Online Learning in Discrete Hidden Markov Models

Roberto Alamino* and Nestor Caticha†

*Neural Computing Research Group,
Aston University
Aston Triangle, Birmingham, B4 7ET, United Kingdom
†Instituto de Física,
Universidade de São Paulo,
CP 66318 São Paulo, SP, CEP 05389-970 Brazil

Abstract. We present and analyze three different online algorithms for learning in discrete Hidden Markov Models (HMMs) and compare their performance with the Baldi-Chauvin Algorithm. Using the Kullback-Leibler divergence as a measure of the generalization error we draw learning curves in simplified situations and compare the results. The performance for learning drifting concepts of one of the presented algorithms is analyzed and compared with the Baldi-Chauvin algorithm in the same situations. A brief discussion about learning and symmetry breaking based on our results is also presented.

Key Words: HMMs, Online Algorithm, Generalization Error, Bayesian Algorithm.

INTRODUCTION

Hidden Markov Models (HMMs) [1, 2] are extensively studied machine learning models for time series with several applications in fields like speech recognition [2], bioinformatics [3, 4] and LDPC codes [5]. They consist of a Markov chain of non-observable *hidden states* $q_t \in S$, $t = 1,...,T$, $S = \{s_1, s_2, ..., s_n\}$, with initial probability vector $\pi_i = \mathcal{P}(q_1 = s_i)$ and transition matrix $A_{ij}(t) = \mathcal{P}(q_{t+1} = s_j | q_t = s_i)$, $i, j = 1,..,n$. At discrete times t, each q_t emmits an *observed state* $y_t \in O$, $O = \{o_1, ..., o_m\}$, with emission probability matrix $B_{i\alpha}(t) = \mathcal{P}(y_t = o_\alpha | q_t = s_i)$, $i = 1,...,n$, $\alpha = 1,...,m$, which are the actual observations of the time series represented, from time $t = 1$ to $t = T$, by the *observed sequence* $y_1^T = \{y_1, y_2, ..., y_T\}$. The q_t's form the so called *hidden sequence* $q_1^T = \{q_1, q_2, ..., q_T\}$. The probability of observing a sequence y_1^T given $\omega \equiv (\pi, A, B)$ is

$$\mathcal{P}(y_1^T | \omega) = \sum_{q_1^T} \mathcal{P}(y_1) \mathcal{P}(y_1 | q_1) \prod_{t=2}^{T} \mathcal{P}(q_{t+1} | q_t) \mathcal{P}(y_t | q_t). \tag{1}$$

The *learning process* consists in presenting a series to the HMM which adapts its parameters in order to produce sequences that mimic it. Depending on how data is presented, it can range from *offline*, when the whole data is given and parameters are calculated all at once, to *online*, when the data is given only by parts and a partial calculation of the parameters is made.

We study a scenario with a data set generated by a HMM of unknown parameters. This is an extension of the student-teacher scenario extensively studied in neural networks. The performance of the learning process, as a function of the number of observations, is given by how *far*, measured by a suitable criterion, is the student from the teacher. Here we use the naturally arising *Kullback-Leibler divergence* (KL-divergence), which although not accessible in practice since it needs knowledge from the teacher, is a simple extension of the idea of generalization error and therefore can be very informative.

We propose three algorithms and compare them with the *Baldi-Chauvin Algorithm* (BC) [6]: the *Baum-Welch Online Algorithm* (BWO), an adaptation of the offline *Baum-Welch Reestimation Formulas* (BW) [1], then, starting from a Bayesian formulation, an approximation called the *Bayesian Online Algorithm* (BOnA), which can be simplified further without noticeable deterioration of performance to a *Mean Posterior Algorithm* (MPA). The last two methods, inspired by the work of Amari [7] and Opper [8] are essentially mean field methods [9] in which a manifold of tractable distributions to be used as priors is introduced and the new datum leads, through Bayes theorem, to a non-tractable posterior. The key inference step is to take as the new prior, not the posterior itself, but the distribution in the manifold which is the closest in some sense.

The paper is organized as follows: first, BWO is introduced and analyzed. Next, we derive BOnA for HMMs and, from it, MPA. We compare the behaviour of MPA and BC with respect to learning drifting concepts and then present a discussion about learning and symmetry breaking based upon our results followed by our conclusions.

