Comparative Analysis of Viterbi Training and Maximum Likelihood Estimation for HMMs

Armen Allahverdyan[*]
Yerevan Physics Institute
Yerevan, Armenia
aarmen@yerphi.am

Aram Galstyan
USC Information Sciences Institute
Marina del Rey, CA, USA
galstyan@isi.edu

Abstract

We present an asymptotic analysis of Viterbi Training (VT) and contrast it with a more conventional Maximum Likelihood (ML) approach to parameter estimation in Hidden Markov Models. While ML estimator works by (locally) maximizing the likelihood of the observed data, VT seeks to maximize the probability of the most likely hidden state sequence. We develop an analytical framework based on a generating function formalism and illustrate it on an exactly solvable model of HMM with one unambiguous symbol. For this particular model the ML objective function is continuously degenerate. VT objective, in contrast, is shown to have only finite degeneracy. Furthermore, VT converges faster and results in sparser (simpler) models, thus realizing an automatic Occam's razor for HMM learning. For more general scenario VT can be worse compared to ML but still capable of correctly recovering most of the parameters.

1 Introduction

Hidden Markov Models (HMM) provide one of the simplest examples of structured data observed through a noisy channel. The inference problems of HMM naturally divide into two classes [20, 9]: *i)* recovering the hidden sequence of states given the observed sequence, and *ii)* estimating the model parameters (transition probabilities of the hidden Markov chain and/or conditional probabilities of observations) from the observed sequence. The first class of problems is usually solved via the maximum a posteriori (MAP) method and its computational implementation known as Viterbi algorithm [20, 9]. For the parameter estimation problem, the prevailing method is maximum likelihood (ML) estimation, which finds the parameters by maximizing the likelihood of the observed data. Since global optimization is generally intractable, in practice it is implemented through an expectation–maximization (EM) procedure known as Baum–Welch algorithm [20, 9].

An alternative approach to parameter learning is Viterbi Training (VT), also known in the literature as segmental K-means, Baum–Viterbi algorithm, classification EM, hard EM, etc. Instead of maximizing the likelihood of the observed data, VT seeks to maximize the probability of the most likely hidden state sequence. Maximizing VT objective function is hard [8], so in practice it is implemented via an EM-style iterations between calculating the MAP sequence and adjusting the model parameters based on the sequence statistics. It is known that VT lacks some of the desired features of ML estimation such as consistency [17], and in fact, can produce biased estimates [9]. However, it has been shown to perform well in practice, which explains its widespread use in applications such as speech recognition [16], unsupervised dependency parsing [24], grammar induction [6], ion channel modeling [19]. It is generally assumed that VT is more robust and faster but usually less accurate, although for certain tasks it outperforms conventional EM [24].

[*]Currently at: *Laboratoire de Physique Statistique et Systemes Complexes*, ISMANS, Le Mans, France.

The current understanding of when and under what circumstances one method should be preferred over the other is not well–established. For HMMs with continuos observations, Ref. [18] established an upper bound on the difference between the ML and VT objective functions, and showed that both approaches produce asymptotically similar estimates when the dimensionality of the observation space is very large. Note, however, that this asymptotic limit is not very interesting as it makes the structure imposed by the Markovian process irrelevant. A similar attempt to compare both approaches on discrete models (for stochastic context free grammars) was presented in [23]. However, the established bound was very loose.

Our goal here is to understand, both qualitatively and quantitatively, the difference between the two estimation methods. We develop an analytical approach based on generating functions for examining the asymptotic properties of both approaches. Previously, a similar approach was used for calculating entropy rate of a hidden Markov process [1]. Here we provide a non-trivial extension of the methods that allows to perform comparative asymptotic analysis of ML and VT estimation. It is shown that both estimation methods correspond to certain free-energy minimization problem at different *temperatures*. Furthermore, we demonstrate the approach on a particular class of HMM with one unambiguous symbol and obtain a closed–form solution to the estimation problem. This class of HMMs is sufficiently rich so as to include models where not all parameters can be determined from the observations, i.e., the model is not *identifiable* [7, 14, 9].

We find that for the considered model VT is a better option if the ML objective is degenerate (i.e., not all parameters can be obtained from observations). Namely, not only VT recovers the identifiable parameters but it also provides a simple (in the sense that non-identifiable parameters are set to zero) and optimal (in the sense of the MAP performance) solution. Hence, VT realizes an automatic Occam's razor for the HMM learning. In addition, we show that the VT algorithm for this model converges faster than the conventional EM approach. Whenever the ML objective is not degenerate, VT leads generally to inferior results that, nevertheless, may be partially correct in the sense of recovering certain (not all) parameters.

2 Hidden Markov Process

Let $\mathcal{S} = \{\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, ...\}$ be a discrete-time, stationary, Markov process with conditional probability

$$\Pr[\mathcal{S}_{k+l} = s_k | \mathcal{S}_{k-1+l} = s_{k-1}] = p(s_k|s_{k-1}), \tag{1}$$

where l is an integer. Each realization s_k of the random variable $\mathcal{S}_k$ takes values $1, ..., L$. We assume that $\mathcal{S}$ is mixing: it has a unique stationary distribution $p_{\rm st}(s)$, $\sum_{r=1}^{L} p(s|r)p_{\rm st}(r) = p_{\rm st}(s)$, that is established from any initial probability in the long time limit.

