Learning to model sequences generated by switching distributions

Yoav Freund
AT&T Bell Labs
600 Mountain Ave.
Murray Hill, NJ, USA

Dana Ron
Computer Science Institute.
Hebrew University
Jerusalem, Israel

Abstract

We study efficient algorithms for solving the following problem, which we call the *switching distributions* learning problem. A sequence $S = \sigma_1\sigma_2\ldots\sigma_n$, over a finite alphabet Σ is generated in the following way. The sequence is a concatenation of K *runs*, each of which is a consecutive subsequence. Each run is generated by independent random draws from a distribution $\vec{p}_i$ over Σ, where $\vec{p}_i$ is an element in a set of distributions $\{\vec{p}_1,\ldots,\vec{p}_N\}$. The learning algorithm is given this sequence and its goal is to find approximations of the distributions $\vec{p}_1,\ldots,\vec{p}_N$, and give an approximate segmentation of the sequence into its constituting runs. We give an efficient algorithm for solving this problem and show conditions under which the algorithm is guaranteed to work with high probability.

1 Introduction

Our work is motivated by a popular model for speech analysis which assumes that the speech signal can be approximately modeled as the concatenation of simple phonetic units. This model, for example, justifies the use of Hidden Markov Models in speech analysis.[1] Learning a model of this type from data involves two interdependent problems. One problem is the *modeling* problem, which is to find good models for the individual elements, or phonemes. The other is the *segmentation* problem, which is to partition the time sequence which represents the speech into short subsequences that correspond to the different phonemes. Clearly, having the exact solution for one of these problems greatly simplifies solving the other. Traditional learning methods used in speech analysis require some amount of speech data to be segmented by hand in order to alleviate the segmentation problem. We are interested in finding learning algorithms that can solve both problems together, using the speech signal alone as the source of training examples. This paper summarizes a first step in addressing the theoretical aspects of this problem.

We define the *switching distributions* learning problem, which is a greatly simplified version of the problem described above. We assume that we are given a sequence $S = \sigma_1\sigma_2\ldots\sigma_M$ over a finite alphabet Σ. The sequence is generated by concatenating K *runs*. Each run is generated by independent random draws from a distribution $\vec{p}_i$ over Σ, where $\vec{p}_i$ is an element in a set of *target* distributions $\vec{p}_1, \ldots, \vec{p}_N$. We assume that the number of distributions, N, is much smaller than the number of runs, K, thus the same distribution is used to generate many different runs.

The modeling problem is to find the N distributions and the segmentation problem is to reconstruct the partitioning of the sequence into runs.

This problem is related to the problem of learning switching concepts studied by Blum and Chalasani [2]. However, in their setup the switching entities are concepts, i.e., mappings from a domain to $\{0,1\}$, while in our setup the switching entities are distributions over a single space.

In this work we give an efficient algorithm for learning switching distributions. We describe several variants of the algorithm, each of which is guaranteed to succeed under slightly different conditions regarding the process which is generating the sequence.

Our algorithm works in the following general way. It starts by finding rough approximations of the distributions $\vec{p}_1, \ldots, \vec{p}_N$. This is done by finding short subsequences of S which, with high probability, are generated mostly by a single distribution. Starting with these approximations of the distributions, the algorithm iterates the following two steps.

1. Using the approximate distributions, the algorithm finds an approximate segmentation of the sequence.
2. From the approximate segmentation, the algorithm finds new estimates of the distributions.

Our analysis shows that if the errors in the initial estimates of the distributions are smaller than a constant factor of the distance between the target distributions, then the number of mistakes in the segmentation is, with high probability, smaller than $O(K\log(NM))$. The error in the re-estimated distributions is $O(\sqrt{(K+D)\log(NM)/M})$. In other words, if the sequence is long enough with respect to the number of runs (and other parameters of the source which we specify later), then even rough initial estimates of the target distributions are sufficient to achieve very accurate estimates within a single iteration of the algorithm.

This approach is similar to that of the Baum-Welch algorithm [1] which is the algorithm commonly used in practice to solve segmentation and recognition problems in speech analysis. The

[1] For an extensive review of speech analysis and the use of Hidden Markov Model see Rabiner and Juang [5].

computational complexity of each iteration of the Baum-Welch algorithm is well understood. However, as far as we know, there are no known bounds on the convergence rate of the algorithm, which determines the overall computation time. This paper is thus a step towards reaching a more complete understanding on the conditions that are required for this type of algorithms to converge rapidly.

The paper is organized as follows. In the Section 2 we give the exact statement of the switching distributions problem. In Section 3 we describe the general algorithm that we use to solve the problem. The details of the algorithm depend on the number of distributions N, and on the amount of information given to the algorithm concerning the different parameters of the problem. In Section 4 we present our main results. Our strongest result is for $N = 2$ and is given in Section 5. In Section 6 we give a general algorithm for $N > 2$, and in Section 7 we describe how to treat unspecified parameters.

2 Description of the Problem

We are interested in the following problem. Let $\Sigma = \{1 \ldots D\}$ be an alphabet of size D. Let $\boldsymbol{\eta} = \eta(1) \ldots \eta(M)$, $\eta(i) \in [1..N]$ be the *target segmentation* sequence containing at most K *runs*, where a *run* is a consecutive subsequence $\eta(i) \ldots \eta(i+s)$ consisting of $s+1$ repetitions of a single index in $[1..N]$. Let $\{\vec{p}_i\}_{j=1}^N$ be the set of *target probability vectors* where for each j, $1 \leq j \leq N$, $\vec{p}_j$ is a D dimensional probability vector defined over Σ. We denote by $\vec{p}_j(\sigma)$ the probability assigned to σ by $\vec{p}_j$ (i.e., the σ coordinate of $\vec{p}_j$).

We assume that a sample sequence $S = \sigma_1 \ldots \sigma_M$ of elements from Σ is generated in the following manner. For each $1 \leq i \leq n$, the element σ_i is chosen independently at random according to the distribution defined by $\vec{p}_{\eta(i)}$. We are interested in algorithms which, given such a sequence, construct a hypothesis $(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N)$, where $\tilde{\boldsymbol{\eta}}$ is a hypothesis segmentation sequence and $\{\tilde{p}_j\}_{j=1}^N$ is a set of distributions. The error of a hypothesis $(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N)$, with respect to the target $(\boldsymbol{\eta}, \{\vec{p}_j\}_{j=1}^N)$, is defined as follows

$$err_d(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{i=1}^{M} d\left(\vec{p}_{\eta(i)}, \tilde{p}_{\tilde{\eta}(i)}\right) \ ,$$

where $d(\cdot,\cdot)$ is a measure of divergence between distribution vectors. We give results with respect to three divergence measures: $d_1(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\|_1$, $d_2(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\|_2$, and $d_3(\vec{x},\vec{y}) = \|\vec{x}-\vec{y}\|_2^2$. We use err_1, err_2, err_3 to denote the errors of a hypothesis with respect to d_1, d_2 and d_3 respectively. As for any distribution vectors $\vec{x}, \vec{y}$, $d_1(\vec{x},\vec{y}) \geq d_2(\vec{x},\vec{y}) \geq d_3(\vec{x},\vec{y})$, err_1 is the most sensitive measure of the error of the hypothesis and err_3 is the least sensitive measure.

Intuitively, a hypothesis with small error is one which defines a sequence of distributions that is very similar to the one which generated the sequence S. If the target distributions $\{\vec{p}_j\}_{j=1}^N$, are all far from each other, then the fact that the error of a hypothesis is small implies that to each target distribution there corresponds a hypothesis distribution which is close to it, and that this hypothesis distribution is matched to it on most of the sequence. This means that a hypothesis which has small error solves both the segmentation and the modelling problems described above.

