QUASI-NEWTON METHOD FOR MAXIMUM LIKELIHOOD ESTIMATION OF HIDDEN MARKOV MODELS

Olivier Cappé, Vincent Buchoux, Eric Moulines

ENST Dpt. Signal / CNRS URA 820
46 rue Barrault, 75634 Paris Cedex 13, France
email: cappe, buchoux, moulines @sig.enst.fr

ABSTRACT

Hidden Markov models (HMMs) are used in many signal processing applications including speech recognition, blind equalization of digital communications channels, etc. The most widely used method for maximum likelihood estimation of HMM parameters is the forward-backward (or Baum-Welch) algorithm which is an early example of application of the Expectation-Maximization (EM) principle. In this contribution, an alternative fast-converging approach for maximum likelihood estimation of HMM parameters is described. This new techniques is based on the use of general purpose quasi-Newton optimization methods as well as on an efficient purely recursive algorithm for computing the log-likelihood and its derivative.

1. INTRODUCTION

Hidden Markov models (HMMs) are used in many applications, including (among many others) speech processing [9], digital communications [3] or biological signal processing [1]. As of today, the most efficient, and by far the most popular, approach for estimating the parameters of HMMs is based on the Expectation-Maximization (EM) principle [9], [6]. The EM algorithm, as formulated by Dempster *et al.* (1977) [2], is an iterative optimization method for maximum likelihood estimation of statistical models that involve unobserved (or latent) data. The application of EM to Hidden Markov Models is not straightforward and requires an additional inductive algorithm known as the *forward-backward*, introduced by Baum *et al.* (1970) [9], [6].

The main advantages of EM are its immediate applicability to a large class of statistical models as well as its ease of implementation. On the other hand, there is no reason why EM should be systematically preferred to other alternative optimization approaches (as was already pointed out by several discutants the original paper by Dempster *et al.* [2]). In particular, EM is known to converge very slowly in some models [4], [10], [6]. Many of the solutions proposed to speed up the convergence of EM are based on the observation that EM provides an indirect way of computing the gradient of the log-likelihood according to the Fisher relation

$$\nabla_\theta \log p(x;\theta)|_{\theta_0} = \nabla_\theta \mathrm{E}\left[\log p(x,y;\theta)|x;\theta_0\right]|_{\theta_0} \tag{1}$$

where x denotes the observed data, y the unobserved data and θ the parameters of the model [4], [6].

For HMMs however, the use of (1) implies a complete iteration of the forward backward procedure introduced by Baum *et al*. In the present work, we describe an alternative recursive procedure for computing the log-likelihood, which is attractive from a computational point of view. The paper is organized as follows: in sections 1 and 2, the computation of the log-likelihood and its derivatives is detailed; sections 3 and 4 are devoted to the comparison of the proposed method with the standard EM-based approach.

2. COMPUTING THE LIKELIHOOD

Let us consider an HMM with vector valued observations $\mathbf{x}_t$ and associated unobservable state $s_t \in \{1,\ldots,N\}$, where N denotes the number of states. $\mathbf{\Pi}$: $\Pi_{ij} \triangleq P(s_{t+1} = j|s_t = i)$ is the transition matrix associated with the Markov process and $f_i(\mathbf{x}_t)$ denotes the state conditional distribution $p(\mathbf{x}_t|s_t = i)$. The notation $\mathbf{X}_r^s$ will be used as a shorthand for the subsequence $(\mathbf{x}_r, \mathbf{x}_{r+1}, \ldots \mathbf{x}_s)$. In the following, the dependence upon the HMM parameters is implicit in all expressions and omitted for notational simplicity. For a sequence of observations of length T, Bayes rules can be used to rewrite the log-likelihood as

$$\log p(\mathbf{X}_1^T) = \sum_{t=1}^{T} \log p(\mathbf{x}_t|\mathbf{X}_1^{t-1}) \tag{2}$$

where $p(\cdot|\mathbf{X}_1^0)$ is by convention equivalent to the unconditional distribution $p(\cdot)$. The HMM structure makes it possible to rewrite (2) as