Let random variables $\mathcal{X}_i$, with realizations $x_i = 1, .., M$, be noisy observations of $\mathcal{S}_i$: the (time-invariant) conditional probability of observing $\mathcal{X}_i = x_i$ given the realization $\mathcal{S}_i = s_i$ of the Markov process is $\pi(x_k|s_k)$. Defining $\mathbf{x} \equiv (x_N, ..., x_1)$, $\mathbf{s} \equiv (s_N, ..., s_0)$, the joint probability of $\mathcal{S}, \mathcal{X}$ reads

$$P(\mathbf{s}, \mathbf{x}) = T_{s_N\, s_{N-1}}(x_N)...T_{s_1\, s_0}(x_1)\, p_{\rm st}(s_0), \tag{2}$$

where the $L \times L$ transfer-matrix $T(x)$ with matrix elements $T_{s_i\, s_{i-1}}(x)$ is defined as

$$T_{s_i\, s_{i-1}}(x) = \pi(x|s_i)\, p(s_i|s_{i-1}). \tag{3}$$

$\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, ...\}$ is called a hidden Markov process. Generally, it is not Markov, but it inherits stationarity and mixing from $\mathcal{S}$ [9]. The probabilities for $\mathcal{X}$ can be represented as follows:

$$P(\mathbf{x}) = \sum\nolimits_{ss'} [\mathbb{T}(\mathbf{x})]_{ss'}\, p_{\rm st}(s'), \quad \mathbb{T}(\mathbf{x}) \equiv T(x_N)T(x_{N-1}) \dots T(x_1), \tag{4}$$

where $\mathbb{T}(\mathbf{x})$ is a product of transfer matrices.

3 Parameter Estimation

3.1 Maximum Likelihood Estimation

The unknown parameters of an HMM are the transition probabilities $p(s|s')$ of the Markov process and the observation probabilities $\pi(x|s)$; see (2). They have to be estimated from the observed

sequence $\mathbf{x}$. This is standardly done via the maximum-likelihood approach: one starts with some trial values $\hat{p}(s|s')$, $\hat{\pi}(x|s)$ of the parameters and calculates the (log)-likelihood $\ln \hat{P}(\mathbf{x})$, where $\hat{P}$ means the probability (4) calculated at the trial values of the parameters. Next, one maximizes $\ln \hat{P}(\mathbf{x})$ over $\hat{p}(s|s')$ and $\hat{\pi}(x|s)$ for the given observed sequence $\mathbf{x}$ (in practice this is done via the Baum-Welch algorithm [20, 9]). The rationale of this approach is as follows. Provided that the length N of the observed sequence is long, and recaling that $\mathcal{X}$ is mixing (due to the analogous feature of $\mathcal{S}$) we get probability-one convergence (law of large numbers) [9]:

$$\ln \hat{P}(\mathbf{x}) \to \sum\nolimits_{\mathbf{y}} P(\mathbf{y}) \ln \hat{P}(\mathbf{y}), \tag{5}$$

where the average is taken over the true probability $P(...)$ that generated $\mathbf{x}$. Since the relative entropy is non-negative, $\sum_{\mathbf{x}} P(\mathbf{x}) \ln[P(\mathbf{x})/\hat{P}(\mathbf{x})] \geq 0$, the global maximum of $\sum_{\mathbf{x}} P(\mathbf{x}) \ln \hat{P}(\mathbf{x})$ as a function of $\hat{p}(s|s')$ and $\hat{\pi}(x|s)$ is reached for $\hat{p}(s|s') = p(s|s')$ and $\hat{\pi}(x|s) = \pi(x|s)$. This argument is silent on how unique this global maximum is and how difficult to reach it.

3.2 Viterbi Training

An alternative approach to the parameter learning employs the maximal a posteriori (MAP) estimation and proceeds as follows: Instead of maximizing the likelihood of observed data (5) one tries to maximize the probability of the most likely sequence [20, 9]. Given the joint probability $\hat{P}(\mathbf{s}, \mathbf{x})$ at trial values of parameters, and given the observed sequence $\mathbf{x}$, one estimates the generating state-sequence $\mathbf{s}$ via maximizing the a posteriori probability

$$\hat{P}(\mathbf{s}|\mathbf{x}) = \hat{P}(\mathbf{s}, \mathbf{x})/\hat{P}(\mathbf{x}) \tag{6}$$

over $\mathbf{s}$. Since $\hat{P}(\mathbf{x})$ does not depend on $\mathbf{s}$, one can maximize $\ln \hat{P}(\mathbf{s}, \mathbf{x})$. If the number of observations is sufficiently large $N \to \infty$, one can substitute $\max_{\mathbf{s}} \ln \hat{P}(\mathbf{s}, \mathbf{x})$ by its average over $P(...)$ [see (5)] and instead maximize (over model parameters)

$$\sum\nolimits_{\mathbf{x}} P(\mathbf{x}) \max\nolimits_{\mathbf{s}} \ln \hat{P}(\mathbf{s}, \mathbf{x}). \tag{7}$$