A learning algorithm for the switching distributions problem receives as input a single sequence S, together with an accuracy parameter $\epsilon > 0$ and a reliability parameter $\delta > 0$. After time polynomial in M, K, N, D, $1/\epsilon$, and $\log(1/\delta)$, the algorithm outputs a hypothesis. We require that there exists a polynomial $q(K, N, D, 1/\epsilon, \log(1/\delta))$, such that if $M \geq q(K, N, D, 1/\epsilon, \log(1/\delta))$, then, with probability at least $1-\delta$ the error of the hypothesis is smaller than ϵ.

3 The general algorithm

In this section we describe an efficient algorithm for solving the switching distributions learning problem. Some elements of the algorithm are left unspecified. These parts are implemented differently for the case of two target distributions ($N = 2$) and for the general case of more than two distributions, and are described in detail in the following sections. The reason for the two different implementations is that we were able to derive better results for the case $N = 2$ by using procedures which exploit the fact that the sequence is generated by no more than two distributions.

General algorithm for learning switching distributions

Input: A sequence $S = \sigma_1, \sigma_2, \ldots, \sigma_M$ and
D - The size of the alphabet, N - The number of unknown distributions, K - the maximal number of runs, and γ - the minimal fraction of the sequence that corresponds to each distribution.

1. **Initialization:** Find initial approximations of the N target distributions: $\tilde{p}_1^1, \ldots, \tilde{p}_N^1$.
2. **Do for** $t = 1, 2, \ldots, T$
   (a) Set the cost vectors $\vec{c}_1^t, \ldots, \vec{c}_N^t$ as functions of the distributions $\tilde{p}_1^t, \ldots, \tilde{p}_N^t$.
   (b) Find the segmentation $\tilde{\boldsymbol{\eta}}^t = \tilde{\eta}^t(1), \ldots, \tilde{\eta}^t(M)$ which minimizes the total cost:
   $$C\left(\{\vec{c}_j^t\}_{j=1}^N, \tilde{\boldsymbol{\eta}}^t, S\right) = \sum_{i=1}^{M} \vec{c}_{\tilde{\eta}^t(i)}^t(\sigma_i)$$
   (c) Calculate the new estimates of the distributions. For $j \in [1, \ldots, N]$ set $\tilde{p}_j^{t+1}$ to the empirical distribution of the elements of S for which $\tilde{\boldsymbol{\eta}}^t(i) = j$.
3. **Output** the segmentation $\tilde{\boldsymbol{\eta}}^T$, and the distributions $\tilde{p}_1^{T+1}, \ldots, \tilde{p}_N^{T+1}$.

Figure 1: The general learning algorithm.

The algorithm is described in Figure 1. It consists of two parts. In the first part the algorithm finds rough approximations of the target distributions. This is implemented, as will be shown in more detail in the following sections, by locating short subsequences that appear to be generated each by a single distribution. The second part of the algorithm is an iterative part in which the approximate distributions are used to generate an approximate segmentation, and this segmentation is then used to re-estimate the distributions. These two steps are repeated T times, for some $T \geq 1$. On iteration t, a *cost vector* $\vec{c}_j^t$ is associated with each approximate distribution. This cost vector is used to calculate a total cost for any hypothesis segmentation of the sequence as defined in step 2b of the algorithm.

Finding the segmentation with K runs that minimizes the total cost for a given set of cost vectors can be performed in time $O(KM^3)$ using dynamic programming as follows. First, one calculates the cost of each consecutive subsequence of S according to each single cost vector. Then, for $k = 2, \ldots, K$, one finds the best segmentation of each consecutive subsequence of S with at most k runs by concatenating subsequences with $\lfloor k/2 \rfloor$ runs with subsequences with $\lceil k/2 \rceil$ runs.

The cost vectors are chosen so that with high probability the segmentation with the lowest total cost does not differ significantly from the target segmentation. More specifically, our goal is to select cost vectors that satisfy the following two properties: (1) The *expected cost* of the target segmentation is smaller than the expected cost of any other segmentation (with at most K runs). (2) With high probability, the segmentation that minimizes the cost on a sample sequence S, has

a small number of segmentation errors. Once the segmentation with the lowest total cost is found, the N probability distributions are re-estimated, and the process is repeated.

The key property of this iterative process, as we shall show in the following sections, is that if the error of the initial estimates of the distributions is smaller than some threshold, then the iterative process increases the accuracy of the models very rapidly. We shall show that this threshold need not depend on the approximation parameter ϵ, but rather is of the order of the smallest distance between any pair of target distributions.

4 Summary of Results

Before we present our results, we add the following notation. Let γ denote the minimal fraction of elements in S corresponding to any single distributions. Namely, if for $j \in [1, \ldots, N]$, n_j is the number of elements in S corresponding to the distribution $\vec{p}_j$, then $\gamma \stackrel{\text{def}}{=} \min_j(n_j)/M$. Let L denote (a lower bound on) the length of the shortest run in S.

We summarize our results in four theorems which correspond to different variants of the algorithm described in Section 3. In the first variant we assume that there are only two target distributions, and that the algorithm receives as input (an upper bound on) K, and (a lower bound on) γ. The error of the algorithm's hypothesis is measured with respect to d_1, which is the most sensitive distance measure we use.

Theorem 1 (Main Theorem for the case of two distributions) *There exists a switching distributions learning algorithm, which is given as input the sequence S together with K and γ, such that if $N = 2$ and if the length M of S satisfies*

$$\frac{M}{\ln M} = \Omega\left(\frac{K \cdot (D + \ln \frac{1}{\delta})}{\epsilon^2 \max(\gamma, \epsilon)}\right) ,$$

then with probability at least $1 - \delta$, the algorithm outputs a hypothesis $(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_1, \tilde{p}_2\})$ such that $err_1(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_1, \tilde{p}_2\}) \leq \epsilon$.

Note that if γ is large with respect to ϵ, then the dependence of $M/\ln M$ on ϵ is $\Omega(1/\epsilon^2)$, and if γ is small then the dependence is $\Omega(1/\epsilon^3)$.

In the second variant, we remove the assumption on N, but we assume that the algorithm is given (an upper bound on) K and (a lower bound on) L. In this case we require that the target distributions are well separated. We measure the separation between the target distributions by $\Delta_2 \stackrel{\text{def}}{=} \min_{j \neq j'} \|\vec{p}_{j'} - \vec{p}_j\|_2$. Our requirements on S are that γ and Δ_2 are polynomially related to ϵ, and that L grows logarithmically with M. The error of the algorithm's hypothesis is measured with respect to d_2.

Theorem 2 (Main Theorem for general N) *There exists a switching distributions learning algorithm, which is given as input the sequence S together with K and L such that if the length M of S satisfies*

$$\frac{M}{\ln M} = \Omega\left(\frac{K \ln N + D + \ln \frac{1}{\delta}}{\gamma \min(\epsilon^2, \Delta_2^3 \epsilon)}\right) ,$$

and the minimum length L of the runs in S satisfies

$$\frac{L}{\ln L} = \Omega\left(\frac{\ln M + D + \ln \frac{1}{\delta}}{\Delta_2^2}\right) ,$$

then with probability at least $1-\delta$*, the algorithm outputs a hypothesis* $(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N)$ *such that* $err_2(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N) \leq \epsilon$.

The third variant of our algorithm needs no input other than S and makes only the assumption that $N \leq 2$. The error of the algorithm's hypothesis is measured with respect to d_3.

Theorem 3 (Main Theorem for two distributions and unspecified parameters) *There exists a switching distributions learning algorithm, such that given the sequence* S*, if* $N = 2$ *and the length* M *of the sequence* S *satisfies*

$$\frac{M}{\ln M} = \Omega\left(\frac{K \cdot (D + \ln \frac{1}{\delta})}{\epsilon^2 \max(\gamma, \epsilon)}\right) ,$$

then with probability at least $1-\delta$*, the algorithm outputs a hypothesis* $(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_1, \tilde{p}_2\})$ *such that* $err_3(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_1, \tilde{p}_2\}) \leq \epsilon$.

In our fourth variant we do not assume that the algorithm is given any of the parameters of the problem. However we still require the existence of a lower bound on L which grows logarithmically with M, and that γ and Δ_2 are polynomially related to ϵ. The error of the algorithm's hypothesis is measured with respect to d_3.