BAUM-WELCH ONLINE ALGORITHM

The *Baum-Welch Online Algorithm* (BWO) is an adaptation of BW to online situations where in each iteration of BW, which is a step towards a maximum of the average over the hidden sequences of $\mathcal{P}(q,y)$, y becomes y^p, the p-th observed sequence. Multiplying the BW increment by a learning rate η_{BW} we get the update equations

$$\hat{\omega}^{p+1} = \hat{\omega}^p + \eta_{BW}\hat{\Delta}\omega^p, \tag{2}$$

with $\hat{\Delta}\omega^p$ the BW variations of ω calculated with y^p. The complexity of BWO is polynomial in n and T.

In figure 1, the HMM learns sequences generated by a teacher with $n = 2$, $m = 3$ and $T = 2$ for different η_{BW}. Initial students have matrices with all entries set to the same value, what we call a *symmetric initial student*. We took averages over 500 random teachers and distances are given by the KL-divergence between two HMMs ω_1 and ω_2

$$d_{KL}(\omega_1,\omega_2) \equiv \sum_{y_1^T} \mathcal{P}(y_1^T|\omega_1) \ln\left[\frac{\mathcal{P}(y_1^T|\omega_1)}{\mathcal{P}(y_1^T|\omega_2)}\right]. \tag{3}$$

We see that after a certain number of sequences the HMM stops learning, which is particular to the symmetric initial student and disappears for a non-symmetric one.

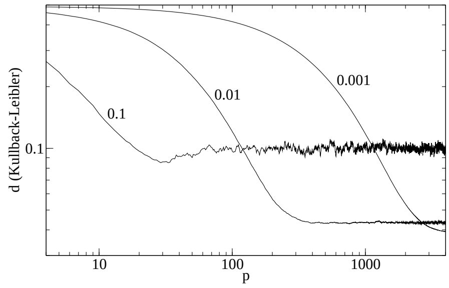

FIGURE 1. Log-log curves of BWO for three different η_{BW} indicated next to the curves.

Denoting the variation of the parameters in BC by Δ, in BW by $\hat{\Delta}$, in BWO by $\tilde{\Delta}$, and with $\gamma_t(i) \equiv \mathcal{P}(q_t = s_i|y^p, \omega^p)$, we have to first order in λ

$$\begin{aligned} \Delta\pi_i &= \frac{\lambda\eta_{BC}}{n}\hat{\Delta}\pi_i = \frac{\lambda}{n}\frac{\eta_{BC}}{\eta_{BW}}\tilde{\Delta}\pi_i, \\ \Delta A_{ij} &= \frac{\lambda\eta_{BC}}{n}\left[\sum_{t=1}^{T-1}\gamma_t(i)\right]\hat{\Delta}A_{ij} = \frac{\lambda}{n}\frac{\eta_{BC}}{\eta_{BW}}\left[\sum_{t=1}^{T-1}\gamma_t(i)\right]\tilde{\Delta}A_{ij}, \\ \Delta B_{i\alpha} &= \frac{\lambda\eta_{BC}}{n}\left[\sum_{t=1}^{T}\gamma_t(i)\right]\hat{\Delta}B_{i\alpha} = \frac{\lambda}{n}\frac{\eta_{BC}}{\eta_{BW}}\left[\sum_{t=1}^{T}\gamma_t(i)\right]\tilde{\Delta}B_{i\alpha}. \end{aligned} \tag{4}$$

For $\eta_{BW} \approx \lambda\eta_{BC}/n$ and small λ, variations in BC are proportional to those in BWO, but with different effective learning rates for each matrix depending on y^p. Simulations show that actual values are of the same order of approximated ones.

THE BAYESIAN ONLINE ALGORITHM

The Bayesian Online Algorithm (BOnA) [8] uses Bayesian inference to adjust ω in the HMM using a data set $D_P = \{y^1, ..., y^P\}$. At each sequence we update a prior distribution by Bayes' theorem, which takes a prior in a parametric family to a posterior no longer in it. BOnA projects it back by minimizing its KL-divergence with the projected distribution. Parameters are estimated as the means in the projected distribution.

