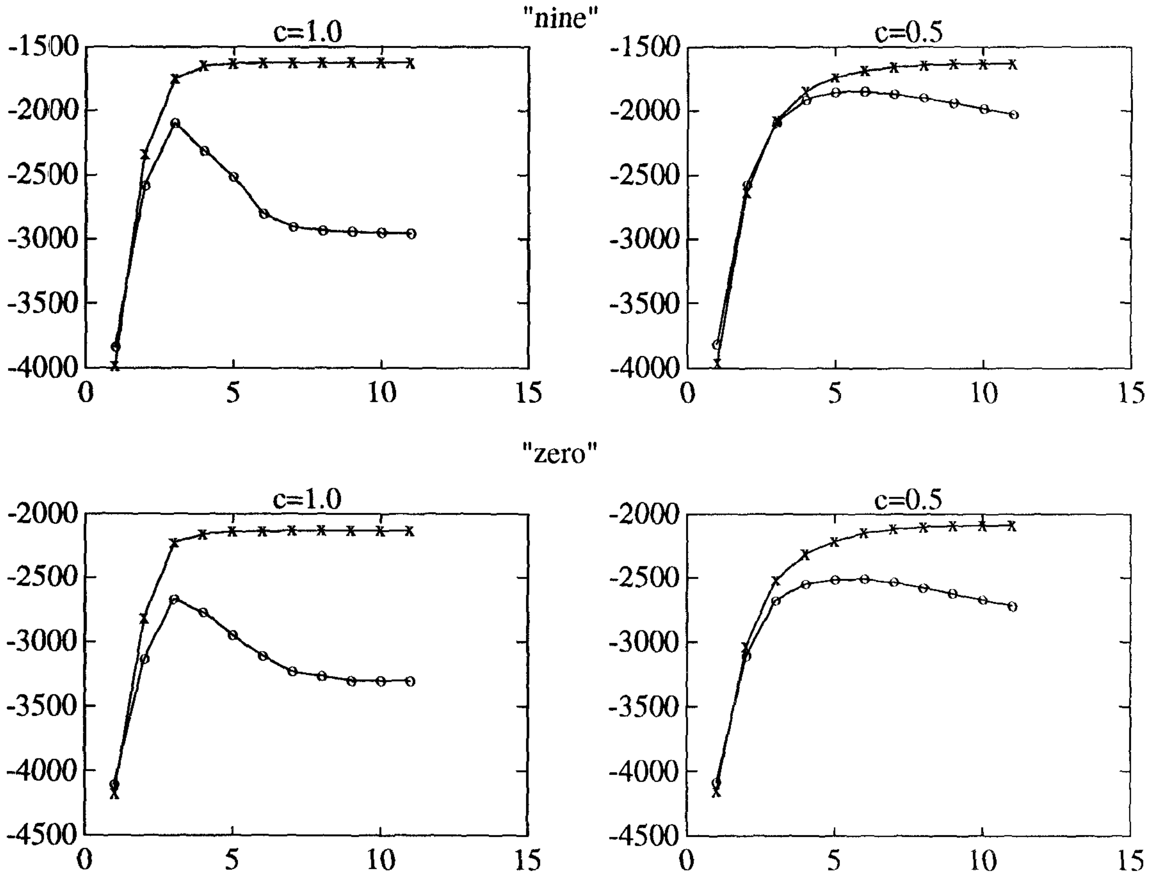

Figure 1: Baum Welch Algorithm: Log-likelihood Scores on Training and Validation Sets vs. Training Iteration for two values of c in two training tasks. Curves -x- plot the score $\sum_{I \in \mathcal{T}} \log Q^p(I)$ on the training set; curves -o- plot the score $\sum_{I \in \mathcal{V}} \log Q^p(I)$ on the validation set. The utterances modeled are (top) "nine" and (bottom) "zero". The test and validation sets each contain 10 utterances.

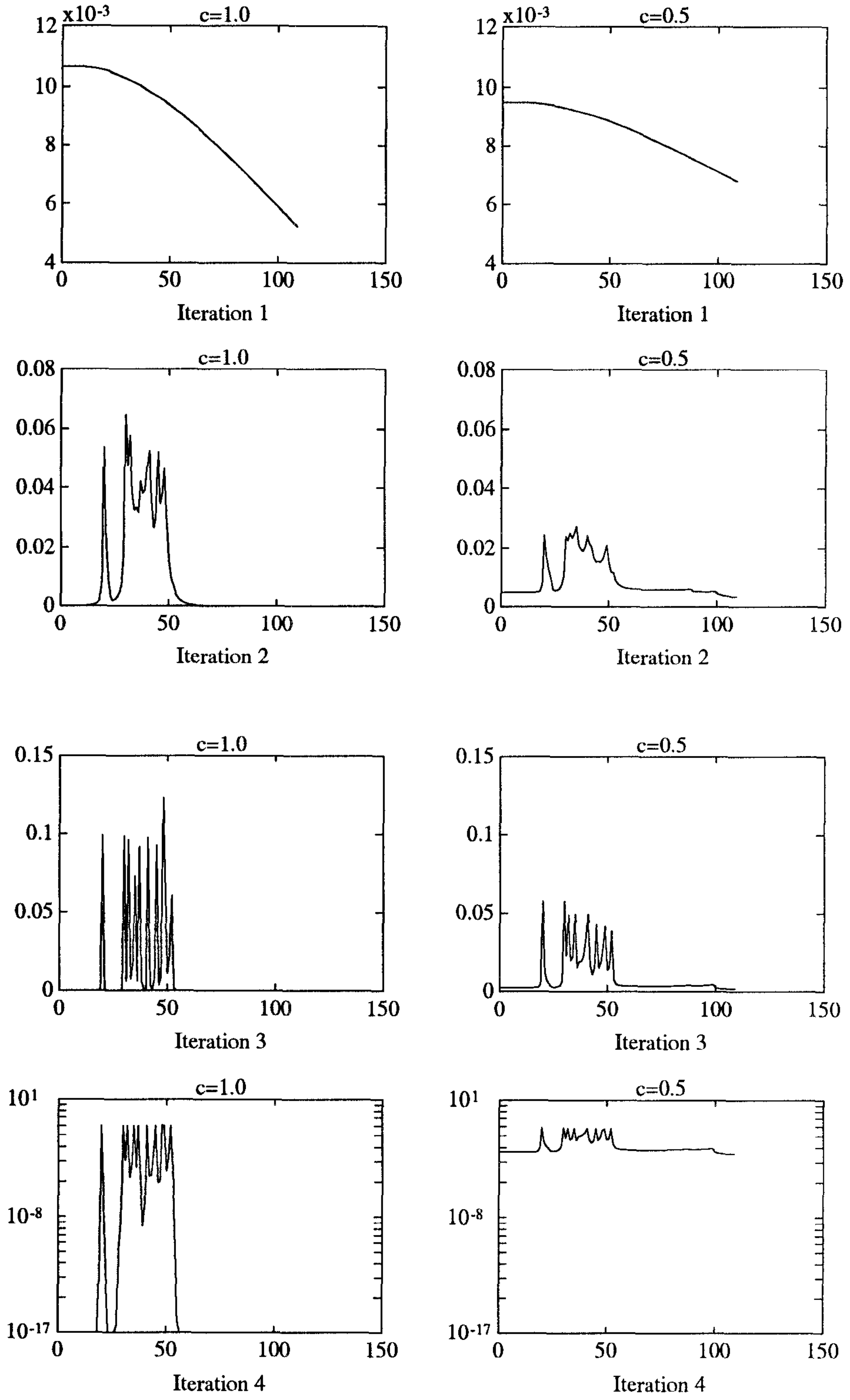

Figure 2: Duration Distribution d_1: Likelihood vs. duration by training iteration for two values of c in the Baum Welch algorithm; obtained in modeling "nine".