Theorem 4 (Main Theorem for general N and unspecified parameters) *There exists a switching distributions learning algorithm, such that given the sequence* S*, if the length* M *of the* S *satisfies*

$$\frac{M}{\ln M} = \Omega\left(\frac{K \ln N + D + \ln \frac{1}{\delta}}{\gamma \min(\epsilon^2, \Delta_2^3 \epsilon)}\right) ,$$

and the minimum length L *of the runs in* S *satisfies*

$$\frac{L}{\ln L} = \Omega\left(\frac{\ln M + D + \ln \frac{1}{\delta}}{\Delta_2^2}\right) ,$$

then with probability at least $1-\delta$*, the algorithm outputs a hypothesis* $(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N)$ *such that* $err_3(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N) \leq \epsilon$.

The theorems presented above result from an analysis of a *single* iteration of step 2 of the learning algorithm. It is natural to ask whether by increasing the number of iterations, we could significantly weaken the requirements on M. Our analysis does not support such a claim, and we shall later discuss this question briefly.

5 The Case of Two Distributions

In this section we consider the case in which there are two target probability distributions $\vec{p}_1$ and $\vec{p}_2$ over an alphabet Σ of size D. The L_1 distance between the two vectors, $\|\vec{p}_1 - \vec{p}_2\|_1$, plays an important role in our analysis, and it is denoted for short by Δ_1. We assume that the algorithm knows (an upper bound on) K, the number of switches in the sequence, and (a lower bound on) γ, the minimum between the fraction of elements in the target sequence, $\boldsymbol{\eta}$, generated by $\vec{p}_1$, and the fraction generated by $\vec{p}_2$. In Section 7 we give bounds on the additional error incurred when we remove this assumption

In Figure 2 we describe how we get initial approximations $\tilde{p}_1^0$ and $\tilde{p}_2^0$ of $\vec{p}_1$ and $\vec{p}_2$ respectively, and we define the pair of cost vectors $\vec{c}_1^t$ and $\vec{c}_2^t$, given approximations $\tilde{p}_1^t$ and $\tilde{p}_2^t$ of $\vec{p}_1$ and $\vec{p}_2$.

The initial approximation procedure is based on the following two facts. The first is that by definition of K and γ, both for $\vec{p}_1$ and for $\vec{p}_2$ there exists a subsequence of S of length $\gamma M/K$ that was generated *solely* according to that probability distribution. The second fact is that, in expectation, the distance between pairs of empirical distributions defined based on pairs of subsequences is maximized when one of the subsequences was generated according to $\vec{p}_1$ and the second according to $\vec{p}_2$. We are thus able to show that the pair of distributions chosen are good initial approximations of $\vec{p}_1$ and $\vec{p}_2$.

As we show in the analysis, the choice of cost vectors given in Figure 2 has the two properties described in the end of Section 3. We cannot show that this choice is optimal, however, it can be verified that if we substitute in our analysis *any* other pair of cost vectors, then the bound we get on the resulting segmentation error is at least a fourth times the bound given in Lemma 5.2

Initial estimates for two distributions.

1. Set $\ell = \gamma M/K$.
   (If L, the minimum length of any run in S, is known, set $\ell = \max(\gamma M/K, L)$).
2. For each $i \in [1, \ldots, M-\ell+1]$ and each $\sigma \in [1, \ldots, D]$, let $f_i(\sigma)$ be the fraction of the elements in $\sigma_i \sigma_{i+1} \ldots \sigma_{i+\ell-1}$ which are equal to σ. Let $\vec{f}_i$ denote the vector $\langle f_i(1), f_i(2), \ldots, f_i(D) \rangle$.
3. Find the pair of indices $1 \leq i_1 < i_2 \leq M - \ell + 1$, for which $\|\vec{f}_{i_1} - \vec{f}_{i_2}\|_1$ is maximized.
4. Set $\tilde{p}_1^0 = f_{i_1}$ and $\tilde{p}_2^0 = f_{i_2}$.

Choice of cost vectors for two distributions

1. Given estimates $\tilde{p}_1^t$ and $\tilde{p}_2^t$, let $\Sigma_1 \stackrel{\text{def}}{=} \{\sigma \in \Sigma \ : \ \tilde{p}_1^t(\sigma) > \tilde{p}_2^t(\sigma)\}$, and for $j \in \{1, 2\}$, let $\tilde{p}_j^t(\Sigma_1) \stackrel{\text{def}}{=} \sum_{\sigma \in \Sigma} \tilde{p}_j^t(\sigma)$.
2. Let $\vec{d}^t$ be defined as follows:
$$\forall \sigma \in \Sigma_1, \ \vec{d}^t(\sigma) = \frac{\tilde{p}_1^t(\Sigma_1) + \tilde{p}_2^t(\Sigma_1)}{2} - 1 \ , \quad \forall \sigma \in \Sigma - \Sigma_1, \ \vec{d}^t(\sigma) = \frac{\tilde{p}_1^t(\Sigma_1) + \tilde{p}_2^t(\Sigma_1)}{2} \ .$$
3. Set $\vec{c}_1^t = \vec{d}^t, \ \ \vec{c}_2^t = \vec{0}$.

Figure 2: The initialization procedure and the choice of the cost vectors for the case of two distributions.

Our main result for the case of two distributions is stated in Theorem 1 and its proof is based on the following three lemmas. All proofs are given in Appendix B. We use the following notation. For a given segmentation $\boldsymbol{\eta}'$, and for $j, j' \in \{1, 2\}$, let $n_{j,j'}(\boldsymbol{\eta}')$ denote the number of symbols in S that belong to a run corresponding to $\vec{p}_j$ in $\boldsymbol{\eta}$, and to a run corresponding to $\vec{p}_{j'}$ in $\boldsymbol{\eta}'$. Let $n_{j'}(\boldsymbol{\eta}') = n_{1,j'}(\boldsymbol{\eta}') + n_{2,j'}(\boldsymbol{\eta}')$.

First, we show that if M, the length of the sequence, is large enough, then the initial estimates $\tilde{p}_1^0$ and $\tilde{p}_2^0$ of $\vec{p}_1$ and $\vec{p}_2$ are guaranteed to have small error.

Lemma 5.1 (Initialization error for two distributions) *If for some ρ, $0 < \rho \leq \Delta_1/2$, the length M of the sequence S satisfies*

$$\frac{M}{\ln M} \geq \frac{2 \cdot K(D + \ln \frac{1}{\delta})}{\rho^2 \cdot \gamma} \ ,$$

then with probability at least $1-\delta$*, there exists a one-to-one mapping* $\phi : \{1,2\} \to \{1,2\}$ *such that for* $j \in \{1,2\}$, $\|\tilde{p}^0_{\phi(j)} - \vec{p}_j\|_1 \leq 3\rho$.

Secondly, we show that if the errors of the estimates $\tilde{p}^t_1$ and $\tilde{p}^t_2$ are not too large then, with high probability, only a small number of mistakes exist in the segmentation with the lowest cost, defined using the corresponding cost vectors.

Lemma 5.2 (Segmentation error for two distributions) *Let*

$$\rho = \min_{\phi} \max_{j \in \{1,2\}} \|\tilde{p}^t_{\phi(j)} - \vec{p}_j\|_1 \ ,$$

where ϕ *ranges over the (two) one-to-one mappings from* $\{1,2\}$ *to* $\{1,2\}$*. If* $\rho < \frac{1}{6}\Delta_1$*, then with probability at least* $1-\delta$

$$\frac{n_{1,2}(\tilde{\boldsymbol{\eta}}^t) + n_{2,1}(\tilde{\boldsymbol{\eta}}^t)}{M} < \frac{32(\ln(1/\delta) + K\ln(2M))}{(\Delta_1 - 6\rho)^2 M} \ ,$$

where $\tilde{\boldsymbol{\eta}}^t$ *is the t'th hypothesis segmentation.*

In the third lemma we show that if the segmentation $\tilde{\boldsymbol{\eta}}^t$ has a small number of errors, then good new estimates of $\vec{p}_1$ and $\vec{p}_2$ can be computed using $\tilde{\boldsymbol{\eta}}^t$.