To relate (7) to the free energy concept (see e.g. [2, 4]), we define an auxiliary (Gibbsian) probability

$$\hat{\rho}_\beta(\mathbf{s}|\mathbf{x}) = \hat{P}^\beta(\mathbf{s}, \mathbf{x}) / \left[\sum\nolimits_{\mathbf{s}'} \hat{P}^\beta(\mathbf{s}', \mathbf{x})\right], \tag{8}$$

where $\beta > 0$ is a parameter. As a function of $\mathbf{s}$ (and for a fixed $\mathbf{x}$), $\hat{\rho}_{\beta\to\infty}(\mathbf{s}|\mathbf{x})$ concentrates on those $\mathbf{s}$ that maximize $\ln \hat{P}(\mathbf{s}, \mathbf{x})$:

$$\hat{\rho}_{\beta\to\infty}(\mathbf{s}|\mathbf{x}) \to \frac{1}{\mathcal{N}} \sum\nolimits_j \delta[\mathbf{s}, \widetilde{\mathbf{s}}^{[j]}(\mathbf{x})], \tag{9}$$

where $\delta(s, s')$ is the Kronecker delta, $\widetilde{\mathbf{s}}^{[j]}(\mathbf{x})$ are equivalent outcomes of the maximization, and $\mathcal{N}$ is the number of such outcomes. Further, define

$$F_\beta \equiv -\frac{1}{\beta} \sum\nolimits_{\mathbf{x}} P(\mathbf{x}) \ln \sum\nolimits_{\mathbf{s}} \hat{P}^\beta(\mathbf{s}, \mathbf{x}). \tag{10}$$

Within statistical mechanics Eqs. 8 and 10 refer to, respectively, the *Gibbs distribution* and *free energy* of a physical system with Hamiltonian $H = -\ln P(\mathbf{s}, \mathbf{x})$ coupled to a thermal bath at inverse temperature $\beta = 1/T$ [2, 4]. It is then clear that ML and Viterbi Training correspond to minimizing the free energy Eq. 10 at $\beta = 1$ and $\beta = \infty$, respectively. Note that $\beta^2 \partial_\beta F = -\sum_{\mathbf{x}} P(\mathbf{x}) \sum_{\mathbf{s}} \rho_\beta(\mathbf{s}|\mathbf{x}) \ln \rho_\beta(\mathbf{s}|\mathbf{x}) \geq \mathbf{0}$, which yields $F_1 \leq F_\infty$.

3.3 Local Optimization

As we mentioned, global maximization of neither objective is feasible in the general case. Instead, in practice this maximization is *locally* implemented via an EM-type algorithm [20, 9]: for a given observed sequence $\mathbf{x}$, and for some initial values of the parameters, one calculates the expected value of the objective function under the trial parameters (E-step), and adjusts the parameters to maximize this expectation (M-step). The resulting estimates of the parameters are now employed as new trial parameters and the previous step is repeated. This recursion continues till convergence.

For our purposes, this procedure can be understood as calculating certain statistics of the hidden sequence averaged over the Gibbs distribution Eqs. 8. Indeed, let us introduce $f_\gamma(\mathbf{s}) \equiv e^{\beta\gamma\sum_{i=1}^{N}\delta(s_{i+1},a)\delta(s_i,b)}$ and define

$$\beta F_\beta(\gamma) \equiv -\sum\nolimits_{\mathbf{x}} P(\mathbf{x})\, \ln \sum\nolimits_{\mathbf{s}} \hat{P}^\beta(\mathbf{s},\mathbf{x}) f_\gamma(\mathbf{s}). \tag{11}$$

Then, for instance, the (iterative) Viterbi estimate of the transition probabilities are given as follows:

$$\widetilde{P}(\mathcal{S}_{k+1}=a,\, \mathcal{S}_k=b) = -\partial_\gamma [F_\infty(\gamma)]|_{\gamma\to 0}. \tag{12}$$

Conditional probabilities for observations are calculated similarly via a different indicator function.

4 Generating Function

Note from (4) that both $P(\mathbf{x})$ and $\hat{P}(\mathbf{x})$ are obtained as matrix-products. For a large number of multipliers the behavior of such products is governed by the multiplicative law of large numbers. We now recall its formulation from [10]: for $N\to\infty$ and $\mathbf{x}$ generated by the mixing process $\mathcal{X}$ there is a probability-one convergence:

$$\frac{1}{N}\ln||\mathbb{T}(\mathbf{x})|| \to \frac{1}{N}\sum\nolimits_{\mathbf{y}} P(\mathbf{y}) \ln\lambda[\mathbb{T}(\mathbf{y})], \tag{13}$$

where $||...||$ is a matrix norm in the linear space of $L\times L$ matrices, and $\lambda[\mathbb{T}(\mathbf{x})]$ is the maximal eigenvalue of $\mathbb{T}(\mathbf{x})$. Note that (13) does not depend on the specific norm chosen, because all norms in the finite-dimensional linear space are equivalent; they differ by a multiplicative factor that disappears for $N\to\infty$ [10]. Eqs. (4, 13) also imply $\sum_{\mathbf{x}}\lambda[\mathbb{T}(\mathbf{x})]\to 1$. Altogether, we calculate (5) via its probability-one limit

$$\frac{1}{N}\sum\nolimits_{\mathbf{x}} P(\mathbf{x})\ln\hat{P}(\mathbf{x}) \to \frac{1}{N}\sum\nolimits_{\mathbf{x}}\lambda[\mathbb{T}(\mathbf{x})]\ln\lambda[\hat{\mathbb{T}}(\mathbf{x})]. \tag{14}$$

Note that the multiplicative law of large numbers is normally formulated for the maximal singular value. Its reformulation in terms of the maximal eigenvalue needs additional arguments [1].

Introducing the generating function

$$\Lambda^N(n,N) = \sum\nolimits_{\mathbf{x}}\lambda[\mathbb{T}(\mathbf{x})]\lambda^n\left[\hat{\mathbb{T}}(\mathbf{x})\right], \tag{15}$$

where n is a non-negative number, and where $\Lambda^N(n,N)$ means $\Lambda(n,N)$ in degree of N, one represents (14) as

$$\frac{1}{N}\sum\nolimits_{\mathbf{x}}\lambda[\mathbb{T}(\mathbf{x})]\ln\lambda[\hat{\mathbb{T}}(\mathbf{x})] = \partial_n\Lambda(n,N)|_{n=0}, \tag{16}$$

where we took into account $\Lambda(0,N)=1$, as follows from (15).