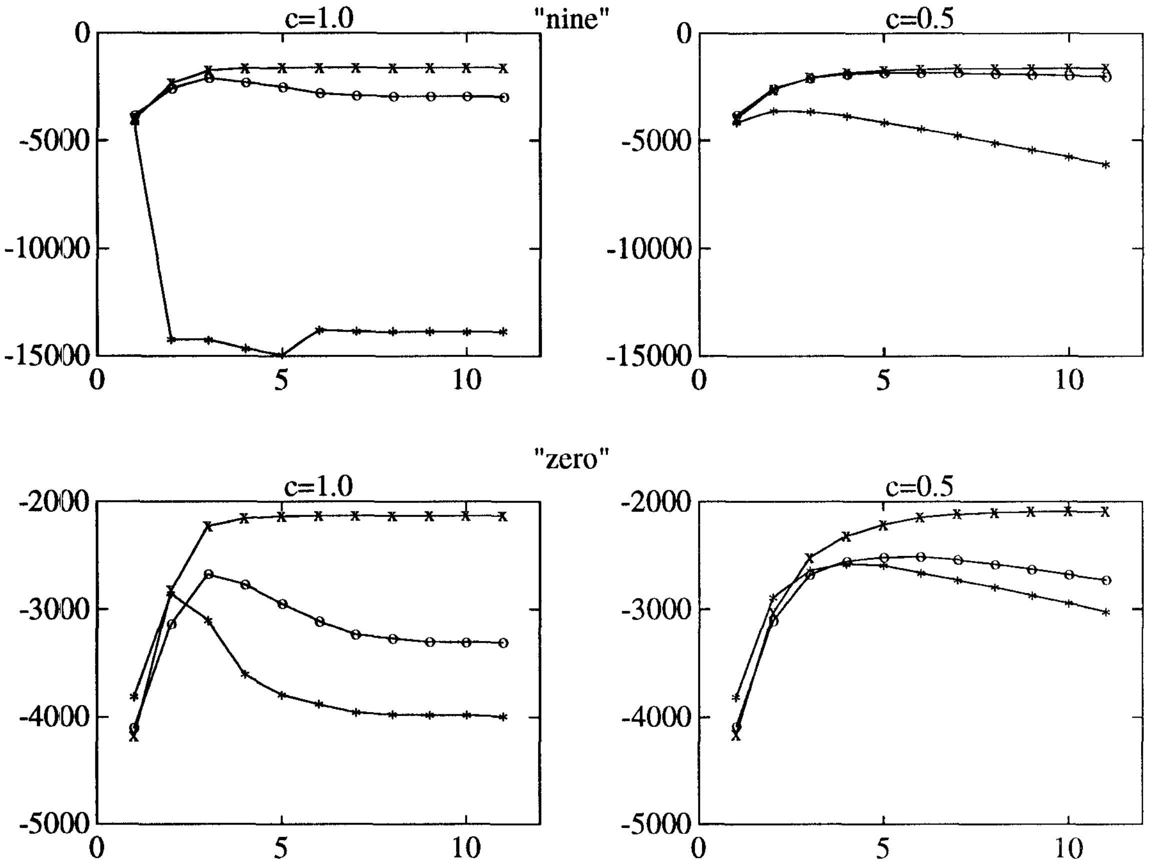

Figure 3: Baum Welch Algorithm: Log-likelihood Scores on Training, In-Class and Cross-Class Validation Sets vs. Training Iteration for two values of c. Curve $-x-$ plots values of $\sum_{I \in \mathcal{T}} \log Q^p(I)$; curve $-o-$ plots values of $\sum_{I \in \mathcal{V}} \log Q^p(I)$; curve -*- plots values of $\sum_{I \in \bar{\mathcal{V}}} \log Q^p(I)$, where $\bar{\mathcal{V}}$ is the validation set of the other model.

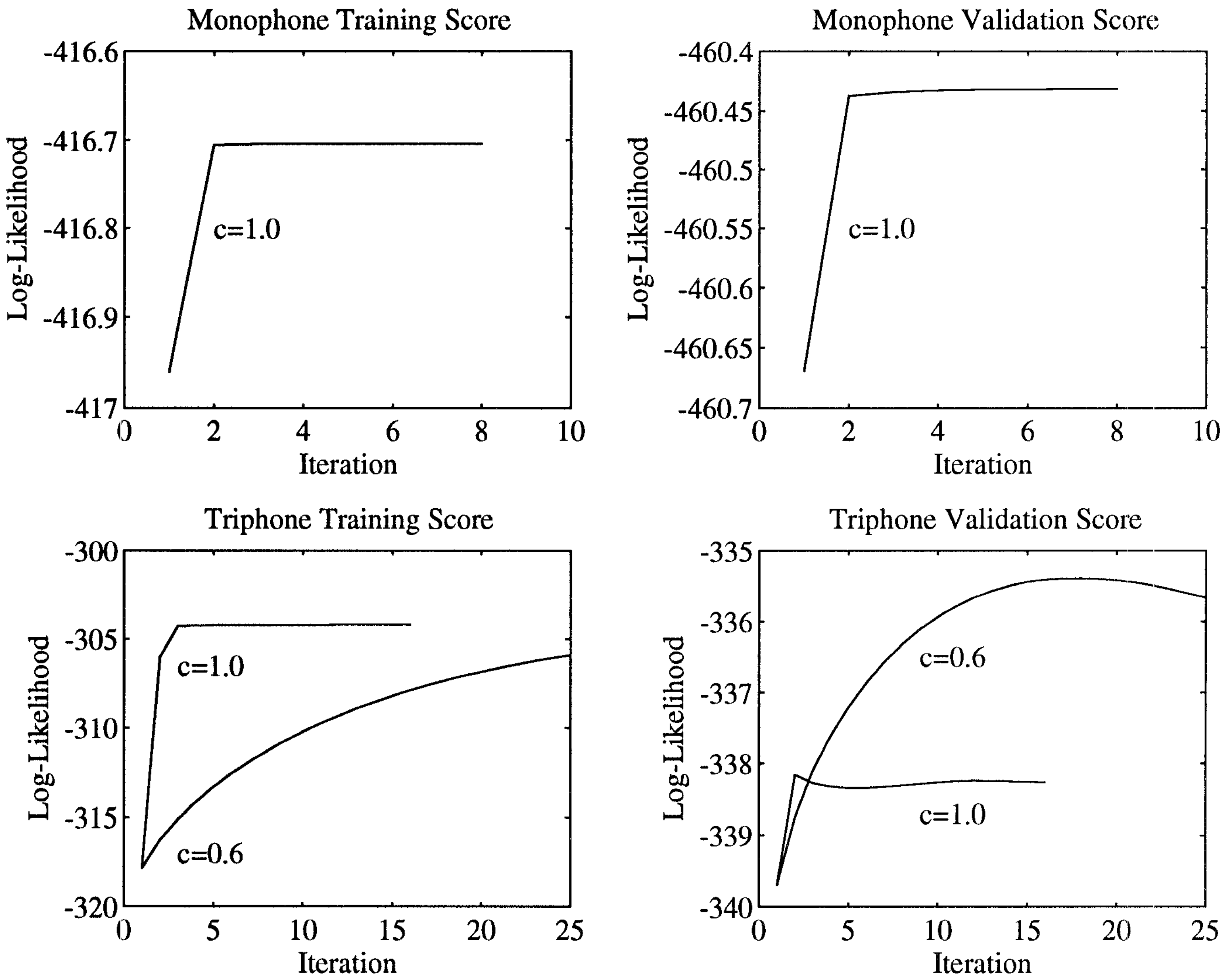

Figure 4: (Top) Training and validation scores in training a monophone model for /k/ using the $c = 1$ Baum Welch algorithm for $|\mathcal{T}| = 296$ and $|\mathcal{V}| = 78$. No overtraining is evident. (Bottom) Training and validation scores in training a triphone model for /cl-k-ix/ using the Baum Welch algorithm for $|\mathcal{T}| = 25$ and $|\mathcal{V}| = 7$ for $c = 1.0$ and $c = 0.6$. Overtraining is evident, but generalization performance is improved for $c = 0.6$. The final, trained monophone model is used as the initial triphone model.