Lemma 5.3 (Reevaluation error for two distributions) *Let*

$$\beta = \max(n_{1,2}(\tilde{\boldsymbol{\eta}}^t), n_{2,1}(\tilde{\boldsymbol{\eta}}^t))/M \ .$$

If $\beta < \gamma$*, then with probability at least* $1-\delta$*, there exists a one-to-one mapping* $\phi : \{1,2\} \to \{1,2\}$ *such that for* $j \in \{1,2\}$

$$\|\tilde{p}^{t+1}_{\phi(j)} - \vec{p}_j\|_1 \leq \frac{\beta}{\gamma - \beta}\Delta_1 + \sqrt{\frac{2(K\ln(2M) + D\ln(M+1) + \ln(1/\delta)}{(\gamma-\beta)M}} \ .$$

Theorem 1 summarizes the convergence properties of the algorithm when step 3 of the algorithm is executed once. It is interesting to consider the convergence properties that are implied by Lemmas 5.2 and 5.3 in a little more detail.

According to Lemma 5.2, if the errors of the estimates of $\vec{p}_1$ and $\vec{p}_2$ are smaller than a constant fraction of $\Delta_1 = \|\vec{p}_1 - \vec{p}_2\|_1$, then the number of segmentation errors can be decreased to an arbitrarily small fraction of M by increasing M, which in turn decreases the error of the new estimates of the distributions that result from the segmentation. Intuitively, this means that there is a "basin of attraction" of the estimates of the distributions, whose "size" is proportional to the distance Δ_1. If the algorithm starts with an estimate of the distribution that is within this basin of attraction then the estimate it gives in the following iteration is very accurate.

On the other hand, the bounds given in Lemma 5.2 do not improve significantly even if there is no error in the estimates. This is because there are always segmentation errors that are the result of the randomness of the sequence S. Thus, according to our analysis, there is no advantage in iterating the algorithm more than a single time. It is unclear whether a better analysis of the algorithm would show that such an advantage exists.

6 The Case of a General Number of Distributions

In this section we consider the case in which there are $N > 2$ target probability distributions $\vec{p}_1, \ldots, \vec{p}_N$. In this case the problem of finding good cost vectors to be associated with the different distributions is more complicated. We have to choose N cost vectors, one per distribution, but have to satisfy $\binom{N}{2}$ sets of requirements, because each *pair* of distributions has to be distinguished well by the corresponding pair of cost vectors. The choice of the cost vectors that we have found is described in Figure 3. This choice is a generalization of the squared loss in the binary case ($D = 2$), and allows us to bound the error of the algorithm according to err_2 (which is weaker, i.e., less sensitive, than err_1).

The initialization procedure that is used in the two-distribution case cannot be applied to the general case. The initialization procedure that we suggest requires that the segmentation sequence $\boldsymbol{\eta}$ is such that *all* of the runs in $\boldsymbol{\eta}$ are longer than some integer parameter L. This allows the algorithm to assume that in each segment of S of length L there is at most one switch between distributions. Consequently, it can identify if both parts were generated almost solely by the same probability distribution. The initialization procedure is described in Figure 3.

We assume that the algorithm receives the parameters K and L as input. This assumption is removed (at some additional cost) in the next section.

The Proof of Theorem 2 is very similar to the proof of Theorem 1 and follows from the lemmas given below whose proofs are provided in Appendix C.

Initial estimates of the distributions. (general case)

1. Set $\ell = \frac{1}{2}L$.
2. Set $\theta = 2\sqrt{\frac{2(\ln M + D\ln(\ell+1) + \ln\frac{1}{\delta})}{\ell}}$.
3. For each $i \in 1 \ldots M-\ell+1$ and each $\sigma \in [1, \ldots, D]$, let $f_i(\sigma)$ be the fraction of the elements in $\sigma_i, \sigma_{i+1}, \ldots, \sigma_{i+\ell-1}$ which are equal to σ. Let $\vec{f}_i$ denote the vector $\langle f_i(1), f_i(2), \ldots, f_i(D)\rangle$.
4. Let A be the set of vectors $(\vec{f}_i + \vec{f}_{i+\ell})/2$ for all $1 \le i \le M-\ell$ such that $\|\vec{f}_i - \vec{f}_{i+\ell}\|_2 \le \theta$.
5. Start with B as an empty set and repeat the following until A becomes empty:
   - Select an element $\vec{f}$ from A and add it to B.
   - Remove from A all elements $\vec{g}$ such that $\|\vec{f} - \vec{g}\|_2 \le 3\theta$.
6. Output the set B as the set of distribution estimates.

Choice of cost vectors. (general case)

- With each estimate $\tilde{p}_j^t$ associate the cost vector

$$\vec{c}_j^t(\sigma) \quad = \quad \sum_{\sigma' \neq \sigma} (\tilde{p}_j^t(\sigma'))^2 + (1 - \tilde{p}_j^t(\sigma))^2 = \sum_{\sigma'=1}^{D} (\tilde{p}_j^t(\sigma'))^2 + 1 - 2\tilde{p}_j^t(\sigma)$$

Figure 3: The initialization procedure and the choice of the cost vectors for the general case

Lemma 6.1 (Initialization error for general N) *If for some $\rho < \Delta/14$, the minimal length L of runs in S satisfies*

$$\frac{L}{\ln L} \ge \frac{4\left(\ln M + D + \ln\frac{1}{\delta}\right)}{\rho^2} ,$$

then, with probability at least $1-\delta$*, the initialization procedure described in Figure 3 generates a set of estimates* $B = \{\tilde{p}_1^0, \ldots, \tilde{p}_N^0\}$ *which approximates* $\{\vec{p}_1, \ldots, \vec{p}_N\}$ *in the following sense. There exists a one-to-one mapping* ϕ *from* $1..N$ *to* $1..N$ *such that for all* $1 \leq j \leq N$

$$\|\vec{p}_j - \tilde{p}_{\phi(j)}^0\|_2 \leq 6\rho$$

Lemma 6.2 (Segmentation error for general N) *Let*

$$\rho = \min_{\phi} \max_{j \in [1,\ldots,N]} \|\tilde{p}_{\phi(j)}^t - \vec{p}_j\|_1 \ ,$$

where ϕ *ranges over all one-to-one mappings from* $[1, \ldots, N]$ *to* $[1, \ldots, N]$*. If* $\rho < \Delta/14$*, then with probability at least* $1-\delta$ *the segmentation* $\tilde{\boldsymbol{\eta}}^t$ *that minimizes the total cost satisfies*

$$\frac{\sum_{j \neq j'} n_{j,j'}(\tilde{\boldsymbol{\eta}}^t)}{M} \leq \frac{8(\ln \frac{1}{\delta} + K \ln(NM))}{M(\Delta - 2\rho)^4}$$

Lemma 6.3 (Reevaluation error for general N) *Let*

$$\beta = \frac{\max_{j \neq j'}(n_{j,j'}(\tilde{\boldsymbol{\eta}}^t))}{M} \ .$$

If $\beta < \gamma$*, then with probability at least* $1-\delta$*, there exists a one-to-one mapping* $\phi : [1, \ldots, N] \to [1, \ldots, N]$*, such that for every* $j \in [1, \ldots, N]$

$$\|\tilde{p}_{\phi(j)}^{t+1} - \vec{p}_j\|_2 \leq \frac{\beta}{\gamma - \beta}\Delta + \sqrt{\frac{2(K \ln(NM) + D \ln(M+1) + \ln(1/\delta)}{(\gamma - \beta)M}} \ .$$

7 Treating Unspecified parameters

So far we have assumed that our learning algorithms receive as input parameters that describe properties of the sequence S. In the two distribution case the algorithm receives as input γ and K, and in the multiple distribution case the algorithm receives K and L. In this section we show that these assumptions can be removed, with some increase in the error, if we measure the error of the hypothesis with err_3. Recall that err_3 is always smaller than err_1 and err_2, thus, the bounds we previously got for cases where the algorithm receives additional input, can be used here.