The behavior of $\Lambda^N(n,N)$ is better understood after expressing it via the zeta-function $\xi(z,n)$ [1]

$$\xi(z,n) = \exp\left[-\sum\nolimits_{m=1}^{\infty}\frac{z^m}{m}\Lambda^m(n,m)\right], \tag{17}$$

where $\Lambda^m(n,m)\geq 0$ is given by (15). Since for a large N, $\Lambda^N(n,N)\to\Lambda^N(n)$ [this is the content of (13)], the zeta-function $\xi(z,n)$ has a zero at $z=\frac{1}{\Lambda(n)}$:

$$\xi(1/\Lambda(n),\, n) = 0. \tag{18}$$

Indeed for z close (but smaller than) $\frac{1}{\Lambda(n)}$, the series $\sum_{m=1}^{\infty}\frac{z^m}{m}\Lambda^m(n,m)\to\sum_{m=1}^{\infty}\frac{[z\Lambda(n)]^m}{m}$ almost diverges and one has $\xi(z,n)\to 1-z\Lambda(n)$. Recalling that $\Lambda(0)=1$ and taking $n\to 0$ in $0=\frac{\mathrm{d}}{\mathrm{d}n}\xi(\frac{1}{\Lambda(n)},n)$, we get from (16)

$$\frac{1}{N}\sum\nolimits_{\mathbf{x}}\lambda[\mathbb{T}(\mathbf{x})]\ln\lambda[\hat{\mathbb{T}}(\mathbf{x})] = \frac{\partial_n\xi(1,0)}{\partial_z\xi(1,0)}. \tag{19}$$

For calculating $-\beta F_\beta$ in (10) we have instead of (19)

$$-\frac{\beta F_\beta}{N} = \frac{\partial_n\xi^{[\beta]}(1,0)}{\partial_z\xi^{[\beta]}(1,0)}, \tag{20}$$

where $\xi^{[\beta]}(z,n)$ employs $\hat{T}^\beta_{s_i\, s_{i-1}}(x) = \hat{\pi}^\beta(x|s_i)\,\hat{p}^\beta(s_i|s_{i-1})$ instead of $\hat{T}_{s_i\, s_{i-1}}(x)$ in (19).

Though in this paper we restricted ourselves to the limit $N\to\infty$, we stress that the knowledge of the generating function $\Lambda^N(n,N)$ allows to analyze the learning algorithms for any finite N.

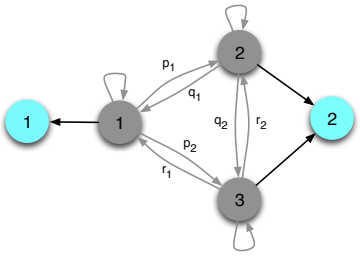

Figure 1: *The hidden Markov process (21–22) for $\epsilon = 0$. Gray circles and arrows indicate on the realization and transitions of the internal Markov process; see (21). The light circles and black arrows indicate on the realizations of the observed process.*

5 Hidden Markov Model with One Unambiguous Symbol

5.1 Definition

Given a L-state Markov process $\mathcal{S}$, the observed process $\mathcal{X}$ has two states 1 and 2; see Fig. 1. All internal states besides one are observed as 2, while the internal state 1 produces, respectively, 1 and 2 with probabilities $1-\epsilon$ and ϵ. For $L = 3$ we obtain from (1) $\pi(1|1) = 1 - \pi(2|1) = 1 - \epsilon$, $\pi(1|2) = \pi(1|3) = \pi(2|1) = 0$, $\pi(2|2) = \pi(2|3) = 1$. Hence 1 is unambiguous: if it is observed, the unobserved process $\mathcal{S}$ was certainly in 1; see Fig. 1. The simplest example of such HMM exists already for $L = 2$; see [12] for analytical features of entropy for this case. We, however, describe in detail the $L = 3$ situation, since this case will be seen to be generic (in contrast to $L = 2$) and it allows straightforward generalizations to $L > 3$. The transition matrix (1) of a general $L = 3$ Markov process reads

$$\mathbb{P} \equiv \{\, p(s|s') \,\}_{s,s'=1}^{3} = \begin{pmatrix} p_0 & q_1 & r_1 \\ p_1 & q_0 & r_2 \\ p_2 & q_2 & r_0 \end{pmatrix}, \quad \begin{pmatrix} p_0 \\ q_0 \\ r_0 \end{pmatrix} = \begin{pmatrix} 1-p_1-p_2 \\ 1-r_1-r_2 \\ 1-r_1-r_2 \end{pmatrix} \tag{21}$$

where, e.g., $q_1 = p(1|2)$ is the transition probability $2 \to 1$; see Fig. 1. The corresponding transfer matrices read from (3)

$$T(1) = (1-\epsilon) \begin{pmatrix} p_0 & q_1 & r_1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad T(2) = \mathbb{P} - T(1). \tag{22}$$

Eq. (22) makes straightforward the reconstruction of the transfer-matrices for $L \geq 4$. It should also be obvious that for all L only the first row of $T(1)$ consists of non-zero elements.

For $\epsilon = 0$ we get from (22) the simplest example of an aggregated HMM, where several Markov states are mapped into one observed state. This model plays a special role for the HMM theory, since it was employed in the pioneering study of the non-identifiability problem [7].