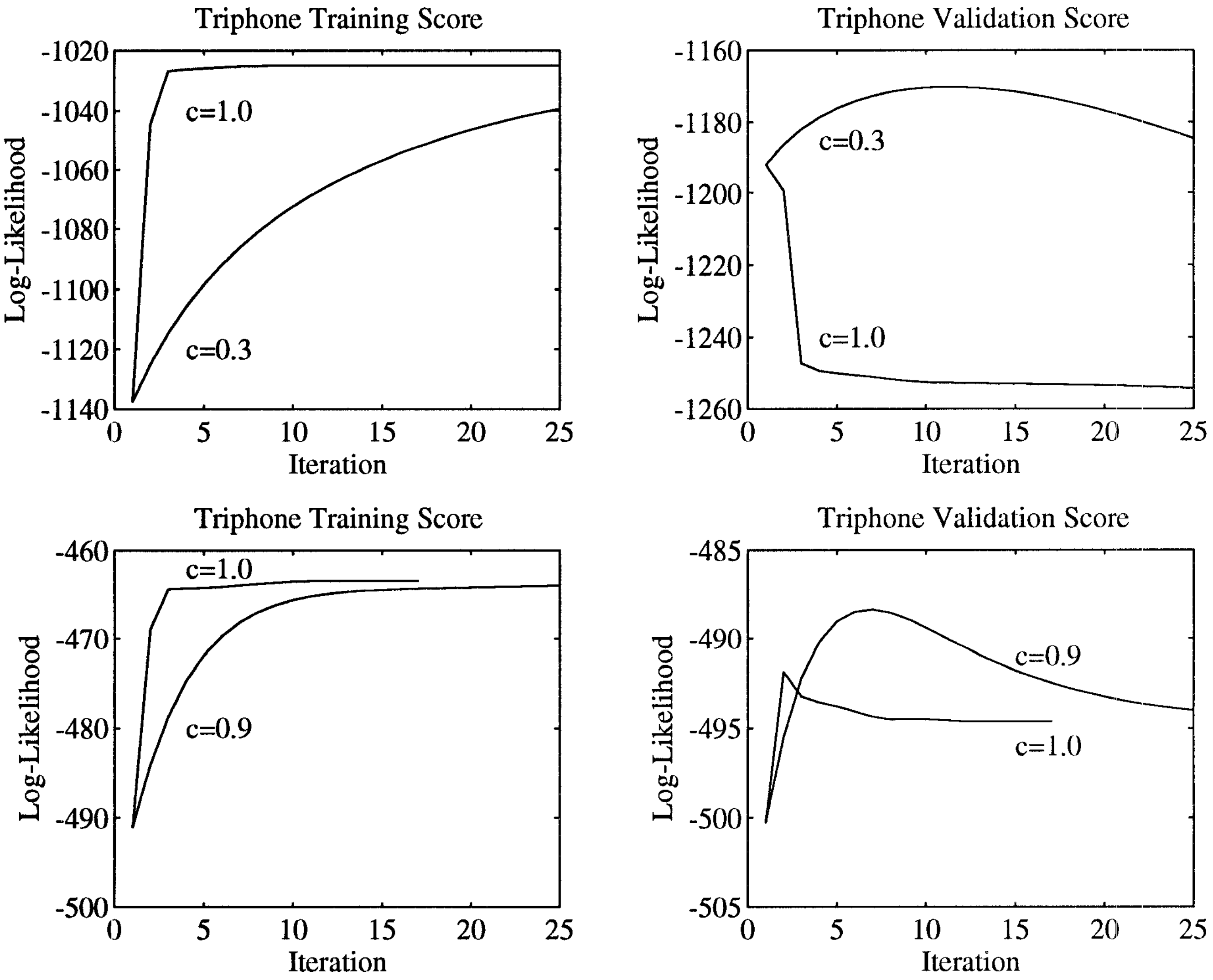

Figure 5: (Top) Training and validation scores in training a triphone model for /t-ay-m/ using the Baum Welch algorithm for $|\mathcal{T}| = 6$ and $|\mathcal{V}| = 6$ for $c = 1.0$ and $c = 0.3$. (Bottom) Training and validation scores in training a triphone model for /ax-vcl-b/ using the Baum Welch algorithm for $|\mathcal{T}| = 33$ and $|\mathcal{V}| = 11$ for $c = 1.0$ and $c = 0.9$.

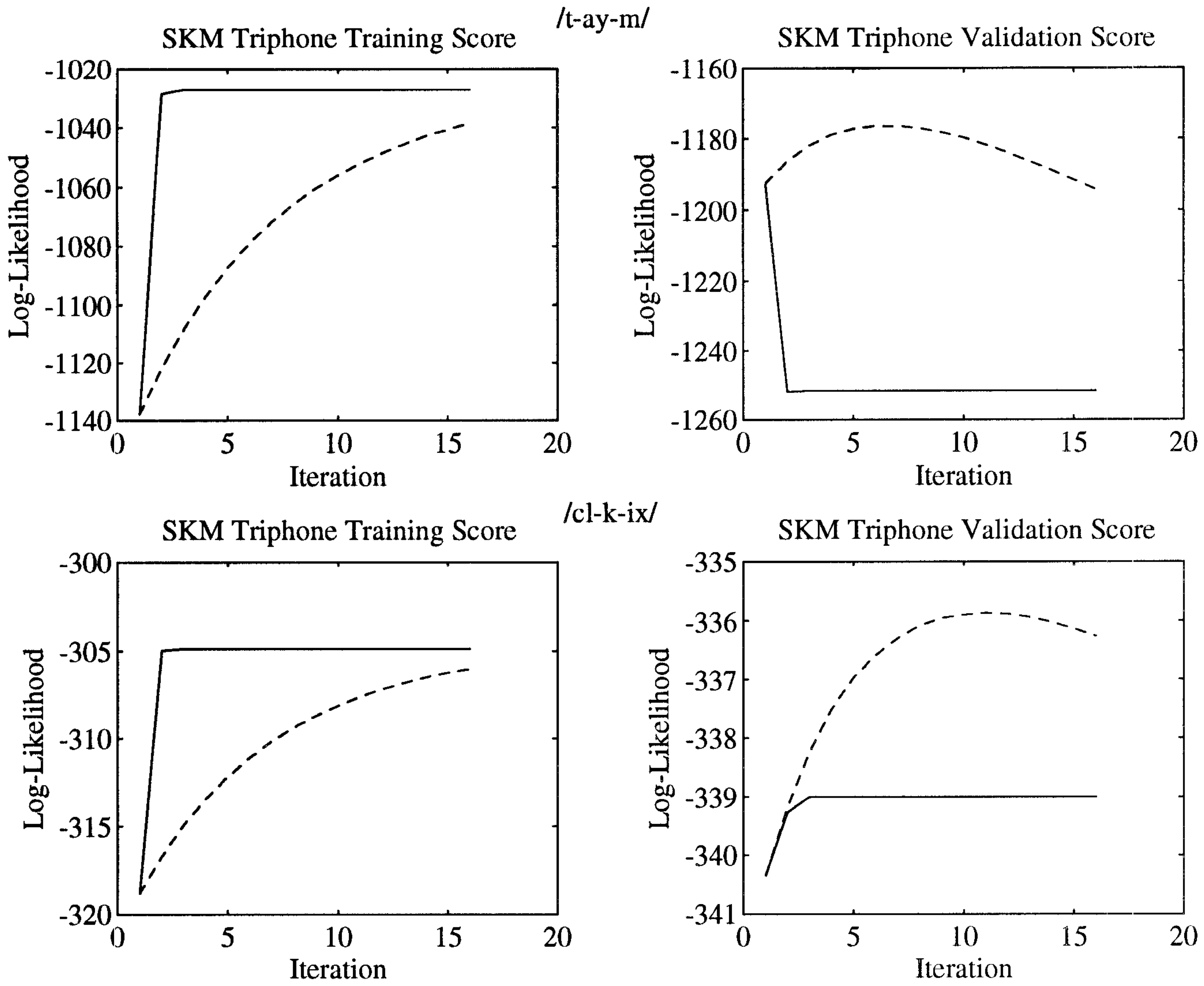

Figure 6: (Top) Training and validation scores in training a triphone model for /t-ay-m/ using the Segmental K-Means algorithm for $|\mathcal{T}| = 6$ and $|\mathcal{V}| = 6$ for $c = 1.0$ ('–') and $c = 0.1$ ('– –'). (Bottom) Training and validation scores in training a triphone model for /cl-k-ix/ using the SKM algorithm for $|\mathcal{T}| = 25$ and $|\mathcal{V}| = 7$ for $c = 1.0$ ('–') and $c = 0.1$ ('– –').

Table 2: Selected Phoneme and Triphone Frequencies from Dialect Region 1 of the TIMIT Database

Phoneme	Training Set	Test Set
/k/	296	78
/cl - k - ix/	25	7
vcl	613	164
ax-vcl-b	33	11
/ay/	163	45
/t-ay-m/	6	6

A context-dependent model is trained using only instances of the phoneme found in that context. A shortcoming of this approach is that it is difficult to find enough instances of a phoneme in all possible contexts to train reliable context-dependent models.

Typically a reliable monophone (context-independent) model is first trained using the usually large amount of context-indepent training data. This model is then used as an initial model in further training with context-dependent data.

An example is given here of training triphone models using moment decay to prevent overtraining. The experiment is conducted using data taken from Dialect Region 1 of the TIMIT database [?]. A model will be trained for the phoneme /k/ in the context /cl - k - ix/, as in "kettle". The occurences of several phonemes sampled from dialect region 1 are given in Table 2 It is clear that specifying the context reduces the amount of training data.

The model used is a three-state left-to-right model with no skips allowed. The observation distributions are single mixture, diagonal covariance Gaussians. The state transition probabilities are reestimated according to Equation 139. The means and covariances are reestimated according to Equations 148 and 149. The observations are 12 liftered, Mel-Frequency Cepstral coefficients, a frame energy term, and their first difference coefficients. The features and model reestimates are computed using the HTK Hidden Markov Model Toolkit [44].

An initial model for /k/ is found using a Segmental K-Means initialization procedure which makes use of all the training data. Only the observation distributions are varied. This model is used as the initial model in the Baum-Welch training algorithm. Although the training algorithm is essentially allowed to find a fixed point, no overtraining is evident in the context independent model (Figure 4,top).

The resulting monophone model for /k/ is used to initialize the Baum-Welch reestimation of /cl - k + ix/. As shown in Figure 4, bottom, overtraining occurs by the second iteration in the usual $c = 1$ Baum-Welch reestimation. When $c = 0.1$, however, overtraining is postponed until the 11th iteration and the overall score of the validation set is much improved.

Examples in training other triphones using the Baum-Welch and modified Segmental K-Means (only means and variances updated) are also given in Figures 5 and 6.

8 Conclusion

A uniform description of the Baum Welch, Segmental K-Means and N-Best maximum likelihood Hidden Markov Model training algorithms has been developed by describing these algorithms as instances of the Alternating Minimization procedure. It is shown that these algorithms can be distinguished by a single parameter which describes a varying restriction on the set of desirable distributions that defines each algorithm. This restriction determines the size of the support of the *a posteriori* desired distributions.

A procedure has been presented which is intended to improve the generalization of statistical models trained from data using the Alternating Minimization procedure under a maximum likelihood criterion. It is compared to other reestimation techniques which incorporate penalties to enforce parameter smoothness, and it is stressed that in this algorithm the penalty, or discounting, is applied to the data set, not the model set. While the confidence constraint may slow the training algorithms, in the examples given it involves very little additional calculation.