The idea is to try all possible settings of the unknown parameters, and then select the best resulting hypothesis. While the different algorithms receive, as input different subsets of γ, K and L, the only variable that is controlled by these parameters is ℓ, the length of the segments that are used for initialization. As there are at most M possible settings for ℓ, we need to run the algorithm M times and then compare the hypotheses.

The different hypotheses are compared in terms of the total cost defined in Figure 3 and the hypothesis with the minimal total cost is selected.

More formally, suppose that there are m hypotheses $\left\{(\tilde{\boldsymbol{\eta}}_r, \{\tilde{p}_{r,j}\}_{j=1}^N)\right\}_{r=1}^m$. For each $1 \leq r \leq m$ we calculate the total cost $C\left(\{\vec{c}_{r,j}\}_{j=1}^N, \tilde{\boldsymbol{\eta}}_r, S\right)$, according to the cost vectors $\{\vec{c}_{r,j}\}_{j=1}^N$ as defined in Figure 3 for the case of more than two distributions. We then select the hypothesis with the minimal total cost as our final hypothesis.

The following lemma shows that the error of the selected hypothesis is, with high probability, only slightly larger than that of the best hypothesis in the set. Its proof is given in Appendix D.

Lemma 7.1 *Let S be a sequence generated according to the segmentation $\boldsymbol{\eta}$ and the distributions $\vec{p}_1, \ldots, \vec{p}_N$. Let $\left\{(\tilde{\boldsymbol{\eta}}_r, \{\tilde{p}_{r,j}\}_{j=1}^N)\right\}_{r=1}^m$ be a set of hypotheses and let $u = \mathrm{argmin}_r C(\{\vec{c}_{r,j}\}_{j=1}^N, \tilde{\boldsymbol{\eta}}_r, S)$. Then, with probability at least $1-\delta$ with respect to the distribution of S*

$$err_3(\tilde{\boldsymbol{\eta}}_u, \{\tilde{p}_{u,j}\}_{j=1}^N) \leq \min_{r=1..m} err_3(\tilde{\boldsymbol{\eta}}_r, \{\tilde{p}_{r,j}\}_{j=1}^N) + \sqrt{\frac{2\ln m + \ln(1/\delta) + K\ln N + K\ln M}{M}}$$

Applying Lemma 7.1 to Theorem 1 we get Theorem 3. In this case we add to the M candidate hypotheses that correspond to the choices of ℓ the hypothesis $(\boldsymbol{\eta}_{single}, \{\tilde{p}\})$, where $\boldsymbol{\eta}_{single}$ consists of a single run, and $\tilde{p}$ is the empirical distribution of the sequence. This hypothesis is accurate if either γ or Δ_1 are smaller than $\epsilon/8$, and thus we can remove any assumption on γ and Δ_1. Applying Lemma 7.1 to Theorem 2 we get Theorem 4. In this case we still need to assume a lower bound on L in order to assure that at least one of the executions of the algorithm succeeds.

In both cases the requirement on M that we have to make in order that the extra error be smaller than ϵ, is $M = \Omega\left(1/\epsilon^2(K\ln N + \ln(1/\delta))\right)$. This requirement is subsumed by the requirements in Theorems 1 and 2.

8 Acknowledgements

The work described here is motivated and closely related to current work on the analysis of speech that is carried out in collaboration with Yoram Singer.

References

[1] Leonard E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, 73:360–363, 1967.

[2] Avrim Blum and Prasad Chalasani. Learning switching concepts. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 231–242, July 1992.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[4] Nick Littlestone. Notes on the derivation of chernoff-type bounds for sums of random variables. Unpublished Manuscript, 1990.

[5] L. R. Rabiner and B. H. Juang. An introduction to hidden markov models. *IEEE ASSP Magazine*, 3(1):4–16, January 1986.

A Useful Inequalities

In the proofs of our theorems and lemmas appearing in the following appendices we apply several well known inequalities that are given here as lemmas. The first is a Chernoff/Hoeffding type bound, derived by Littlestone [4], and the second is due to Sanov ([3],page 292).

Lemma A.1 *For $m > 0$, let $X_1, X_2, ...X_m$ be m independent random variables where $a_i \leq X_i \leq b_i$. Let $p = \sum_i E[X_i]/m$. Then, for $w > 0$,*

$$Pr[\sum_{i=1}^{m} X_i \geq pm + w] \leq exp\left(-\frac{2w^2}{\sum_{i=1}^{m}(b_i - a_i)^2}\right) .$$

Lemma A.2 (Sanov's Inequality) *For an alphabet Σ of size D, let $\vec{p}$ be a D dimensional probability vector defined over Σ. Let T be a random sample of size m generated according to $\vec{p}$, and let $\vec{X}_T$, the* type *of T, be a D dimensional probability vector defined as follows: the d'th coordinate of $\vec{X}_T$ is the relative frequency of the symbol d in S. Then, for $\alpha > 0$,*

$$Pr[D_{KL}(\vec{X}_T||\vec{p}) \geq \alpha] \leq (m+1)^D 2^{-\alpha m} ,$$

where $D_{KL}[\vec{X}_T||\vec{p}]$ is the **Kullback Leibler (KL) divergence** *between the distributions and is defined as follows:*

$$D_{KL}[\vec{X}_T||\vec{p}] \stackrel{\text{def}}{=} \sum_{\sigma \in \Sigma} \log \vec{p}(\sigma) \frac{\vec{X}_T(\sigma)}{\vec{p}(\sigma)} .$$

One more useful inequality is the following:

Lemma A.3 *Let $\vec{p}_1$ and $\vec{p}_2$ be two probability vectors, then:*

$$D_{KL}(\vec{p}_1||\vec{p}_2) \geq \frac{1}{2\ln 2}\|\vec{p}_1 - \vec{p}_2\|_1^2 \geq \frac{1}{2\ln 2}\|\vec{p}_1 - \vec{p}_2\|_2^2$$

B Proofs for Section 5

Proof of Theorem 1: In what follows we assume that both γ and Δ are greater than $\epsilon/8$. If $\gamma < \epsilon/8$, it is not hard to verify that under our assumption on the size of M, with probability at least $1 - \delta$ the hypothesis which constitutes of a single run, together with the corresponding probability distribution (defined based on the complete sequence), has error bounded by ϵ. This is also true if $\Delta < \epsilon/8$, but since we do not assume that the algorithm known Δ, we cannot simply output the single run hypothesis. However, it is easily verified that in this case, with high probability, *any* hypothesis segmentation $\boldsymbol{\eta}'$, for which $\min(n_1(\boldsymbol{\eta}'), n_2(\boldsymbol{\eta}')) \geq \gamma n/2$ (together with the corresponding probability distributions), has error at most ϵ. Thus we need only add a simple test at the end of our algorithm, in which if $\min(n_1(\tilde{\boldsymbol{\eta}}^0), n_2(\tilde{\boldsymbol{\eta}}^0)) < \gamma n/2$, then we set $\tilde{\boldsymbol{\eta}}^0$ to be the single run segmentation. It follows from the rest of this proof that such an event does not occur if $\Delta \geq \epsilon/8$.