5.2 Solution of the Model

For this model $\xi(z,n)$ can be calculated exactly, because $T(1)$ has only one non-zero row. Using the method outlined in the supplementary material (see also [1, 3]) we get

$$\xi(z,n) = 1 - z(t_0 \hat{t}_0^n + \tau_0 \hat{\tau}_0^n) + \sum\nolimits_{k=2}^{\infty} [\tau \hat{\tau}^n \hat{t}_{k-2}^n t_{k-2} - \hat{t}_{k-1}^n t_{k-1}] z^k \tag{23}$$

where τ and $\hat{\tau}$ are the largest eigenvalues of $T(2)$ and $\hat{T}(2)$, respectively

$$t_k \equiv \langle 1|T(1)T(2)^k|1\rangle = \sum\nolimits_{\alpha=1}^{L} \tau_\alpha^k \psi_\alpha, \tag{24}$$

$$\psi_\alpha \equiv \langle 1|T(1)|R_\alpha\rangle\langle L_\alpha|1\rangle, \quad \langle 1| \equiv (1,0,\ldots,0). \tag{25}$$

Here $|R_\alpha\rangle$ and $\langle L_\alpha|$ are, respectively right and left eigenvalues of $T(2)$, while $\tau_1, \ldots, \tau_L$ ($\tau_L \equiv \tau$) are the eigenvalues of $T(2)$. Eqs. (24, 25) obviously extend to hatted quantities.

We get from (23, 19):

$$\xi(1,n) = (1-\hat{\tau}^n\tau)\left(1-\sum\nolimits_{k=0}^{\infty}\hat{t}_k^n t_k\right), \tag{26}$$

$$\frac{\partial_n\xi(1,0)}{\partial_z\xi(1,0)} = \frac{\sum_{k=0}^{\infty}t_k\ln[\hat{t}_k]}{\sum_{k=0}^{\infty}(k+1)t_k}. \tag{27}$$

Note that for $\epsilon = 0$, t_k are return probabilities to the state 1 of the L-state Markov process. For $\epsilon > 0$ this interpretation does not hold, but t_k still has a meaning of probability as $\sum_{k=0}^{\infty}t_k = 1$.

Turning to equations (19, 27) for the free energy, we note that as a function of trial values it depends on the following $2L$ parameters:

$$(\hat{\tau}_1,\ldots,\hat{\tau}_L,\hat{\psi}_1,\ldots,\hat{\psi}_L). \tag{28}$$

As a function of the true values, the free energy depends on the same $2L$ parameters (28) [without hats], though concrete dependencies are different. For the studied class of HMM there are at most $L(L-1)+1$ unknown parameters: $L(L-1)$ transition probabilities of the unobserved Markov chain, and one parameter ϵ coming from observations. We checked numerically that the Jacobian of the transformation from the unknown parameters to the parameters (28) has rank $2L-1$. Any $2L-1$ parameters among (28) can be taken as independent ones.

For $L > 2$ the number of *effective* independent parameters that affect the free energy is smaller than the number of parameters. So if the number of unknown parameters is larger than $2L-1$, neither of them can be found explicitly. One can only determine the values of the effective parameters.

6 The Simplest Non-Trivial Scenario

The following example allows the full analytical treatment, but is generic in the sense that it contains all the key features of the more general situation given above (21). Assume that $L = 3$ and

$$q_0 = \hat{q}_0 = r_0 = \hat{r}_0 = 0, \qquad \epsilon = \hat{\epsilon} = 0. \tag{29}$$

Note the following explicit expressions

$$t_0 = p_0,\ t_1 = p_1q_1 + p_2r_1,\ t_2 = p_1r_1q_2 + p_2q_1r_2, \tag{30}$$

$$\tau = \tau_3 = \sqrt{q_2r_2},\quad \tau_2 = -\tau,\quad \tau_1 = 0, \tag{31}$$

$$\psi_3 - \psi_2 = t_1/\tau,\quad \psi_3 + \psi_2 = t_2/\tau^2, \tag{32}$$

Eqs. (30–32) with obvious changes $s_i \to \hat{s}_i$ for every symbol s_i hold for $\hat{t}_k$, $\hat{\tau}_k$ and $\hat{\psi}_k$. Note a consequence of $\sum_{k=0}^{2}p_k = \sum_{k=0}^{2}q_k = \sum_{k=0}^{2}r_k = 1$:

$$\tau^2(1-t_0) = 1 - t_0 - t_1 - t_2. \tag{33}$$

6.1 Optimization of F_1

Eqs. (27) and (30–32) imply $\sum_{k=0}^{\infty}(k+1)t_k = \frac{\mu}{1-\tau^2}$,

$$\mu \equiv 1 - \tau^2 + t_2 + (1-t_0)(1+\tau^2) > 0, \tag{34}$$

$$-\frac{\mu F_1}{N} = t_1\ln\hat{t}_1 + t_2\ln\hat{t}_2 + (1-\tau^2)t_0\ln\hat{t}_0 + (1-t_0)\tau^2\ln\hat{\tau}^2\ . \tag{35}$$

The free energy F_1 depends on three independent parameters $\hat{t}_0, \hat{t}_1, \hat{t}_2$ [recall (33)]. Hence, minimizing F_1 we get $\hat{t}_i = t_i$ $(i = 0, 1, 2)$, but we do not obtain a definite solution for the unknown parameters: any four numbers $\hat{p}_1$, $\hat{p}_2$, $\hat{q}_1$, $\hat{r}_1$ satisfying three equations $t_0 = 1 - \hat{p}_1 - \hat{p}_2$, $t_1 = \hat{p}_1\hat{q}_1 + \hat{p}_2\hat{r}_1$, $t_2 = \hat{p}_1\hat{r}_1(1-\hat{q}_1) + \hat{p}_2\hat{q}_1(1-\hat{r}_1)$, minimize F_1.