For a family of the form $P(x) \propto e^{-\sum_i \lambda_i f_i(x)}$, which is obtained by MaxEnt constraining the averages over $P(x)$ of arbitrary functions $f_i(x)$, minimizing the KL-divergence is equivalent to equating these averages to those over the unprojected posterior.

For HMMs, the vector π and each row of A and B are different discrete distributions which we assume independent in order to write the factorized distribution

$$\mathcal{P}(\omega|u) \equiv \mathcal{P}(\pi|\rho)\prod_{i=1}^{n}\mathcal{P}(A^i|a^i)\mathcal{P}(B^i|b^i). \tag{5}$$

where $A^i \equiv (A_{i1}, ..., A_{in})$, $B^i \equiv (B_{i1}, ..., B_{im})$ and $u = (\rho, a, b)$ represents the parameters of the distributions.

As each factor is a distribution over probabilities, the natural choice are the Dirichlet distributions, which for a N-dimensional variable x is

$$\mathcal{D}(x|u) = \frac{\Gamma(u_0)}{\prod_{i=1}^{N}\Gamma(u_i)}\prod_{i=1}^{N} x_i^{u_i-1}, \tag{6}$$

with $u_0 = \sum_i u_i$. These can be obtained from MaxEnt with $f_i(x) = \ln x_i$ [13]:

$$\int d\mu\, \mathcal{D}(x)\ln x_i = \alpha_i, \qquad d\mu \equiv \delta\left(\sum_i x_i - 1\right)\prod_i \theta(x_i)dx_i. \tag{7}$$

The function to be extremized is

$$\mathcal{L} = \int d\mu\, \mathcal{D}\ln\mathcal{D} + \lambda\left(\int d\mu\,\mathcal{D} - 1\right) + \sum_i \lambda_i\left(\int d\mu\,\mathcal{D}\ln x_i - \alpha_i\right), \tag{8}$$

and with $\delta\mathcal{L}/\delta\mathcal{D} = 0$ we get the Dirichlet with normalization $e^{\lambda+1}$ and $u_i = 1-\lambda_i$.

Each factor distribution is separately projected by equating the average of the logarithms in the original posterior Q and in the projected distributions

$$\begin{aligned}
\psi(\rho_i) - \psi\left(\sum_j \rho_j\right) &= \langle \ln \pi_i\rangle_Q \equiv \mu_i(\rho), \qquad (9)\\
\psi(a_{ij}) - \psi\left(\sum_k a_{ik}\right) &= \langle \ln A_{ij}\rangle_Q \equiv \mu_{ij}(a),\\
\psi(b_{i\alpha}) - \psi\left(\sum_\beta b_{i\beta}\right) &= \langle \ln B_{i\alpha}\rangle_Q \equiv \mu_{i\alpha}(b),
\end{aligned}$$

where $\psi(x) = d\ln\Gamma(x)/dx$ is the digamma function. We call a set of N equations

$$\psi(x_i) - \psi\left(\sum_j x_j\right) = \mu_i, \tag{10}$$

with $i = 1, ...N$ a *digamma system* in the variables x_i with coefficients μ_i.

Let us call $P^p(\omega)$ the projected distribution after observation of y^p, and $Q^{p+1}(\omega)$ the posterior distribution (not projected yet) after y^{p+1}. By Bayes' theorem,

$$Q^{p+1}(\omega) = \frac{1}{Z_Q}P^p(\omega)\sum_{q^{p+1}}\mathcal{P}(y^{p+1}, q^{p+1}|\omega), \tag{11}$$

where Z_Q is the normalization.

The calculation of μ's in (9) leads to averages over Dirichlets of the form [10]

$$\mu_i = \left\langle \left[\prod_j x_j^{r_j}\right]\ln x_i\right\rangle = \frac{\Gamma(u_0)}{\prod_j\Gamma(u_j)}\frac{\prod_j\Gamma(u_j+r_j)}{\Gamma(u_0+r_0)}[\psi(u_i+r_i) - \psi(u_0+r_0)]. \tag{12}$$

In order to solve (10), we solve for x_i, sum over i with $x_0 \equiv \sum_i x_i$ and write it as a one-dimensional map

$$x_0^{n+1} = \sum_i \psi^{-1}[\mu_i + \psi(x_0^n)], \tag{13}$$

finding numerically the fixed point by iterating from an arbitrary initial point. We found a unique fixed point except for μ_i's too close to 0, which is rare in most applications.