We next show that with probability at least $1 - \delta$, there exist a mapping ϕ such that

$$\max_{j \in \{1,2\}} \|\tilde{p}^1_{\phi(j)} - \vec{p}_j\|_1 \leq \epsilon/8 , \quad \text{and} \quad \max(n_{1,2}(\tilde{\boldsymbol{\eta}}^0), n_{2,1}(\tilde{\boldsymbol{\eta}}^0))/M \leq \epsilon/(4\Delta) .$$

It follows that the total error is at most

$$(\epsilon/8)(1 - \min(1, \epsilon/(4\Delta))) + (\Delta + \epsilon/4)\min(1, \epsilon/(4\Delta)) .$$

The first term, whose source is the approximation error of the probability distributions, is clearly always less than $\epsilon/8$. The source of the second term is the segmentation error and it can be bounded

by $\epsilon/2$ as follows. If $\Delta \geq \epsilon/4$, then $\Delta + \epsilon/4 \leq 2\Delta$ and we get a bound of $\epsilon/2$ when multiplying by $\min(1, \epsilon/(4\Delta))$. If $\Delta < \epsilon/4$, then $\min(1, \epsilon/(4\Delta)) = 1$, and $\Delta + \epsilon/4 < \epsilon/2$ as well.

In order to get these desired bounds, we "work our way backwards", starting from the reevaluation error of $\tilde{p}^1_{\phi(j)}$, through the segmentation error of $\tilde{\boldsymbol{\eta}}^0$, and to the error of the initial approximation of the probability distributions.

In the bound on the reevaluation error given in Lemma 5.3 we have two terms. If $\beta \leq \gamma/2$, then our requirement on M gives us a bound on the second term which is less than $\epsilon/2$. We next bound the first term by $\epsilon/2$ as well. If $\Delta \leq \epsilon$, then it suffices that $\beta \leq \gamma/3$. Otherwise it suffices to require that $\beta \leq \gamma\epsilon/(2\Delta)$ which is also at most the bound needed for the segmentation error. In both cases, if $\rho \leq \Delta/12$, the inequality in Lemma 5.2, together with our requirement on M, give us the bound we want. Finally, it can easily be verified that our requirement on M gives us the desired bound on ρ when applying Lemma 5.1. ■

Proof of Lemma 5.1: For a given index $i \in [1, \ldots, M - \ell + 1]$, let $S_{i,i+\ell-1} \stackrel{\text{def}}{=} \sigma_i \ldots \sigma_{i+\ell-1}$, and let α_i be such that the fraction of symbols in $S_{i,i+l-1}$ that were generated by $\vec{p}_1$ and $\vec{p}_2$ are $(1-\alpha)$ and α respectively. We say that a vector $\vec{f}_i$ is *pure* if either $\alpha_i = 0$ or $\alpha_i = 1$. For every pair of indices $1 \leq i_1 < i_2 \leq n - \ell + 1$, let $e_{i_1 i_2} = \|\vec{f}_{i_1} - \vec{f}_{i_2}\|_1$. Clearly, the expected value of $e_{i_1 i_2}$ is maximized for a pair of pure vectors $\vec{f}_{i_1}$ and $\vec{f}_{i_2}$ satisfying $\alpha_{i_1} = 0$ (1) and $\alpha_{i_2} = 1$ (0).

We shall show that for some $\rho < \Delta/2$, (which we set subsequently),

$$\|\vec{f}_i - ((1-\alpha_i)\vec{p}_1 + \alpha_i \vec{p}_2)\|_1 \leq \rho \ , \tag{1}$$

for every i. Let $\vec{f}_{i_a}$ and $\vec{f}_{i_b}$ be the pair of vectors for which $e_{i_a i_b}$ is maximized. Without loss of generality we assume that $\alpha_{i_a} < (1 - \alpha_{i_b})$. Based on our choice of ℓ, there must exist a pair of pure vectors $\vec{f}_{i_1}$ and $\vec{f}_{i_2}$ having $\alpha_{i_1} = 0$, and $\alpha_{i_2} = 1$. Since for this pure pair $e_{i_1 i_2} \geq \Delta - 2\rho$, for the maximizing pair, $\vec{f}_{i_a}$ and $\vec{f}_{i_a}$, $e_{i_a i_b} \geq \Delta - 2\rho$ as well. It follows that $\|\vec{f}_{i_a} - \vec{p}_1\|_1 \leq 3\rho$ and $\|\vec{f}_{i_b} - \vec{p}_2\|_1 \leq 3\rho$.

It remains to show that the selection of ρ assumed in Equation 1 exists. Applying Inequality A.2 and using the bound on the KL divergence given in Inequality A.3, we have that for every $\vec{f}_i$,

$$Pr_S\left[\left\|\vec{f}_i - ((1-\alpha_i)\vec{p}_1 + \alpha_i\vec{p}_2\right\|_1 > \rho\right] < (\ell+1)^D \exp\left(-\frac{1}{2}\rho^2\ell\right) \ . \tag{2}$$

There are at most M such vectors, and hence, by setting M as in the statement of the lemma, we find that, with probability at least $1 - \delta$, $\|\vec{f}_i - ((1-\alpha_i)\vec{p}_1 + \alpha_i\vec{p}_2)\|_1 \leq \rho$ for every i as required. ■.

Proof of Lemma 5.2: By our definition of the cost vectors $\vec{c}^t_1$ and $\vec{c}^t_2$, we have that for any given segmentation $\boldsymbol{\eta}'$,

$$E_S\left[C\left(\{\vec{c}^t_1, \vec{c}^t_2\}, \boldsymbol{\eta}, S\right) - C\left(\{\vec{c}^t_1, \vec{c}^t_2\}, \boldsymbol{\eta}', S\right)\right] = n_{1,2}(\boldsymbol{\eta}')\vec{p}_1 \cdot \vec{d}^t - n_{2,1}(\boldsymbol{\eta}')\vec{p}_2 \cdot \vec{d}^t \ , \tag{3}$$

where $\vec{d}^t = \vec{c}^t_1 - \vec{c}^t_2$ is as defined in Figure 2. We first verify that $\vec{p}_1 \cdot \vec{d}^t < 0$ and $\vec{p}_2 \cdot \vec{d}^t > 0$, and hence for every segmentation $\boldsymbol{\eta}' \neq \boldsymbol{\eta}$, $E_S\left[C\left(\{\vec{c}^t_1, \vec{c}^t_2\}, \boldsymbol{\eta}, S\right)\right] < E_S\left[C\left(\{\vec{c}^t_1, \vec{c}^t_2\}, \boldsymbol{\eta}', S\right)\right]$. For $j \in \{1, 2\}$, let $\vec{e}^t_j = \tilde{p}^t_j - \vec{p}_j$, where we assume that $\phi(j) = j$, and $\|\vec{e}^t_j\|_1 \leq \rho (< \frac{1}{6}\|\vec{p}_1 - \vec{p}_2\|_1)$. Then,

$$\begin{aligned}
\vec{p}_1 \cdot \vec{d}^t &= \tilde{p}^t_1 \cdot \vec{d}^t - \vec{e}^t_1 \cdot \vec{d}^t & (4)\\
&= \sum_{\sigma \in \Sigma_1} \tilde{p}^t_1(\sigma)\left(\frac{\tilde{p}^t_1(\Sigma_1) + \tilde{p}^t_2(\Sigma_1)}{2} - 1\right) + \sum_{\sigma \in \Sigma - \Sigma_1} \tilde{p}^t_1(\sigma)\frac{\tilde{p}^t_1(\Sigma_1) + \tilde{p}^t_2(\Sigma_1)}{2} - \vec{e}^t_1 \cdot \vec{d}^t & (5)
\end{aligned}$$

$$
\begin{aligned}
&= \tilde{p}_1^t(\Sigma_1)\left(\frac{\tilde{p}_1^t(\Sigma_1)+\tilde{p}_2^t(\Sigma_1)}{2}-1\right)+\left(1-\tilde{p}_1^t(\Sigma_1)\right)\frac{\tilde{p}_1^t(\Sigma_1)+\tilde{p}_2^t(\Sigma_1)}{2}-\vec{e}_1^t\cdot\vec{d}^t && (6)\\
&= \frac{\tilde{p}_2^t(\Sigma_1)-\tilde{p}_1^t(\Sigma_1)}{2}-\vec{e}_1^t\cdot\vec{d}^t && (7)\\
&\leq -\frac{1}{4}\|\tilde{p}_1^t-\tilde{p}_2^t\|_1+\rho && (8)\\
&\leq -\frac{1}{4}\Delta+\frac{3}{2}\rho && (9)\\
&< 0\ . && (10)
\end{aligned}
$$