6.2 Optimization of F_∞

In deriving (35) we used no particular feature of $\{\hat{p}_k\}_{k=0}^2$, $\{\hat{q}_k\}_{k=1}^2$, $\{\hat{r}_k\}_{k=1}^2$. Hence, as seen from (20), the free energy at $\beta > 0$ is recovered from (35) by equating its LHS to $-\frac{\beta F_\beta}{N}$ and by taking in

its RHS: $\hat{t}_0 \to \hat{p}_0^\beta$, $\hat{\tau}^2 \to \hat{q}_2^\beta \hat{r}_2^\beta$, $\hat{t}_1 \to \hat{p}_1^\beta \hat{q}_1^\beta + \hat{p}_2^\beta \hat{r}_1^\beta$, $\hat{t}_2 \to \hat{p}_1^\beta \hat{r}_1^\beta \hat{q}_2^\beta + \hat{p}_2^\beta \hat{q}_1^\beta \hat{r}_2^\beta$. The zero-temperature free energy reads from (35)

$$-\frac{\mu F_\infty}{N} = (1-\tau^2) t_0 \ln \hat{t}_0 + (1-t_0)\tau^2 \ln \hat{\tau}^2 \quad + \quad t_1 \ln \max[\hat{p}_1 \hat{q}_1, \hat{p}_2 \hat{r}_1] \\ + \quad t_2 \ln \max[\hat{p}_2 \hat{q}_1 \hat{r}_2, \hat{p}_1 \hat{r}_1 \hat{q}_2]. \tag{36}$$

We now minimize F_∞ over the trial parameters $\hat{p}_1, \hat{p}_2, \hat{q}_1, \hat{r}_1$. This is not what is done by the VT algorithm; see the discussion after (12). But at any rate both procedures aim to minimize the same target. VT recursion for this models will be studied in section 6.3 — it leads to the same result. Minimizing F_∞ over the trial parameters produces four distinct solutions:

$$\{\hat{\sigma}_i\}_{i=1}^4 = \{\hat{p}_1 = 0,\, \hat{p}_2 = 0,\, \hat{q}_1 = 0,\, \hat{r}_1 = 0\}. \tag{37}$$

For each of these four solutions: $\hat{t}_i = t_i$ ($i = 0, 1, 2$) and $F_1 = F_\infty$. The easiest way to get these results is to minimize F_∞ under conditions $\hat{t}_i = t_i$ (for $i = 0, 1, 2$), obtain $F_1 = F_\infty$ and then to conclude that due to the inequality $F_1 \leq F_\infty$ the conditional minimization led to the global minimization. The logics of (37) is that the unambiguous state tends to get detached from the ambiguous ones, since the probabilities nullifying in (37) refer to transitions from or to the unambiguous state.

Note that although minimizing either F_∞ and F_1 produces correct values of the independent variables t_0, t_1, t_2, in the present situation minimizing F_∞ is preferable, because it leads to the four-fold degenerate set of solutions (37) instead of the continuously degenerate set. For instance, if the solution with $\hat{p}_1 = 0$ is chosen we get for other parameters

$$\hat{p}_2 = 1 - t_0, \;\; \hat{q}_1 = \frac{t_2}{1 - t_0 - t_1}, \;\; \hat{r}_1 = \frac{t_1}{1 - t_0}. \tag{38}$$

Furthermore, a more elaborate analysis reveals that for each fixed set of correct parameters only one among the four solutions Eq. 37 provides the best value for the quality of the MAP reconstruction, i.e. for the overlap between the original and MAP-decoded sequences.

Finally, we note that minimizing F_∞ allows one to get the correct values t_0, t_1, t_2 of the independent variables $\hat{t}_0$, $\hat{t}_1$ and $\hat{t}_2$ only if their number is less than the number of unknown parameters. This is not a drawback, since once the number of unknown parameters is sufficiently small [less than four for the present case (29)] their exact values are obtained by minimizing F_1. Even then, the minimization of F_∞ can provide partially correct answers. Assume in (36) that the parameter $\hat{r}_1$ is known, $\hat{r}_1 = r_1$. Now F_∞ has three local minima given by $\hat{p}_1 = 0$, $\hat{p}_2 = 0$ and $\hat{q}_1 = 0$; cf. with (37). The minimum with $\hat{p}_2 = 0$ is the global one and it allows to obtain the exact values of the two effective parameters: $\hat{t}_0 = 1 - \hat{p}_1 = t_0$ and $\hat{t}_1 = \hat{p}_1 \hat{q}_1 = t_1$. These effective parameters are recovered, because they do not depend on the known parameter $\hat{r}_1 = r_1$. Two other minima have greater values of F_∞, and they allow to recover only one effective parameter: $\hat{t}_0 = 1 - \hat{p}_1 = t_0$. If in addition to $\hat{r}_1$ also $\hat{q}_1$ is known, the two local minimia of F_∞ ($\hat{p}_1 = 0$ and $\hat{p}_2 = 0$) allow to recover $\hat{t}_0 = t_0$ only. In contrast, if $\hat{p}_1 = p_1$ (or $\hat{p}_2 = p_2$) is known exactly, there are three local minima again—$\hat{p}_2 = 0$, $\hat{q}_1 = 0$, $\hat{r}_1 = 0$—but now none of effective parameters is equal to its true value: $\hat{t}_i \neq t_i$ ($i = 0, 1, 2$).