This technique is presented as a modification to the maximum likelihood criterion. However, it only relies upon a minimum divergence formulation and can be applied to other models and techniques which can be formulated in this framework, such as training under a Maximum Mutual Information criterion [16] which may be useful in improving the cross-class rejection described in the examples.

The technique tunes the rate at which the EM algorithm abandons prior information. It may be useful in applications such as speaker adaptation where a small amount of speaker dependent data is used to refine speaker independent models. It may also be useful in the opposite task of merging well-trained speaker-dependent models into robust, speaker-independent systems

A manner for choosing the correct value of c is not known; a fixed value may not even be appropriate. While it appears that lower values of c encourage generalization, lower values of c also slow the training algorithm. In this algorithm, there is a clear trade-off between generalization and speed of learning.

References

[1] S.E.Levinson. Structural methods in automatic speech recognition. *Proceedings of the I.E.E.E.*, 73(11):1625–1650, November 1985.

[2] S.E.Levinson, L.R.Rabiner, and M.M.Sondhi. An introuction to the application of the theory of probabilistic functions of a Markov process to automatic speech recognition. *The Bell Sysem Technical Journal*, 64(4):1035–1074, April 1984.

[3] Lawrence R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the I.E.E.E.*, 77(2):257–286, February 1989.

[4] L.R.Rabiner and B.H.Juang. An introduction to Hidden Markov Models. *I.E.E.E. A.S.S.P. Magazine*, pages 4–16, January 1986.

[5] A.P.Dempster, N.M.Laird, and D.B.Rubin. Maximum likelihood from incomplete data. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[6] I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. *Statistics and Decisions, Supplementary Issue Number 1*, pages 205–237, 1984.

[7] Neriv Merhav and Yariv Ephraim. Hidden Markov Modeling using a dominant state sequence with application to speech recognition. *Computer Speech and Language*, 5:327–339, 1991.

[8] Neriv Merhav and Yariv Ephraim. Maximum likelihood Hidden Markov Modeling using a dominant sequence of states. *I.E.E.E. Transactions on Signal Processing*, 39(9):2111–2115, November 1991.

[9] Richard Schwartz and Yen Lu Chow. The N-Best algorithm: An efficient and exact procedure for finding the N most likely sentence hypotheses. In *I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pages 81–84, 1990.

[10] Richard Schwartz et. al. New uses for the N-best sentence hypotheses within the BYBLOS speech recognition system. In *I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pages 1–4, 1992.

[11] L.E.Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[12] B.-H. Juang L.R.Rabiner. The segmental K-means algorithm for estimating parameters of Hidden Markov Models. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 38(9):1639–1641, Sept 1990.

[13] L.R. Rabiner, J.G. Wilpon, and B.-H.Juang. A segmental k-means training procedure for connected word recognition. *AT&T Technical Journal*, pages 21–40, May 1986.

[14] Solomon Kullback. *Information theory and statistics.* Wiley, New York, 1959.

[15] J. E. Shore and R.W. Johnson. Axiomatic derivation of the principle of maximum entropy and minimum cross-entropy. *I.E.E.E. Transactions on Information Theory*, 26(1):26–37, January 1980.

[16] Yariv Ephraim and Lawrence R. Rabiner. On the relations between modeling approaches for information sources. *I.E.E.E. Transactions on Information Theory*, 36(2):372–380, March 1990.

[17] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[18] Imre Csiszár and Janos Körner. *Information theory : coding theorems for discrete memoryless systems.* Academic Press, Orlando, 1981.

[19] Michael I. Miller and Donald L. Snyder. The role of likelihood and entropy in incomplete-data problems: Applications to estimating point-process intensities and Toeplitz constrained covariances. *Proceedings of the I.E.E.E.*, 75(7):892–907, July 1985.

[20] Y. Vardi, L.A. Shepp, and L. Kaufman. A statistical model for Positron Emission Tomography. *Journal of the American Statistical Association*, 80(39):8–20, March 1985.

[21] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 36:111–147, 1974.

[22] Kevin J. Coakley. A cross-validation procedure for stopping the EM algorithm and deconvolution of neutron depth profiling spectra. *I.E.E.E. Transactions on Nuclear Science*, 38(1):9–1, February 1991.

[23] Eric B. Baum and David Haussler. What size net gives valid generalization? *Neural Computation*, 1, 1989.

[24] Anders Krogh, Michael Brown, I. Saira Mian, Kimmen Sjölander, and David Haussler. Hidden Markov Models in computational biology: applications to protein modeling. *submitted for publication*, December 1992.

[25] Peter J. Green. On the use of the EM algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society*, 52(3):443–452, 1990.

[26] B. W. Silverman, M. C. Jones, J. D. Wilson, and D. W. Nychka. A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *Journal of the Royal Statistical Society*, 52(2):271–324, 1990.

[27] Richard Schwartz et. al. Robust smoothing methods for discrete Hidden Markov Models. In *I.E.E.E. International Conference on Acoustics, Speech and Signal Processing*, pages 548–551, 1989.

[28] Isaac Meilijson. A fast improvement to the EM algorithm on its own terms. *Journal of the Royal Statistical Society*, 51(1):127–138, 1989.

[29] Jerome R. Bellegarda and David Nahamoo. Tied mixture continuous parameter modeling for speech recognition. *I.E.E.E. Transactions of Acoustics, Speech and Signal Processing*, 38(12):2033–2045, December 1990.

[30] Bradley Efron. The geometry of exponential families. *The Annals of Statistics*, 6(2):362–376, 1978.

[31] Shun-Ichi Amari. *Differential-Geometrical methods in statistics.* Springer-Verlag, New York, 1985.

[32] S.E. Levinson. Continuously variable duration Hidden Markov Models for automatic speech recognition. *Computer Speech and Language*, 1(1):29–46, March 1986.

[33] Louis A. Liporace. Maximum likelihood estimation for multivariate observations of markov sources. *I.E.E.E. Transactions on Information Theory*, 28(5):729–734, September 1982.

[34] B.-H. Juang. Maximum-likelihood estimation for mixture multivariate stochastic observations of markov chains. *AT&T Technical Journal*, 64(6):1235–1249, July-August 1985.

[35] Stephen José Hanson and Lorein Y. Pratt. Comparing biases for minimal network construction with back-propagation. In David S. Touretzky, editor, *Advances in Neural Information Processings Systems 1*, pages 177–185, 1988.

[36] B.D.Ripley. Statistical aspects of neural networks. In *Seminarie European de statistique*. Chapman and Hall, 1992. to appear; available in neruoprose archives.

[37] William A. Gale and Kenneth W. Church. What's wrong with adding one? Unpublished Manuscript.

[38] Adrian W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika*, 71(2):353–360, 1984.

[39] David Haussler, Anders Krogh, I. Saira Mian, and Kimmen Sjolander. Protein modeling using Hidden Markov Models: analysis of globins. Technical Report UCSC-CRL-92-23, Computer and Information Sciences University of California, Santa Cruz, Santa Cruz, CA 95064, September 1992.

[40] Geoffrey E. Hinton, Terrence J. Sejnowski, and David H. Ackley. Boltzmann Machines: Constraint satisfaction networks that learn. Technical Report CMU-CS-84-119, Carnegie-Mellon University, Pittsburgh, PA 15213, May 1984.

[41] William J. Byrne. Alternating Minimization and Boltzmann Machine learning. *I.E.E.E. Transactions on Neural Networks*, 3(4):612–620, 1992.

[42] Ehud Weinstein, Meir Feder, and Alan V. Oppenheim. Sequential algorithms for parameter estimation based on the Kullback-Leibler information measure. *I.E.E.E. Transactions on Acoustics, Speech, and Signal Processing*, 38(9):1652–1654, September 1990.

[43] William Byrne, John Robinson, and Shihab Shamma. The auditory processing and recognition of speech. In *Proceedings of the Speech and Natural Language Workshop*, pages 325–331, October 1989.

[44] P.C. Woodland and S.J.Young. Benchmark DARPA RM results with the HTK portable HMM toolkit. In *Proceedings of the Speech and Natural Language Workshop*, September 1992.