BOnA suffers from a common problem of Bayesian algorithms: due to the sum over hidden variables, the complexity scales exponentially in T. Also, the calculation of several digamma functions is very time consuming. In the next section, we develop an approximation that runs faster, although still with exponential complexity in T, which is not a problem for we can fix T with the algorithm scaling polynomially in n.

MEAN POSTERIOR APPROXIMATION

The Mean Posterior Approximation (MPA) is a simplification of BOnA inspired in its results for gaussians, where we match the first and second moments of posterior and projected distributions. Noting this, instead of minimizing d_{KL} we just match the mean and one of the variances of posterior and projected distributions as an approximation.

With hatted variables for reestimated values, the matching of the moments gives [10]

$$\begin{aligned}
\hat{\rho}_i &= \langle \pi_i \rangle_Q \frac{\langle \pi_1 \rangle_Q - \langle \pi_1^2 \rangle_Q}{\langle \pi_1^2 \rangle_Q - \langle \pi_1 \rangle_Q^2}, \\
\hat{a}_{ij} &= \langle a_{ij} \rangle_Q \frac{\langle a_{i1} \rangle_Q - \langle a_{i1}^2 \rangle_Q}{\langle a_{i1}^2 \rangle_Q - \langle a_{i1} \rangle_Q^2}, \\
\hat{b}_{i\alpha} &= \langle b_{i\alpha} \rangle_Q \frac{\langle b_{i1} \rangle_Q - \langle b_{i1}^2 \rangle_Q}{\langle b_{i1}^2 \rangle_Q - \langle b_{i1} \rangle_Q^2}.
\end{aligned} \tag{14}$$

The complexity is again of order n^T, but the simplifications heavily reduce the real computational time making it better for practical applications.

Figure 2 compares MPA and BOnA. The initial difference gets smaller with time and both come closer relatively fast. We used $n = 2$, $m = 3$ and $T = 2$ and averaged over 150 random teachers with symmetric initial students. The computational time for BOnA was 340min, and for MPA, 5s in a 1GHz processor. Figure 3a compares MPA to BC and figure 3b to BWO. In both cases MPA has superior generalization. We used $n = 2$, $m = 3$, $T = 2$, symmetric initial students and took averages over 500 random teachers.

LEARNING DRIFTING CONCEPTS

We tested BC and MPA for changing teachers. In figure 4a, the teacher changes at random after each 500 sequences ($\lambda = 0.01$, $\eta_{BC} = 10.0$). In figure 4b, each time a sequence is observed, a small random quantity is added to the teacher. Both simmulations used $n = 2$, $m = 3$ and were averaged over 200 runs.

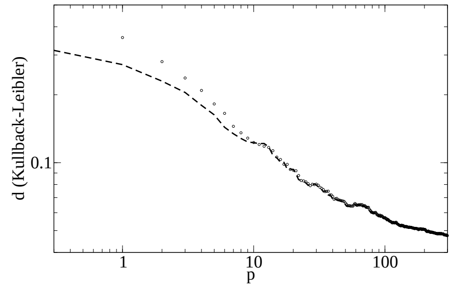

FIGURE 2. Comparison in log-log scale of MPA (dashed line) and BOnA (circles).

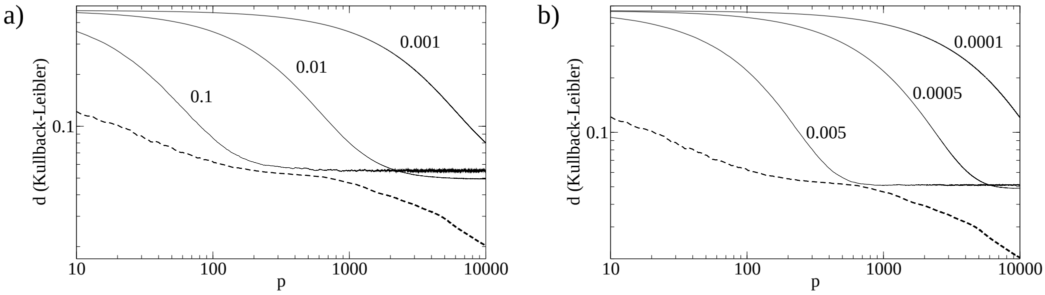

FIGURE 3. a) Comparison between MPA (dashed) and BC (continuous). Values of λ are indicated next to the curves. $\eta_{BC} = 0.5$. b) Comparison between MPA (dashed) and BWO (continuous). Values of η_{BW} are indicated next to the curves. Both scales are log-log.