6.3 Viterbi EM

Recall the description of the VT algorithm given after (12). For calculating $\widetilde{P}(\mathcal{S}_{k+1} = a, \mathcal{S}_k = b)$ via (11, 12) we modify the transfer matrix element in (15, 17) as $\hat{T}_{ab}(k) \to \hat{T}_{ab}(k) e^\gamma$, which produces from (11, 12) for the MAP-estimates of the transition probabilities

$$\widetilde{p}_1 = \frac{t_1 \hat{\chi}_1 + t_2 \hat{\chi}_2}{t_1 + t_2 + t_0(1-\tau^2)}, \quad \widetilde{p}_2 = 1 - t_0 - \widetilde{p}_1, \tag{39}$$

$$\widetilde{q}_1 = \frac{t_1 \hat{\chi}_1 + t_2 (1 - \hat{\chi}_2)}{t_1 \hat{\chi}_1 + t_2 + (1-t_0)\tau^2}, \quad \widetilde{q}_2 = 1 - \widetilde{q}_1 \tag{40}$$

$$\widetilde{r}_1 = \frac{t_1 (1 - \hat{\chi}_1) + t_2 \hat{\chi}_2}{t_2 + t_1 (1 - \hat{\chi}_1) + (1-t_0)\tau^2} \quad \widetilde{r}_2 = 1 - \widetilde{r}_1, \tag{41}$$

where $\hat{\chi}_1 \equiv \frac{\hat{p}_1^\beta \hat{q}_1^\beta}{\hat{p}_1^\beta \hat{q}_1^\beta + \hat{p}_2^\beta \hat{r}_1^\beta}$, $\hat{\chi}_2 \equiv \frac{\hat{p}_1^\beta \hat{r}_1^\beta \hat{q}_2^\beta}{\hat{p}_1^\beta \hat{r}_1^\beta \hat{q}_2^\beta + \hat{p}_2^\beta \hat{r}_2^\beta \hat{q}_1^\beta}$. The $\beta \to \infty$ limit of $\hat{\chi}_1$ and $\hat{\chi}_2$ is obvious: each of them is equal to 0 or 1 depending on the ratios $\frac{\hat{p}_1 \hat{q}_1}{\hat{p}_2 \hat{r}_1}$ and $\frac{\hat{p}_1 \hat{r}_1 \hat{q}_2}{\hat{p}_2 \hat{r}_2 \hat{q}_1}$. The EM approach amounts to

starting with some trial values $\hat{p}_1$, $\hat{p}_2$, $\hat{q}_1$, $\hat{r}_1$ and using $\widetilde{p}_1$, $\widetilde{p}_2$, $\widetilde{q}_1$, $\widetilde{r}_1$ as new trial parameters (and so on). We see from (39–41) that the algorithm converges just in one step: (39–41) are equal to the parameters given by one of four solutions (37)—which one among the solutions (37) is selected depends on the on initial trial parameters in (39–41)—recovering the correct effective parameters (30–32); e.g. cf. (38) with (39, 41) under $\hat{\chi}_1 = \hat{\chi}_2 = 0$. Hence, VT converges in one step in contrast to the Baum-Welch algorithm (that uses EM to locally minimize F_1) which, for the present model, obviously does not converge in one step. There is possibly a deeper point in the one-step convergence that can explain why in practice VT converges faster than the Baum-Welch algorithm [9, 21]: recall that, e.g. the Newton method for local optimization works precisely in one step for quadratic functions, but generally there is a class of functions, where it performs faster than (say) the steepest descent method. Further research should show whether our situation is similar: the VT works just in one step for this exactly solvable HMM model that belongs to a class of models, where VT generally performs faster than ML.

We conclude this section by noting that the solvable case (29) is generic: its key results extend to the general situation defined above (21). We checked this fact numerically for several values of L. In particular, the minimization of F_∞ nullifies as many trial parameters as necessary to express the remaining parameters via independent effective parameters $t_0, t_1, \ldots$. Hence for $L = 3$ and $\epsilon = 0$ two such trial parameters are nullified; cf. with discussion around (28). If the true error probability $\epsilon \neq 0$, the trial value $\hat{\epsilon}$ is among the nullified parameters. Again, there is a discrete degeneracy in solutions provided by minimizing F_∞.

7 Summary

We presented a method for analyzing two basic techniques for parameter estimation in HMMs, and illustrated it on a specific class of HMMs with one unambiguous symbol. The virtue of this class of models is that it is exactly solvable, hence the sought quantities can be obtained in a closed form via generating functions. This is a rare occasion, because characteristics of HMM such as likelihood or entropy are notoriously difficult to calculate explicitly [1]. An important feature of the example considered here is that the set of unknown parameters is not completely identifiable in the maximum likelihood sense [7, 14]. This corresponds to the zero eigenvalue of the Hessian for the ML (maximum-likelihood) objective function. In practice, one can have weaker degeneracy of the objective function resulting in very small values for the Hessian eigenvalues. This scenario occurs *often* in various models of physics and computational biology [11]. Hence, it is a drawback that the theory of HMM learning was developed assuming complete identifiably [5].