A I-Projection of Q^p onto $\mathcal{D}_c^N$

From Equation 6, P^p should be found to minimize

$$D(P \parallel Q^p) = \sum_{I \in \mathcal{T}} \sum_{S \in \mathcal{S}} P(I,S) \log \frac{P(I,S)}{Q^p(I,S)} + \sum_{I \in \mathcal{T}^c} \sum_{S \in \mathcal{S}} P(I,S) \log \frac{P(I,S)}{Q^p(I,S)}. \tag{174}$$

The minimization can be performed independently on each sum if the conditions for membership of P^p in $\mathcal{D}_c^N$

$$\begin{aligned} P(I) &= c\ \hat{P}(I) \quad \forall I \in \mathcal{T} & (175) \\ |P^{-1}(\cdot|I)| &\leq N \quad \forall I \in \mathcal{T} & (176) \\ P(\mathcal{T}^c) &= 1-c & (177) \end{aligned}$$

are not violated. Through the log-sum inequality [18], the sum on $\mathcal{T}^c$ satisfies the inequality

$$\sum_{I \in \mathcal{T}^c} \sum_{S} P(I,S) \log \frac{P(I,S)}{Q^p(I,S)} \geq P(\mathcal{T}^c) \log \frac{P(\mathcal{T}^c)}{Q^p(\mathcal{T}^c)} \tag{178}$$

and achieves this lower bound if and only if

$$\frac{P(I,S)}{P(\mathcal{T}^c)} = \frac{Q^p(I,S)}{Q^p(\mathcal{T}^c)} \quad I \in \mathcal{T}^c. \tag{179}$$

Satisfying Equation 177 implies that

$$P^p(I,S) = (1-c)\,\frac{Q^p(I,S)}{Q^p(\mathcal{T}^c)} \quad I \in \mathcal{T}^c. \tag{180}$$

P^p is found for $I \in \mathcal{T}$ in the same way as Equation 25 is derived. The sum on $\mathcal{T}$ obeys

$$\sum_{I\in\mathcal{T}}\sum_{S\in\mathcal{S}} P(I,S)\,\log\frac{P(I,S)}{Q^p(I,S)} = \sum_{I\in\mathcal{T}}\sum_{S\in P^{-1}(\cdot|I)} P(I,S)\,\log\frac{P(I,S)}{Q^p(I,S)} \tag{181}$$

$$\geq \sum_{I\in\mathcal{T}} P(I)\,\log\frac{P(I)}{\sum_{S\in P^{-1}(\cdot|I)} Q^p(I,S)} \tag{182}$$

with equality if and only if

$$\frac{P(I,S)}{P(I)} = \frac{Q^p(I,S)}{\sum_{S\in P^{-1}(\cdot|I)} Q^p(I,S)} \quad I \in \mathcal{T}. \tag{183}$$

Satisfying Equations 175 and 176 and minimizing the term on the right hand side requires that

$$P^p(I,S) = \begin{cases} c\ \hat{P}(I)\ \frac{Q^p(I,S)}{\sum_{S\in B^p_N(I)} Q^p(I,S)} & S \in P^{-1}(\cdot|I),\ I \in \mathcal{T} \\ 0 & S \notin P^{-1}(\cdot|I),\ I \in \mathcal{T} \end{cases}. \tag{184}$$

Combining the distributions which minimize the two sums yields the I-projection of Q^p onto $\mathcal{D}_c$

$$P^p(I,S) = \begin{cases} c\ \hat{P}(I)\ \frac{Q^p(I,S)}{\sum_{S\in B^p_N(I)} Q^p(I,S)} & S \in P^{-1}(\cdot|I),\ I \in \mathcal{T} \\ 0 & S \notin P^{-1}(\cdot|I),\ I \in \mathcal{T} \\ (1-c)\ \frac{Q^p(I,S)}{Q^p(\mathcal{T}^c)} & I \in \mathcal{T}^c \end{cases}. \tag{185}$$

The divergence from Q^p to $\mathcal{D}_c$ is found by substituting Equations 178 and 182 into Equation 174

$$D(P^p \parallel Q^p) = \sum_{\mathcal{T}} P^p(I)\log\frac{P^p(I)}{\sum_{S\in B^p_N(I)} Q^p(I,S)} + P^p(\mathcal{T}^c)\log\frac{P^p(\mathcal{T}^c)}{Q^p(\mathcal{T}^c)} \tag{186}$$

$$= \sum_{\mathcal{T}} c\hat{P}(I)\log c\hat{P}(I) - \sum_{\mathcal{T}} c\hat{P}(I)\log(\sum_{S\in B^p_N(I)} Q^p(I,S)\ ) + (1-c)\log\frac{1-c}{Q^p(\mathcal{T}^c)} \tag{187}$$

$$= c\log c + c\sum_{\mathcal{T}}\hat{P}(I)\log\hat{P}(I) + (1-c)\log\frac{1-c}{Q^p(\mathcal{T}^c)} - c\sum_{\mathcal{T}}\hat{P}(I)\log(\sum_{S\in B^p_N(I)} Q^p(I,S)). \tag{188}$$

B VDHMM Reestimation

In this section, the association parameter a is fixed at some value in $[0,1]$. The updated Hidden Markov Model $Q^{p+1} = \{b^{p+1}, d^{p+1}\}$ is found from the current desired distribution P^p by solving

$$\min_Q D(P^p \parallel Q) = \min_Q \sum_{I,S} P^p(I,S)\ \ \log\frac{P^p(I,S)}{Q(I,S)} = \max_Q \sum_{I,S} P^p(I,S)\ \ \log Q(I,S) \tag{189}$$

$$
\begin{aligned}
&= \max_Q [\sum_{I,S} P^p(I,S) \ \ \log Q(I|S) + \sum_S P^p(S) \ \ \log Q(S)\,] && (190)\\
&= \max_Q [\ \sum_{I,S} P^p(I,S) \ \ \sum_t \log b_{S_t}(I_t) + \sum_S P^p(S) \ \ \sum_n \log d_n(\tau_n) \ \,] \\
&= \max_b \sum_{I,S} P^p(I,S) \ \ \sum_t \log b_{S_t}(I_t) + \max_d \sum_S P^p(S) \ \ \sum_n \log d_n(\tau_n) \\
&= \max_b \sum_{I,S} P^p(I,S) \ \ \sum_t \sum_i \sum_n \delta_n(S_t)\delta_i(I_t) \log b_n(i) + \max_d \sum_S P^p(S) \ \ \sum_n \sum_\tau \delta_\tau(\tau_n) \log d_n(\tau) \\
&= \sum_n \max_{b_n} \sum_i \log b_n(i) \sum_t \sum_{I,S} P^p(I,S) \ \ \delta_n(S_t)\delta_i(I_t) + \sum_n \max_{d_n} \sum_\tau \log d_n(\tau) \sum_S P^p(S) \ \ \delta_\tau(\tau_n) \\
&= \sum_n \max_{b_n} \sum_i \log b_n(i) \ \ \sum_t P^p(I_t = i, S_t = n) + \sum_n \max_{d_n} \sum_\tau \log d_n(\tau) \ \ P^p(\tau_n = \tau). && (191)
\end{aligned}
$$

The distribution Q^{p+1} which maximizes these terms is ([2], Lemma 2)

$$
\begin{aligned}
d_n^{p+1}(\tau) &= P^p(\tau_n = \tau) && (192)\\
b_n^{p+1}(i) &= \frac{\sum_t P^p(I_t = i, S_t = n)}{\sum_t P^p(S_t = n)} = \frac{\sum_t P^p(I_t = i, S_t = n)}{\bar{\tau}_n^p} && (193)
\end{aligned}
$$

where $\bar{\tau}_n^p = \sum_t P^p(S_t = n) = \sum \tau P^p(\tau_n = \tau)$.

For $c = 1$, the usual algorithm, Q^{p+1} is found from P^p by substituting

$$
P^p(I,S) = \hat{P}(I) \ \ Q^p(S|B_N^p(I), I) \tag{194}
$$

into Equations 192 and 193 to obtain

$$
\begin{aligned}
\tilde{d}_n^{p+1}(\tau) &= \sum_{I \in \mathcal{T}} \hat{P}(I) \sum_{S:\tau_n = \tau} Q^p(S|B_N^p(I), I) && (195)\\
\tilde{b}_n^{p+1}(i) &= \frac{1}{\bar{\tau}_n^p} \sum_t \sum_{I \in \mathcal{T}: I_t = i} \hat{P}(I) \sum_{S:S_t = n} Q^p(S|B_N^p(I), I) && (196)
\end{aligned}
$$

where the tilde denotes that these are the $c = 1$ reestimates from Q^p.

For $c < 1$, the reestimate of the duration distribution can be found by substituting Equation 51 into Equations 192 and 193 and following the derivation of Equation 100

$$
\begin{aligned}
d_n^{p+1}(\tau) &= c \ \tilde{d}_n^{p+1}(\tau) + \frac{1-c}{1 - Q^p(\mathcal{T})} \ [d_n^p(\tau) - Q^p(\mathcal{T}, \tau_n = \tau)] && (197)\\
b_n^{p+1}(i) &= c \, \tilde{b}_n^{p+1}(i) + \frac{1-c}{1 - Q^p(\mathcal{T})} \ [b_n^p(i) - \frac{1}{\bar{\tau}_n^p} \sum_{I \in \mathcal{T}} \sum_t Q^p(I, S : I_t = i, S_t = n)]. && (198)
\end{aligned}
$$

B.1 HMM Reestimation with Constrained Duration Densities

When $Q(S) = \prod_{n,n'=1}^N a_{n,n'}^{\#_{n,n'}(S)}$, finding Q^{p+1} from P^p proceeds from Equation 190 by finding the parameters a which maximize

$$
\begin{aligned}
\sum_S P^p(S) \log Q(S) &= \sum_S P^p(S) \sum_{n,n'} \#_{n,n'}(S) \log \ a_{n,n'} && (199)\\
&= \sum_{n,n'} \log \ a_{n,n'} \sum_S P^p(S) \#_{n,n'}(S) && (200)\\
&= \sum_{n,n'} \#_{n,n'}^{p+1} \log \ a_{n,n'} && (201)\\
&= \sum_n [\sum_{n'} a_{n,n'} \log \ a_{n,n'}\,] && (202)
\end{aligned}
$$

where

$$\#^{p+1}_{n,n'} = \sum_{S} P^p(S)\#_{n,n'}(S) \tag{203}$$

and also

$$\#^{p+1}_{n} = \sum_{n'}\sum_{S} P^p(S)\#_{n,n'}(S). \tag{204}$$

If a^{p+1} is chosen as

$$a^{p+1}_{n,n'} = \frac{\#^{p+1}_{n,n'}}{\#^{p+1}_{n}} \tag{205}$$

then the bracketed term above is maximized ([2], Lemma 2) and Q^{p+1} is found.