Similarly

$$\vec{p}_2\cdot\vec{d}^t \geq \frac{1}{4}\Delta-\frac{3}{2}\rho > 0\ . \tag{11}$$

It remains to bound the probability that $C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta}',S\right) < C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta},S\right)$ for a segmentation $\boldsymbol{\eta}'$ such that $n_{1,2}(\boldsymbol{\eta}')+n_{2,1}(\boldsymbol{\eta}') \geq \frac{8(\ln(1/\delta)+K\ln(2M))}{(\Delta-6\rho)^2}$. The difference in the total cost, $C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta}',S\right)-C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta},S\right)$, is a sum of the contributions of the elements of S for which $\eta(i)\neq\tilde{\eta}(i)$. These contributions are independent and the range os the values of each one is exactly 1. Thus the expected value of the difference is

$$E_S\left[C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta},S\right)-C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta}',S\right)\right] = n_{1,2}(\boldsymbol{\eta}')\vec{p}_1\cdot\vec{d}^t-n_{2,1}(\boldsymbol{\eta}')\vec{p}_2\cdot\vec{d}^t\ , \tag{12}$$

and the probability that the total cost of $\tilde{\boldsymbol{\eta}}$ is smaller than that of $\boldsymbol{\eta}$ is upper bounded by

$$
\begin{aligned}
Pr_S & \left[C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta}',S\right) < C\left(\{\vec{c}_1^t,\vec{c}_2^t\},\boldsymbol{\eta},S\right)\right] && (13)\\
&\leq exp\left(-\frac{\left(n_{1,2}\vec{p}_1\cdot\vec{d}^t-n_{2,1}\vec{p}_2\cdot\vec{d}^t\right)^2}{2(n_{1,2}+n_{2,1})}\right) && (14)\\
&= exp\left(-\frac{n_{1,2}+n_{2,1}}{2}\left(\frac{n_{1,2}}{n_{1,2}+n_{2,1}}\vec{p}_1\cdot\vec{d}^t-\frac{n_{2,1}}{n_{1,2}+n_{2,1}}\vec{p}_2\cdot\vec{d}^t\right)^2\right)\ . && (15)
\end{aligned}
$$

From our assumption on the size of $n_{1,2}+n_{2,1}$, and by substituting our bounds on $\vec{p}_1\vec{d}^t$ and $\vec{p}_2\vec{d}^t$, the probability above is bounded by $\delta/((2M)^K)$. Since the total number of possible segmentation containing at most K runs, is at most $M^K\cdot 2^K$, we get the statement of the lemma. desired bound. ■

Proof of Lemma 5.3: For a given segmentation $\boldsymbol{\eta}'$, for $j\in\{1,2\}$, let $\vec{p}_j^{\boldsymbol{\eta}'}$ be the empirical distribution of the elements σ_i in S for which $\boldsymbol{\eta}'(i)=j$. We define the deviation, $e_j(\boldsymbol{\eta}')$, of $\vec{p}_j^{\boldsymbol{\eta}'}$ from its mean as follows:

$$e_j(\boldsymbol{\eta}') \stackrel{\text{def}}{=} \left\|\vec{p}_j^{\boldsymbol{\eta}'}-\left(\frac{n_{1,j}(\boldsymbol{\eta}')\vec{p}_1+n_{2,j}(\boldsymbol{\eta}')\vec{p}_2}{n_{1,j}(\boldsymbol{\eta}')+n_{2,j}(\boldsymbol{\eta}')}\right)\right\|_1\ . \tag{16}$$

We shall show that there exists $\rho<1$ (which is set subsequently) such that for every $\boldsymbol{\eta}'$, having $\max(n_{1,2}(\boldsymbol{\eta}'),n_{2,1}(\boldsymbol{\eta}'))/M\leq\beta$, $e_1(\boldsymbol{\eta}'),e_2(\boldsymbol{\eta}')\leq\rho$. Thus, in particular, $e_1(\tilde{\boldsymbol{\eta}}^t),e_2(\tilde{\boldsymbol{\eta}}^t)\leq\rho$ and

$$\|\tilde{p}_1^t-\vec{p}_1\|_1 \leq \|\frac{n_{1,1}(\tilde{\boldsymbol{\eta}}^t)\vec{p}_1+n_{2,1}(\tilde{\boldsymbol{\eta}}^t)\vec{p}_2}{n_{2,1}(\tilde{\boldsymbol{\eta}}^t)+n_{1,1}(\tilde{\boldsymbol{\eta}}^t)}-\vec{p}_1\|_1+\rho \tag{17}$$

$$= \frac{n_{2,1}(\tilde{\boldsymbol{\eta}}^t)}{n_{1,1}(\tilde{\boldsymbol{\eta}}^t) + n_{2,1}(\tilde{\boldsymbol{\eta}}^t)}\Delta + \rho \tag{18}$$

$$\leq \frac{n_{2,1}(\tilde{\boldsymbol{\eta}}^t)}{n_1 - n_{1,2}(\tilde{\boldsymbol{\eta}}^t)}\Delta + \rho \tag{19}$$

$$\leq \frac{\beta}{\gamma - \beta}\Delta + \rho \ , \tag{20}$$

where the last inequality follows from our assumptions on $n_{2,1}(\tilde{\boldsymbol{\eta}}^t)$ and n_1. The same bound on $\|\tilde{p}_2^t - \vec{p}_2\|_1$ is obtained analogously.

It remains to bound ρ. Applying Inequality A.2 and using the bound on the KL divergence given in Inequality A.3, we have that with probability at least $1-\delta$, for every segmentation $\boldsymbol{\eta}'$ having $\max(n_{1,2}(\boldsymbol{\eta}'), n_{2,1}(\boldsymbol{\eta}'))/M \leq \beta$, and for $j \in \{1,2\}$,

$$e_j(\boldsymbol{\eta}') \leq \sqrt{\frac{2(K\ln(2M) + D\ln(M+1) + \ln(1/\delta)}{(\gamma-\beta)M}} \ . \tag{21}$$

■

C Proofs for Section 6

Proof of Lemma 6.1: The initialization procedure starts by considering all pairs of windows of the form $\sigma_i, \sigma_{i+1}, \ldots, \sigma_{i+\ell-1}$ and $\sigma_{i+\ell}, \sigma_{i+\ell+1}, \ldots, \sigma_{i+2\ell-1}$. The assumption on the minimal length of runs in $\boldsymbol{\eta}$ implies that each such pair overlaps with at most one switch between runs in $\boldsymbol{\eta}$. We concentrate on some particular pair of windows, whose corresponding estimates are $\vec{f}_i$ and $\vec{f}_{i+\ell}$ and assume, without loss of generality, that the switch is in the second window. Let $\vec{p}_a$ be the distribution which generates the elements before the switch and $\vec{p}_b$ be the distribution after the switch. Then the expected value of $\vec{f}_i$ is $\vec{p}_a$ and the expected value of $\vec{f}_{i+\ell}$ is $(1-\alpha_i)\vec{p}_a + \alpha_i\vec{p}_b$ for some $0 \leq \alpha_i \leq 1$. Using Inequality A.2 and our requirement on L we show that, for all $1 \leq i \leq M-\ell$, the actual value of $\vec{f}_i$ which is associated with each window is close to its expected value. Specifically, with probability at least $1-\delta$, the L_2 distance between each estimate $\vec{f}_i$ and its expected value is at most $\rho = \theta/2$. We thus get that the pair $\vec{f}_i$, $\vec{f}_{i+\ell}$ is added to the set A only if

$$\theta \geq \|\vec{f}_i - \vec{f}_{i+\ell}\|_2 \geq \alpha_i\|\vec{p}_a - \vec{p}_b\|_2 - \theta \ , \tag{22}$$

so that

$$\alpha_i \leq \frac{2\theta}{\|\vec{p}_a - \vec{p}_b\|_2} \ . \tag{23}$$

This implies that the estimate that is added to the set A in this case satisfies

$$\|(\vec{f}_i + \vec{f}_{i+\ell})/2 - \vec{p}_a\|_2 \leq (\alpha_i/2)\|\vec{p}_a - \vec{p}_b\|_2 + \theta/2 \leq \frac{3\theta}{2} \ . \tag{24}$$

Thus the estimates in the set B are all within $(3/2)\theta$ of actual target distributions.