One of our main result is that in contrast to the ML approach that produces continuously degenerate solutions, VT results in finitely degenerate solution that is sparse, i.e., some [non-identifiable] parameters are set to zero, and, furthermore, converges faster. Note that sparsity might be a desired feature in many practical applications. For instance, imposing sparsity on conventional EM-type learning has been shown to produce better results part of speech tagging applications [25]. Whereas [25] had to impose sparsity via an additional penalty term in the objective function, in our case sparsity is a natural outcome of maximizing the likelihood of the best sequence. While our results were obtained on a class of exactly-solvable model, it is plausible that they hold more generally.

The fact that VT provides simpler and more definite solutions—among all choices of the parameters compatible with the observed data—can be viewed as a type of the Occam's razor for the parameter learning. Note finally that statistical mechanics intuition behind these results is that the aposteriori likelihood is (negative) zero-temperature free energy of a certain physical system. Minimizing this free energy makes physical sense: this is the premise of the second law of thermodynamics that ensures relaxation towards a more equilibrium state. In that zero-temperature equilibrium state certain types of motion are frozen, which means nullifying the corresponding transition probabilities. In that way the second law relates to the Occam's razor. Other connections of this type are discussed in [15].

Acknowledgments

This research was supported in part by the US ARO MURI grant No. W911NF0610094 and US DTRA grant HDTRA1-10-1-0086.

References

[1] A. E. Allahverdyan, *Entropy of Hidden Markov Processes via Cycle Expansion*, J. Stat. Phys. **133**, 535 (2008).

[2] A.E. Allahverdyan and A. Galstyan, *On Maximum a Posteriori Estimation of Hidden Markov Processes*, Proc. of UAI, (2009).

[3] R. Artuso. E. Aurell and P. Cvitanovic, *Recycling of strange sets*, Nonlinearity **3**, 325 (1990).

[4] P. Baldi and S. Brunak, *Bioinformatics*, MIT Press, Cambridge, USA (2001).

[5] L. E. Baum and T. Petrie, *Statistical inference for probabilistic functions of finite state Markov chains*, Ann. Math. Stat. **37**, 1554 (1966).

[6] J.M. Benedi, J.A. Sanchez, *Estimation of stochastic context-free grammars and their use as language models*, Comp. Speech and Lang. **19**, pp. 249-274 (2005).

[7] D. Blackwell and L. Koopmans, *On the identifiability problem for functions of finite Markov chains*, Ann. Math. Statist. **28**, 1011 (1957).

[8] S. B. Cohen and N. A. Smith, *Viterbi training for PCFGs: Hardness results and competitiveness of uniform initialization*, Procs. of ACL (2010).

[9] Y. Ephraim and N. Merhav, *Hidden Markov processes*, IEEE Trans. Inf. Th., **48**, 1518-1569, (2002).

[10] L.Y. Goldsheid and G.A. Margulis, *Lyapunov indices of a product of random matrices*, Russ. Math. Surveys **44**, 11 (1989).

[11] R. N. Gutenkunst *et al.*, *Universally Sloppy Parameter Sensitivities in Systems Biology Models*, PLoS Computational Biology, **3**, 1871 (2007).

[12] G. Han and B. Marcus, *Analyticity of entropy rate of hidden Markov chains*, IEEE Trans. Inf. Th., **52**, 5251 (2006).

[13] R. A. Horn and C. R. Johnson, *Matrix Analysis* (Cambridge University Press, New Jersey, USA, 1985).

[14] H. Ito, S. Amari, and K. Kobayashi, *Identifiability of Hidden Markov Information Sources*, IEEE Trans. Inf. Th., **38**, 324 (1992).

[15] D. Janzing, *On causally asymmetric versions of Occam's Razor and their relation to thermodynamics*, arXiv:0708.3411 (2007).

[16] B. H. Juang and L. R. Rabiner, *The segmental k-means algorithm for estimating parameters of hidden Markov models*, IEEE Trans. Acoustics, Speech, and Signal Processing, ASSP-38, no.9, pp.1639-1641, (1990).

[17] B. G. Leroux, *Maximum-Likelihood Estimation for Hidden Markov Models*, Stochastic Processes and Their Applications, **40**, 127 (1992).

[18] N. Merhav and Y. Ephraim, *Maximum likelihood hidden Markov modeling using a dominant sequence of states*, IEEE Transactions on Signal Processing, vol.39, no.9, pp.2111-2115 (1991).

[19] F. Qin, *Restoration of single-channel currents using the segmental k-means method based on hidden Markov modeling*, Biophys J **86**, 14881501 (2004).

[20] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE, **77**, 257 (1989).

[21] L. J. Rodriguez and I. Torres, *Comparative Study of the Baum-Welch and Viterbi Training Algorithms*, Pattern Recognition and Image Analysis, Lecture Notes in Computer Science, **2652/2003**, 847 (2003).

[22] D. Ruelle, *Statistical Mechanics, Thermodynamic Formalism*, (Reading, MA: Addison-Wesley, 1978).

[23] J. Sanchez, J. Benedi, F. Casacuberta, *Comparison between the inside-outside algorithm and the Viterbi algorithm for stochastic context-free grammars*, in Adv. in Struct. and Synt. Pattern Recognition (1996).

[24] V. I. Spitkovsky, H. Alshawi, D. Jurafsky, and C. D. Manning, *Viterbi Training Improves Unsupervised Dependency Parsing*, in Proc. of the 14th Conference on Computational Natural Language Learning (2010).

[25] A. Vaswani, A. Pauls, and D. Chiang, *Efficient optimization of an MDL-inspired objective function for unsupervised part-of-speech tagging*, in Proc. ACL (2010).