The moments are derived in the same way as Equation 100

$$\begin{aligned}
\#^{p+1}_{n,n'} &= \sum_{S} P^p(S)\#_{n,n'}(S) & (206)\\
&= \sum_{I\in\mathcal{T},S} P^p(I,S)\#_{n,n'}(S) + \sum_{I\in\mathcal{T}^c,S} P^p(I,S)\#_{n,n'}(S) & (207)\\
&= c\sum_{I\in\mathcal{T}} \hat{P}(I)\sum_{S}\#_{n,n'}(S)Q^p(S|B^p_N(I),I) + (1-c)\sum_{S}\#_{n,n'}(S)\sum_{I\in\mathcal{T}^c} Q^p(I,S) & (208)\\
&= c\sum_{I\in\mathcal{T}} \hat{P}(I)E_{Q^p(S|I)}[\#_{n,n'}(S)|B^p_N(I),I] + (1-c)\sum_{S}\#_{n,n'}(S)[Q^p(S) - \sum_{I\in\mathcal{T}} Q^p(I,S)] & \\
&= c\tilde{\#}^{p+1}_{n,n'} + (1-c)[\#^p_{n,n'} - \sum_{I\in\mathcal{T}}\sum_{S}\#_{n,n'}(S)Q^p(I,S)] & (209)
\end{aligned}$$

where the usual, $c = 1.0$, reestimate of the moment is

$$\tilde{\#}^{p+1}_{n,n'} = \sum_{I\in\mathcal{T}} \hat{P}(I)E_{Q^p(S|I)}[\#_{n,n'}(S)|B^p_N(I),I]. \tag{210}$$

Under the small data set assumption this becomes

$$\#^{p+1}_{n,n'} = c\,\tilde{\#}^{p+1}_{n,n'} + (1-c)\,\#^{p}_{n,n'}. \tag{211}$$

C I-Projection onto $\mathcal{D}^N_c \cap \mathcal{M}_c$

Denote the I-Projection of Q^p onto $\mathcal{D}^N_c$, defined in Equation 51, as P^*. Because $\mathcal{D}^N_c$ and $\mathcal{D}^N_c \cap \mathcal{M}_c$ are both convex, the I-Projection of P^* onto $\mathcal{M}_c$ is the I-Projection of Q^p onto $\mathcal{D}^N_c \cap \mathcal{M}_c$ ([17] Theorem 3.2). Denote the I-Projection of P^* on $\mathcal{D}^N_c \cap \mathcal{M}_c$ as P^{**}. It is well known that

$$P^{**}(I,S) = P^*(I,S)\ \exp\{\lambda\cdot g(I,S)\ 1_{\mathcal{T}^c}(I)\}. \tag{212}$$

where the Langrange multipliers λ are chosen so that $P^{**} \in \mathcal{D}^N_c \cap \mathcal{M}_c$. Suppose $\lambda = w^0 - w^p$. Evaluating the condition for membership in $\mathcal{M}_c$ yields

$$\begin{aligned}
\sum_{I,S} P^{**}(I,S)\ 1_{\mathcal{T}^c}(I)\ g(I,S) &= \sum_{I\in\mathcal{T}^c,S} P^*(I,S)\ \exp\{\lambda\cdot g(I,S)\} & (213)\\
&= \sum_{I\in\mathcal{T}^c,S} (1-c)\ \frac{Q^p(I,S)}{Q^p(\mathcal{T}^c)}\ \exp\{\lambda\cdot g(I,S)\} & (214)\\
&= \frac{1-c}{Q^p(\mathcal{T}^c)}\sum_{I\in\mathcal{T}^c,S} \exp\{w^p\cdot g(I,S)\}\ \exp\{\lambda\cdot g(I,S)\} & (215)\\
&= \frac{1-c}{Q^p(\mathcal{T}^c)}\sum_{I\in\mathcal{T}^c,S} \exp\{w^0\cdot g(I,S)\} & (216)
\end{aligned}$$

$$\begin{aligned} &= \frac{1-c}{Q^p(\mathcal{T}^c)} \sum_{I \in \mathcal{T}^c, S} Q^0(I,S) && (217) \\ &= (1-c)\, q^0 && (218) \end{aligned}$$

where the last equation assumes that Q^0 and Q^p obey the small data assumptions.

The I-Projection of Q^p onto $\mathcal{D}_c^N \cap \mathcal{M}_c$ is therefore

$$\begin{aligned} P^{**}(I,S) &= P^*(I,S) \exp\{\lambda \cdot g(I,S)\, 1_{\mathcal{T}^c}(I)\} && (219) \\ &= \begin{cases} c\, \hat{P}(I)\, Q^p(S|B_N^p(I), I) & I \in \mathcal{T} \\ (1-c)\, \frac{Q^p(I,S)}{Q^p(\mathcal{T}^c)} \exp\{\lambda \cdot g(I,S)\} & I \in \mathcal{T}^c \end{cases} && (220) \\ &= \begin{cases} c\, \hat{P}(I)\, Q^p(S|B_N^p(I), I) & I \in \mathcal{T} \\ (1-c)\, Q^0(I,S) & I \in \mathcal{T}^c \end{cases} . && (221) \end{aligned}$$

D Sequential Algorithms

From Equation 52,

$$\frac{\delta}{\delta w} D(\mathcal{D}_c^N \parallel Q) = -c \sum_{T} \hat{P}(I)\ \frac{\delta}{\delta w} \log Q(I, B_N(I)) - (1-c) \frac{\delta}{\delta w} \log Q(\mathcal{T}^c). \tag{222}$$

The first term would be evaluated as

$$\frac{\delta}{\delta w} \log Q(I, B_N(I)) = \frac{1}{Q(I, B_N(I))} \frac{\delta}{\delta w} \sum_{B_N(I)} Q(I,S) \tag{223}$$

however because B_N is a function of w it is not necessarily true that

$$\frac{\delta}{\delta w} \sum_{B_N(I)} Q(I,S) = \sum_{B_N(I)} \frac{\delta}{\delta w} Q(I,S). \tag{224}$$

In general, it is too strong an assumption that small changes in parameters will not affect the most likely hidden variables. For the marginal likelihood $Q(I, B_N(I)) = \sum_S Q(I,S)$, so this is not a problem. Only this case will be considered.

The first derivative is somewhat difficult to evaluate directly, so the approach used in Section 2 of [5] will be followed. Define

$$z_w = \sum_{I,S} \exp\ w \cdot g(I,S). \tag{225}$$

so

$$Q(I,S) = \frac{\exp\ w \cdot g(I,S)}{z_w} \tag{226}$$

and

$$Q(I) = \frac{\sum_{S'} \exp\ w \cdot g(I,S')}{z_w}. \tag{227}$$

Note that $Q(S|I) = \frac{Q(I,S)}{Q(I)}$ so

$$\begin{aligned} Q(S|I) &= \frac{\exp\ w \cdot g(I,S)}{\sum_{S'} \exp\ w \cdot g(I,S')} && (228) \\ &= \frac{\exp\ w \cdot g(I,S)}{z_w(I)} && (229) \end{aligned}$$

where

$$z_w(I) = \sum_{S'} \exp\ w \cdot g(I, S') \tag{230}$$

It follows that

$$\begin{align} \log Q(I) &= \log Q(I, S) - \log Q(S|I) \tag{231} \\ &= \log z_w(I) - \log z_w . \tag{232} \end{align}$$

It is now easier to take the derivatives:

$$\begin{align} \frac{\delta}{\delta w} \log z_w(I) &= \frac{1}{z_w(I)} \sum_I \frac{\delta}{\delta w} \exp w \cdot g(I, S) \tag{233} \\ &= \sum_S g(I, S)\, Q(S|I) \tag{234} \end{align}$$

and

$$\begin{align} \frac{\delta}{\delta w} \log z_w &= \frac{1}{z_w} \sum_{I,S} \frac{\delta}{\delta w} \exp\ w \cdot g(I, S) \tag{235} \\ &= \sum_{I,S} g(I, S)\, Q(I, S) \tag{236} \\ &= q \tag{237} \end{align}$$

so that

$$\frac{\delta}{\delta w} \log Q(I, S) = \sum_S g(I, S)\, Q(S|I) - q. \tag{238}$$

Inserting this into the first term of Equation 222 yields

$$\begin{align} \sum_{\mathcal{T}} \hat{P}(I) \frac{\delta}{\delta w} \log Q(I) &= \sum_{\mathcal{T},\mathcal{S}} g(I, S)\, \hat{P}(I)\, Q(S|I) - q \tag{239} \\ &= \tilde{p} - q. \tag{240} \end{align}$$