Figure 4b shows that BC adapts better, but is not *fully* adaptive and we do not know how to modify it. MPA instead derives from Bayesian principles and we can guess the problem by analogy with similar Bayesian algorithms [12]: the variance of the distributions decreases in the process as in the perceptron case, where they turn out to be the learning rates, explaining the memory effect which difficults the learning after changes. Although we cannot prove it yet, we expect that variance and learning rate are similarly related in MPA, which can be used to improve its performance.

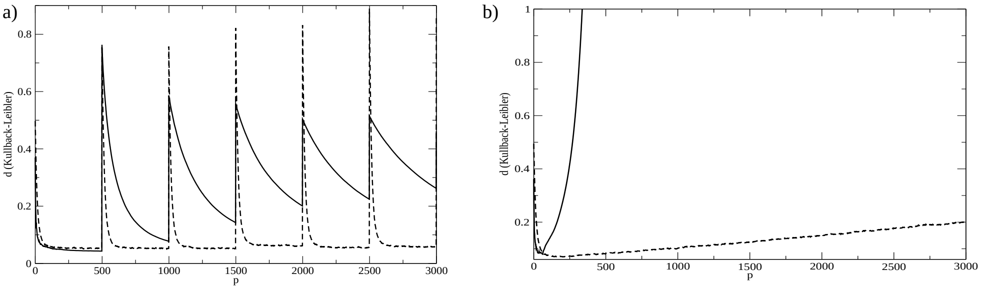

FIGURE 4. Drifting concepts. Continuous lines correspond to MPA and dashed lines to BC. a) Abrupt changes at 500 sequences interval. b) Small random changes at each new sequence.

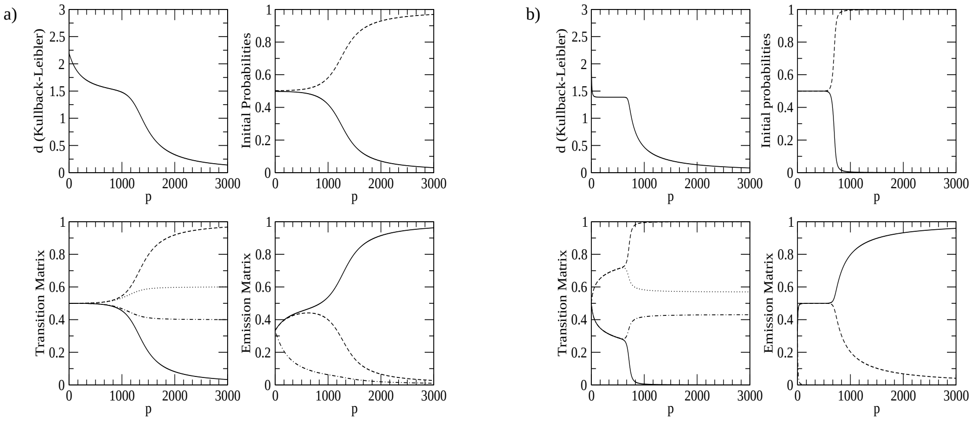

FIGURE 5. KL-divergence and student's parameters for a) BC and b) MPA.

LEARNING AND SYMMETRY BREAKING

Learning from symmetric initial configurations requires that in some point the student parameters break appart from each other and the student's symmetry is broken, marked as a sharp decrease in the generalization error. This point depends on the algorithm and is an important feature in *online* algorithms [11].

Instead of taking averages which smooth out abrupt changes, here we draw curves for only one teacher, rendering the changes visible. Flat pieces before a symmetry breaking are called *plateaux* and occur in situations where it is difficult to break the symmetry.