On the other hand, we are guaranteed that each run contains at least one pair of estimates that are both pure, and it is easy to check that the accuracy of the estimates $\vec{f}_i$ guarantee that this pair will be accepted into A. Thus each target distribution has a least one representative in A.

The goal of step 4 of the initialization procedure is to find a *single* representative for each target distribution. Simple arguments show that the distance between two different representatives of

the same distribution is at most 3θ and the distance between two representatives of two different distributions is at least $\|\vec{p}_a - \vec{p}_b\|_2 - 3\theta$. From the requirement on L in the statement of the lemma we get that $\|\vec{p}_a - \vec{p}_b\|_2 - 3\theta \geq 4\theta$. Thus step 4 generates a set of distributions with one representative per target distribution, as required in the statement of the lemma. ■

Proof of Lemma 6.2: Assume some element in the sequence S is generated by the distribution $\vec{p}_j$ and then assigned a cost c from the cost vector $\vec{c}^t_{j'}$ which corresponds to the approximated distribution $\tilde{p}^t_{j'}$. Define $\vec{e} = \tilde{p}^t_{j'} - \vec{p}_j$. Then the expected value of c is

$$\begin{aligned}
\sum_{\sigma=1}^{D} \vec{p}_j(\sigma)\vec{c}_j(\sigma) &= \sum_{\sigma=1}^{D} \vec{p}_j(\sigma) \left(\sum_{\sigma'=1}^{D} (\tilde{p}^t_{j'}(\sigma'))^2 + 1 - 2\tilde{p}^t_j(\sigma) \right) && (25)\\
&= \sum_{\sigma=1}^{D} (\vec{p}_j(\sigma) + \vec{e}(\sigma))^2 + 1 - 2\sum_{\sigma=1}^{D} \vec{p}_j(\sigma)(\tilde{p}^t_{j'}(\sigma) + \vec{e}(\sigma)) && (26)\\
&= 1 - \sum_{\sigma=1}^{D} (\vec{p}_j(\sigma))^2 + \sum_{\sigma=1}^{D} \vec{e}(\sigma)^2 && (27)\\
&= 1 - \|\vec{p}_j\|_2^2 + \|\vec{e}\|_2^2 \,. && (28)
\end{aligned}$$

Thus the expected contribution of any element to the total cost is a sum of two terms. The first term depends only on the underlying target distribution, and the second term depends on the L_2 norm of the approximation error $\tilde{p}^t_{j'} - \vec{p}_j$.

Similarly to the analysis in the proof of Lemma 5.2 we now consider the difference between the total costs, corresponding to the approximate cost vectors, of two different segmentations. The first is the correct segmentation and the second is the best segmentation which minimizes the total cost on S. Clearly, only the elements on which $\boldsymbol{\eta}$ and $\tilde{\boldsymbol{\eta}}$ differ contribute to the difference in the total cost. From equation (28) we get that the expected total difference is

$$E_S\left[C\left(\{\vec{c}_j\}_{j=1}^N, \boldsymbol{\eta}, S\right) - C\left(\{\vec{c}_j\}_{j=1}^N, \boldsymbol{\eta}', S\right)\right] = \sum_{j' \neq j} n_{j,j'} \left(\|\tilde{p}^t_j - \vec{p}_j\|_2^2 - \|\tilde{p}^t_{j'} - \vec{p}_j\|_2^2 \right) \qquad (29)$$

Thus if $\|\tilde{p}^t_{j'} - \vec{p}_j\|_2 > |\tilde{p}^t_j - \vec{p}_j\|_2$ for all $j' \neq j$, then the expected cost difference is guaranteed to be negative. Using the triangle inequality for the L_2 norm we find that, from the conditions on the minimal separation and on the maximal error, that $\forall j' \neq j$, $\|\tilde{p}^t_{j'} - \vec{p}_j\|_2 - \|\tilde{p}^t_j - \vec{p}_j\|_2 \geq \Delta - 2\rho$. We thus get that the expected total difference in costs is bounded by

$$E_S\left[C\left(\{\vec{c}_j\}_{j=1}^N, \boldsymbol{\eta}, S\right) - C\left(\{\vec{c}_j\}_{j=1}^N, \boldsymbol{\eta}', S\right)\right] \leq -(\Delta - 2\rho)^2 \sum_{j' \neq j} n_{j,j'} \qquad (30)$$

We want to bound the probability that a segmentation with many errors has a total cost which is smaller than that of the correct segmentation. The cost difference is a sum of the cost differences in the places where the segmentations disagree. These are independent random variables. It is easy to check that the coordinates of any cost vector are bounded in the range $[0, 1]$ and thus the cost difference is bounded in $[-1, 1]$. We can thus apply Inequality A.1 and get that for any individual segmentation, the probability that the segmentation achieves smaller total cost than the correct segmentation is upper bounded by

$$Pr_S\left[C\left(\{\vec{c}_j\}_{j=1}^N, \boldsymbol{\eta}', S\right) < C\left(\{\vec{c}_j\}_{j=1}^N, \boldsymbol{\eta}, S\right)\right] \leq \exp\left(-\frac{1}{8}(\Delta - 2\rho)^4 \sum_{j' \neq j} n_{j,j'} \right) \,. \qquad (31)$$

There are less than $M^K N^K$ segmentations with K runs. Combining this with the last equation we get the statement of the lemma. ■

The Proof of Lemma 6.3 is the same as the proof of Lemma 5.3 except for the use of the L_2 norm in place of the L_1 norm.

D Proofs for Section 7

Proof of Lemma 7.1: We first consider the expected value of the total cost of the sequence S for some fixed hypothesis $(\tilde{\boldsymbol{\eta}}, \{\tilde{p}_j\}_{j=1}^N)$. Let σ_i be the ith element in S. Thus σ_i is generated by the distribution $\vec{p}_{\eta(i)}$ and then assigned a cost $c(i)$ according to the cost vector $\vec{c}_{\tilde{\eta}(i)}$ which corresponds to the approximated distribution $\tilde{p}_{\tilde{\eta}(i)}$. Define $\vec{e}_i = \vec{p}_{\eta(i)} - \tilde{p}_{\tilde{\eta}(i)}$. Then similarly to Equation (28), the expected value of $c(i)$ is

$$E[c(i)] = 1 - \|\vec{p}_{\eta(i)}\|_2^2 + \|\vec{e}_i\|_2^2 \ . \tag{32}$$

Summing the expected value of $c(i)$ over $i = 1 \ldots n$, we get the expected total cost of S is $A + \sum_{i=1}^n \|\vec{e}_i\|_2^2$. Where A is a constant independent of the hypothesis.

The range of all of the $c(i)$'s is $[0, 1]$ and they are independent random variables. Thus the deviation of the cost of the sequence from its expected value can be bounded using Inequality A.1 as follows

$$Pr\left[\left|\sum_{i=1}^{M} c(i) - E(\sum_{i=1}^{M} c(i))\right| > a\right] \leq \exp\left(-\frac{2a^2}{M}\right) \ . \tag{33}$$

We now return to the set of m hypotheses. As each of these hypotheses is selected from a set of at most $(NM)^K$ segmentations, we find that by setting a to be $\sqrt{(M/2)\ln(m(NM)^K/\delta)}$ we get that with probability at least $1 - \delta$ the total costs of all $m(NM)^K$ possible hypotheses are within a of their expected values.

Thus the expected total cost of the segmentation that appears to be best is within $2a$ of the expected value of the actual best segmentation. Which gives the statement of the lemma. ■