The derivative of $\log Q(\mathcal{T}^c)$ is found via

$$\begin{align} \frac{\delta}{\delta w} Q(\mathcal{T}^c) &= \sum_{\mathcal{T}^c,\mathcal{S}} \frac{\delta}{\delta w} Q(I, S) = \sum_{\mathcal{T}^c,\mathcal{S}} \frac{\delta}{\delta w} \frac{\exp\ w \cdot g(I, S)}{z_w} \tag{241} \\ &= \sum_{\mathcal{T}^c,\mathcal{S}} \frac{1}{z_w} \frac{\delta}{\delta w} \exp\ w \cdot g(I, S) - \sum_{\mathcal{T}^c,\mathcal{S}} \frac{\exp w \cdot g(I, S)}{a_w^2} \frac{\delta}{\delta w} z_w \tag{242} \\ &= \sum_{\mathcal{T}^c,\mathcal{S}} g(I, S) \frac{\exp\ w \cdot g(I, S)}{z_w} - \sum_{\mathcal{T}^c,\mathcal{S}} Q(I, S) \frac{1}{z_w} \frac{\delta}{\delta w} \sum_{I',S'} \exp\ w \cdot g(I', S') \notag \\ &= \sum_{\mathcal{T}^c,\mathcal{S}} g(I, S)\, Q(I, S) - \sum_{\mathcal{T}^c,\mathcal{S}} Q(I, S) \frac{1}{z_w} \sum_{I',S'} g(I', S') \frac{\exp\ w \cdot g(I', S')}{z_w} \tag{243} \\ &= \sum_{\mathcal{T}^c,\mathcal{S}} g(I, S)\, Q(I, S) - Q(\mathcal{T}^c) \sum_{I',S'} g(I', S') Q(I', S') \tag{244} \\ &= \sum_{\mathcal{T}^c,\mathcal{S}} g(I, S)\, Q(I, S) - Q(\mathcal{T}^c)\, q \tag{245} \\ &= Q(\mathcal{T})\, q - \sum_{\mathcal{T},\mathcal{S}} g(I, S)\, Q(I, S) \tag{246} \end{align}$$

so that

$$\frac{\delta}{\delta w} \log Q(\mathcal{T}^c) = \frac{1}{Q(\mathcal{T}^c)} \left(Q(\mathcal{T})\, q - \sum_{\mathcal{T},\mathcal{S}} g(I, S)\, Q(I, S) \right). \tag{247}$$

Substituting this and Equation 240 into Equation 222 yields

$$\frac{\delta}{\delta w} D(\mathcal{D}_c \,\|\, Q) = -c\,(\tilde{p} - q) + (1-c)\,\frac{1}{Q(\mathcal{T}^c)}(\, Q(\mathcal{T})\, q - \sum_{\mathcal{T},\mathcal{S}} g(I,S)\, Q(I,S)\,). \tag{248}$$

E Reestimation of HMM Gaussian Observation Distributions

To model continuous observations taken from $\Re^D$, the density of a Hidden Markov Model with a single mixture, Gaussian observation density can be written as

$$f(I,S) = \prod_{t=1}^{T} b_{S_t}(I_t) \prod_{n,n'=1}^{N} a_{n,n'}^{\#_{n,n'}(S_t} \tag{249}$$

where the state dependent observation density has the form

$$b_n(I) = \frac{1}{(2\pi)^{D/2}} |\Sigma_n|^{\frac{1}{2}} \exp[-\frac{1}{2}\,(I-\mu)'\,\Sigma_n^{-1}\,(I-\mu)\,]. \tag{250}$$

From a model Q^p parameterized by $\{a^p, \mu^p, \Sigma^p\}$, it is necessary to find a new set of parameters which maximize the auxiliary function, Equation 88. This will be done by differentiating with respect to each parameter, setting the derivatives to zero, and solving for the new parameter set.

Finding μ_n^{p+1}: Following [29], using $\frac{\partial}{\partial \mu_n} \log\, b_n(i) = \Sigma_n^{-1}(i - \mu_n)$, it follows that

$$\begin{aligned}
\frac{\partial}{\partial \mu_n} \log f(I,S) &= \frac{\partial}{\partial \mu_n}[\sum_t \log b_{S_t}(I_t) + \sum_t \log a_{S_t, S_{t,t+1}}\,] && (251)\\
&= \frac{\partial}{\partial \mu_n} \sum_t \sum_{n'} \delta_{n'}(S_t)\, \log b_{n'}(I_t) && (252)\\
&= \sum_t \delta_n(S_t)\, \Sigma_n^{-1}\,(I_t - \mu_n). && (253)
\end{aligned}$$

Differentiation and maximization of the auxiliary function requires finding μ_n^{p+1} as the solution of

$$\frac{\partial}{\partial \mu_n}[\,\frac{c}{|T|} \sum_{i \in T} E_{Q^p}[\log q(S,I)|i = I] + (1-c) E_{Q^p} \log q(S,I)\,] = 0. \tag{254}$$

Consider first the $c = 1$ case which yields the usual, EM, reestimate of μ_n, denoted $\tilde{\mu}_n^{p+1}$. This requires solving

$$\frac{\partial}{\partial \mu_n} \frac{1}{|T|} \sum_{i \in T} E_{Q^p}[\log q(S,I)|i = I] = 0. \tag{255}$$

Evaluating the expectation yields

$$\begin{aligned}
\frac{\partial}{\partial \mu_n} E_{Q^p}[\log Q(S,I)|i = I] &= E_{Q^p}[\sum_t \delta_n(S_t) \Sigma_n^{-1}(I_t - \mu_n)|i = I] && (256)\\
&= \Sigma_n^{-1}[E_{Q^p}[\sum_t \delta_n(S_t) I_t|i = I] - E_{Q^p}[\#_n(S)|i = I]\mu_n] && (257)
\end{aligned}$$

where the final step uses $\#_n(S) = \sum_t \delta_n(S_t)$ to denote the number of occurences of state n in sequence S. The term $\frac{1}{|T|} \sum_{i \in T}\; E_{Q^p}[\#_n(S)|i = I]$ is denoted $\tilde{\#}_n^{p+1}$ so that

$$\frac{\partial}{\partial \mu_n} \frac{1}{|T|} \sum_{i \in T} E_{Q^p}[\,\log Q(S,I)|i = I\,] = \Sigma_n^{-1}[\,\frac{1}{|T|} \sum_{I \in T}\; E_{Q^p}[\sum_t \delta_n(S_t)\, I_t|i = I] - \tilde{\#}_n^{p+1}\, \mu_n\,]. \tag{258}$$

Setting this to zero yields

$$\tilde{\mu}_n^{p+1}\, \tilde{\#}_n^{p+1} = \frac{1}{|T|} \sum_{i \in T} E_{Q^p}[\sum_t \delta_n(S_t)\, I_t | i = I]. \tag{259}$$

For $c < 1$, the derivative of the second term in Equation 254 can be found as

$$\begin{aligned}
\frac{\partial}{\partial \mu_n} E_{Q^p} \log q(S, I) &= E_{Q^p}[\, \Sigma_n^{-1} \sum_t \delta_n(S_t)\, (I_t - \mu_n)] && (260)\\
&= \Sigma_n^{-1}\, E_{Q^p}[\sum_t \delta_n(S_t)\, I_t\,] - \Sigma_n^{-1}\, E_{Q^p}[\sum_t \delta_n(S_t)\,]\, \mu_n && (261)\\
&= \Sigma_n^{-1}\, E_{Q^p}[\sum_t \delta_n(S_t)\, I_t\,] - \Sigma_n^{-1}\, E_{Q^p}[\, \#_n(S)\,]\, \mu_n && (262)\\
&= \Sigma_n^{-1}\, E_{Q^p}[\, E_{Q^p}[\sum_t \delta_n(S_t)\, I_t\, |S_t\,]\,] - \Sigma_n^{-1}\, \#_n^p\, \mu_n && (263)\\
&= \Sigma_n^{-1}\, E_{Q^p}[\sum_t \delta_n(S_t)\, E_{Q^p}[\, I_t\, |S_t\,]\,] - \Sigma_n^{-1}\, \#_n^p\, \mu_n && (264)\\
&= \Sigma_n^{-1}\, E_{Q^p}[\sum_t \delta_n(S_t)\, \mu_{S_t}^p\,] - \Sigma_n^{-1}\, \#_n^p\, \mu_n && (265)\\
&= \Sigma_n^{-1}\, \mu_n^p\, E_{Q^p}[\sum_t \delta_n(S_t)\,] - \Sigma_n^{-1}\, \#_n^p\, \mu_n && (266)\\
&= \Sigma_n^{-1}\, \mu_n^p\, \#_n^p - \Sigma_n^{-1}\, \#_n^p\, \mu_n && (267)
\end{aligned}$$

where $\#_n^p$ is used to denote $E_{Q^p}\, \#_n(S)$.
The derivative of the auxiliary function for $c < 1$ is therefore

$$\frac{\partial}{\partial \mu_n}[\, \frac{c}{|T|} \sum_{i \in T} E_{Q^p}[\log q(S, I)|i = I] + (1 - c) E_{Q^p} \log Q(S, I)\,] = \tag{268}$$

$$c\, \Sigma_n^{-1}(\frac{1}{|T|} \sum_{I \in T} E_{Q^p}[\sum_t \delta_n(S_t)\, I_t | i = I] - \tilde{\#}_n^{p+1}\, \mu_n) + (1 - c)\, \Sigma_n^{-1}\, \#_n^p(\mu_n^p - \mu_n).$$