Figure 5a shows the results for BC ($\lambda = 0.01$, $\eta_{BC} = 1.0$). There are two abrupt changes in d_{KL}: at the beginning of the process and after 1000 sequences. π and A only break their symmetry in the second, while B breaks it at both points. Figure 5b shows that in MPA the second change is stronger and the symmetry breaking affects both B and A. Figure 6 shows BWO with $\eta_{BW} = 0.01$ where only B breaks its symmetry. Note that the more symmetries are broken, the best is the generalization of the algorithm.

In all simulations we set $n = 2$, $m = 3$ and $T = 2$ with a teacher HMM given by

$$\pi = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \qquad A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \qquad B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{15}$$

CONCLUSIONS

We proposed and analyzed three learning algorithms for HMMs: Baum-Welch Online (BWO), Bayesian Online Algorithm (BOnA) and Mean Posterior Approximation (MPA). We showed that when the teacher does not change, MPA has superior performance, but for drifting concepts, the Baldi-Chauvin (BC) algorithm is better, although the Bayesian nature of MPA suggests how to fix this behavior.

The importance of symmetry breaking in learning processes is presented here in a brief discussion where the phenomenon is shown to occur in our models.

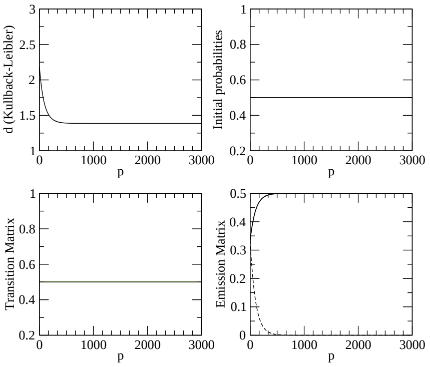

FIGURE 6. KL-divergence and student's parameters for BWO.

Preliminary studies on real data seem to confirm the performance of the algorithms.

ACKNOWLEDGEMENTS

We would like to thank Evaldo Oliveira, Manfred Opper and Lehel Csato for useful discussions. This work was made part in the University of São Paulo with financial support of FAPESP and part in the Aston University with support of Evergrow Project.

REFERENCES

1. Y. Ephraim, N. Merhav, Hidden Markov Processes. IEEE Trans. Inf. Theory **48**, 1518-1569 (2002).
2. L. R. Rabiner, A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proc. IEEE **77**, 257-286 (1989).
3. P. Baldi, S. Brunak, Bioinformatics: The Machine Learning Approach. MIT Press (2001).
4. R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge (1998).
5. J. Garcia-Frias, Deconding of Low-Density Parity-Check Codes Over Finite-State Binary Markov Channels. IEEE Trans. Comm. **52**, 1840-1843 (2004).
6. P. Baldi, Y. Chauvin, Smooth On-Line Learning Algorithms for Hidden Markov Models. Neural Computation **6**, 307-318 (1994).
7. S. Amari, Neural learning in structured parameter spaces - Natural Riemannian gradient. NIPS'96 **9**, MIT Press (1996).
8. M. Opper, A Bayesian Approach to On-line Learning. On-line learning in Neural Networks, edited by D. Saad, Publications of the Newton Institute, Cambridge Press, Cambridge (1998).
9. M. Opper, D. Saad, Advanced Mean Field Methods: Theory and Practice. MIT Press (2001).
10. R. Alamino, N. Caticha, Bayesian Online Algorithms for Learning in Discrete Hidden Markov Models. Submitted to Discrete and Continuous Dynamical Systems.
11. T. Heskes, W. Wiegerinck, W., On-line Learning with Time-Correlated Examples. On-line Learning in Neural Networks, 251-278, edited by David Saad, Cambridge University Press, Cambridge (1998).
12. R. Vicente, O. Kinouchi, N. Caticha. Statistical Mechanics of Online Learning of Drifting Concepts: A Variational Approach. Machine Learning **32**, 179-201 (1998).
13. M. O. Vlad, M. Tsuchiya, P. Oefner, J. Ross. Bayesian analysis of systems with random chemical composition: Renormalization-group approach to Dirichlet distributions and the statistical theory of dilution. Phys. Rev. E **65**, 011112(1)-01112(8) (2001).