Setting this to zero yields

$$\begin{aligned}
\tilde{\mu}_n^{p+1} &= \frac{c\, \frac{1}{|T|}\, \sum_{I \in T}\, E_{Q^p}[\sum_t \delta_n(S_t)\, I_t | i = I] + (1 - c)\, \#_n^p\, \mu_n^p}{c\, \tilde{\#}_n^{p+1} + (1 - c)\, \#_n^p} && (269)\\
&= \frac{c\, \tilde{\mu}_n^{p+1}\, \tilde{\#}_n^{p+1} + (1 - c)\, \#_n^p\, \mu_n^p}{c\, \tilde{\#}_n^{p+1} + (1 - c)\, \#_n^p} && (270)\\
&= \frac{c\, \tilde{\mu}_n^{p+1} + (1 - c)\, (\#_n^p / \tilde{\#}_n^{p+1})\, \mu_n^p}{c + (1 - c)\, (\#_n^p / \tilde{\#}_n^{p+1})}. && (271)
\end{aligned}$$

Finding Σ^{p+1}: After [29], finding Σ^{p+1} requires solving

$$0 = \frac{\partial}{\partial \Sigma_n}[\, \frac{c}{|T|} \sum_{I \in T} E_{Q^p}[\log q(I, S)|i = I] + (1 - c)\, E_{Q^p} \log q(I, S)\,]. \tag{272}$$

Using $\frac{\partial}{\partial \Sigma_n} \log q(I, S) = \Sigma_n - (I - \mu_n)'(I - \mu_n)$, and following the first step in the derivation of μ_n^{p+1}

$$\frac{\partial}{\partial \Sigma_n} \log q(I, S) = \sum_t \delta_n(S_t)[\, \Sigma_n - (I_t - \mu_n)'(I_t - \mu_n)\,]. \tag{273}$$

Inserting this into Equation 272 yields

$$\frac{\partial}{\partial \Sigma_n}[\ \frac{c}{|T|}\sum_{I \in T} E_{Q^p}[\log q(I,S)|i = I] + (1-c)\ E_{Q^p} \log q(I,S)\ ] = \tag{274}$$
$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)|i = I\ ]\ \Sigma_n - \frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \mu_n)'(I_t - \mu_n)|i = I]$$
$$+(1-c)\ E_{Q^p}[\sum_t \delta_n(S_t)]\Sigma_n - (1-c) E_{Q^p}[\sum_t \delta_n(S_t)(I_t - \mu_n)'(I_t - \mu_n)].$$

As in the derivation of μ_n^{p+1}, the first term in Equation 274 is

$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)|i = I\ ] = c\ \tilde{\#}_n^{p+1}. \tag{275}$$

The second term in Equation 274 is evaluated using Equation 259. It follows that

$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \mu_n)'(I_t - \mu_n)|i = I] = \tag{276}$$

$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \tilde{\mu}_n^{p+1} + \tilde{\mu}_n^{p+1} - \mu_n)'(I_t - \tilde{\mu}_n^{p+1} + \tilde{\mu}_n^{p+1} - \mu_n)|i = I] = \tag{277}$$

$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \tilde{\mu}_n^{p+1})'(I_t - \tilde{\mu}_n^{p+1})|i = I] + \tag{278}$$
$$2\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \tilde{\mu}_n^{p+1})'(\tilde{\mu}_n^{p+1} - \mu_n)|i = I] + c\ (\tilde{\mu}_n^{p+1} - \mu_n)'(\tilde{\mu}_n^{p+1} - \mu_n) =$$

$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \tilde{\mu}_n^{p+1})'(I_t - \tilde{\mu}_n^{p+1})|i = I] + \tag{279}$$
$$2\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \tilde{\mu}_n^{p+1})'|i = I](\tilde{\mu}_n^{p+1} - \mu_n) + c\ (\tilde{\mu}_n^{p+1} - \mu_n)'(\tilde{\mu}_n^{p+1} - \mu_n) =$$

$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \tilde{\mu}_n^{p+1})'(I_t - \tilde{\mu}_n^{p+1})|i = I] + \tag{280}$$
$$2\ c\ \tilde{\#}_n^{p+1}(\tilde{\mu}_n^{p+1} - \mu_n)'(\tilde{\mu}_n^{p+1} - \mu_n) + c\ (\tilde{\mu}_n^{p+1} - \mu_n)'(\tilde{\mu}_n^{p+1} - \mu_n) =$$

$$\frac{c}{|T|}\sum_{i \in T} E_{Q^p}[\ \sum_t \delta_n(S_t)\ (I_t - \tilde{\mu}_n^{p+1})'(I_t - \tilde{\mu}_n^{p+1})|i = I] + \tag{281}$$
$$c(2\tilde{\#}_n^{p+1} + 1)(\tilde{\mu}_n^{p+1} - \mu_n)'(\tilde{\mu}_n^{p+1} - \mu_n).$$

The third term in Equation 274 is evaluated using the definition $\#_n^p = E_{Q^p}\#_n(S)$.
The final term in Equation 274 is found as

$$E_{Q^p}[\sum_t \delta_n(S_t)\ (I_t - \mu_n)'(I_t - \mu_n)] = \tag{282}$$

$$E_{Q^p}[\sum_t E_{Q^p}[\delta_n(S_t)\ (I_t - \mu_n)'(I_t - \mu_n)|S_t]] = \tag{283}$$

$$E_{Q^p}[\sum_t \delta_n(S_t)\ E_{Q^p}[(I_t - \mu_n^p + \mu_n^p - \mu_n)'(I_t - \mu_n^p + \mu_n^p - \mu_n)|S_t]] = \tag{284}$$

$$E_{Q^p}[\sum_t \delta_n(S_t)\ E_{Q^p}[(I_t - \mu_n^p)'(I_t - \mu_n^p)|S_t] + (\mu_n^p - \mu_n)'(\mu_n^p - \mu_n)\ ] = \tag{285}$$

$$E_{Q^p}[\sum_t \delta_n(S_t)\ \Sigma_{S_t}^p + (\mu_n^p - \mu_n)'(\mu_n^p - \mu_n)\ ] = \tag{286}$$

$$\#_n^p\ \Sigma_n^p + \#_n^p\ (\mu_n^p - \mu_n)'(\mu_n^p - \mu_n). \tag{287}$$

For the usual, $c = 1$, case, setting the derivative of the auxiluary function to 0 yields

$$\begin{aligned}&\frac{\partial}{\partial \Sigma_n} \frac{1}{|T|} \sum_{I \in T} E_{Q^p}[\log q(I,S) | i = I] = \\ &\qquad \tilde{\#}_n^{p+1} \Sigma_n + \frac{1}{|T|} \sum_{i \in T} E_{Q^p}[\sum_t \delta_n(S_t)\,(I_t - \tilde{\mu}^{p+1})'(I_t - \tilde{\mu}^{p+1}) | i = I]\end{aligned} \tag{288}$$

so that

$$\tilde{\#}_n^{p+1}\, \tilde{\Sigma}_n^{p+1} = \frac{1}{|T|} \sum_{i \in T} E_{Q^p}[\sum_t \delta_n(S_t)\,(I_t - \tilde{\mu}^{p+1})'(I_t - \tilde{\mu}^{p+1}) | i = I]. \tag{289}$$

Using the above in the derivative of the auxiluary function yields

$$\begin{aligned}&\frac{\partial}{\partial \Sigma_n} [\, \frac{c}{|T|} \sum_{I \in T} E_{Q^p}[\log q(I,S) | i = I] + (1-c)\, E_{Q^p} \log q(I,S)\,] = \\ &\qquad c\, \tilde{\#}_n^{p+1} \Sigma_n - c\, [\, \tilde{\#}_n^{p+1}\, \tilde{\Sigma}_n^{p+1} + (2\, \tilde{\#}_n^{p+1} + 1)(\tilde{\mu}_n^{p+1} - \mu_n^p)'(\tilde{\mu}_n^{p+1} - \mu_n^p)\,] + \\ &\qquad (1-c) \#_n^p\, \Sigma_n - (1-c)[\,(\mu_n^p - \mu_n^{p+1})'(\mu_n^p - \mu_n^{p+1}) + \Sigma_n^p]\end{aligned} \tag{290}$$

The reestimate of the covariance under the modified likelihood criterion is therefore

$$\begin{aligned}\Sigma_n^{p+1} \quad = \quad & \frac{c\, \tilde{\#}_n^{p+1}\, \tilde{\Sigma}_n^{p+1} + (1-c)\, \#_n^p\, \Sigma_n^p}{c\, \tilde{\#}_n^{p+1} + (1-c)\, \#_n^p} \\ & + \frac{c\, (2\, \tilde{\#}_n^{p+1} + 1)(\tilde{\mu}_n^{p+1} - \mu_n^p)'(\tilde{\mu}_n^p - \mu_n^p)}{c\, \tilde{\#}_n^{p+1} + (1-c)\, \#_n^p} + \frac{(1-c)\, \#_n^p\, (\mu_n^p - \mu_n^{p+1})'(\mu_n^p - \mu_n^{p+1})}{c\, \tilde{\#}_n^{p+1} + (1-c)\, \#_n^p}.\end{aligned} \tag{291}$$