$$\log p(\mathbf{X}_1^T) = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{N} p(\mathbf{x}_t, s_t = i | \mathbf{X}_1^{t-1}) \right] \\ = \sum_{t=1}^{T} \log \left[\sum_{i=1}^{N} f_i(\mathbf{x}_t) \phi_t(i) \right] \qquad (3)$$

where $\phi_t(i) \triangleq P(s_t = i | \mathbf{X}_1^{t-1})$ denotes the state prediction filter. $\phi_t(i)$ can be updated recursively using (see [11])

$$\phi_1(j) = p(s_1 = j) \\ \phi_{t+1}(j) = \frac{1}{c_t} \sum_{i=1}^{N} f_i(\mathbf{x}_t) \phi_t(i) \Pi_{ij} \quad (t \geq 1) \qquad (4)$$

The normalization factors c_t are given by

$$c_t = \sum_{k=1}^{N} f_k(\mathbf{x}_t) \phi_t(k) \qquad (5)$$

Except for the fact that it is normalized, (4) bear close resemblance with the formulas used for updating the forward variable $\alpha_t(i) \triangleq p(\mathbf{X}_1^t, s_t = i)$ in the Baum-Welch recursions [9], [6].

3. COMPUTING THE GRADIENT

Formulas for computing the derivatives of the log-likelihood can be obtained by simply differentiating (3) and (4). The form of the log-likelihood given by (3) has been used before as a technique for proving the consistence and asymptotic normality of the maximum likelihood estimate in HMMs (see the references in [7]), however the idea of using (3) to effectively compute the gradient of the likelihood is due to Mevel and LeGland [7], [5].

Usually the HMM parameter vector can be splitted in two distinct parts: Parameters $\boldsymbol{\theta}$ of state-conditional distributions $f_i(\cdot)$ and transition parameters that define $\mathbf{\Pi}$. For simplicity we omit here the question of the initial state distribution and assume that it is known *a priori*.

For the distributional parameters $\boldsymbol{\theta}$, the gradient of the log-likelihood is given by

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \log p(\mathbf{X}_1^T) = \sum_{t=1}^{T} \frac{1}{c_t} \times \\ \sum_{i=1}^{N} \left[\phi_t(i) \boldsymbol{\nabla}_{\boldsymbol{\theta}} f_i(\mathbf{x}_t) + f_i(\mathbf{x}_t) \boldsymbol{\nabla}_{\boldsymbol{\theta}} \phi_t(i) \right] \qquad (6)$$

Where $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \phi_t(i)$ can be updated recursively using

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}} \phi_{t+1}(j) = \frac{1}{c_t} \sum_{i=1}^{N} \left(\Pi_{ij} - \phi_{t+1}(j) \right) \times \\ \left[\phi_t(i) \boldsymbol{\nabla}_{\boldsymbol{\theta}} f_i(\mathbf{x}_t) + f_i(\mathbf{x}_t) \boldsymbol{\nabla}_{\boldsymbol{\theta}} \phi_t(i) \right] \qquad (7)$$

For the transition parameters $\boldsymbol{\eta}$, the gradient of the log-likelihood is given by

$$\boldsymbol{\nabla}_{\boldsymbol{\eta}} \log p(\mathbf{X}_1^T) = \sum_{t=1}^{T} \frac{1}{c_t} \sum_{i=1}^{N} f_i(\mathbf{x}_t) \boldsymbol{\nabla}_{\boldsymbol{\eta}} \phi_t(i) \qquad (8)$$

Where $\boldsymbol{\nabla}_{\boldsymbol{\eta}} \phi_t(i)$ can be updated recursively using

$$\boldsymbol{\nabla}_{\boldsymbol{\eta}} \phi_{t+1}(j) = \frac{1}{c_t} \sum_{i=1}^{N} f_i(\mathbf{x}_t) \times \\ \left[\left(\Pi_{ij} - \phi_{t+1}(j) \right) \boldsymbol{\nabla}_{\boldsymbol{\eta}} \phi_t(i) + \phi_t(i) \boldsymbol{\nabla}_{\boldsymbol{\eta}} \Pi_{ij} \right] \qquad (9)$$

(7) and (9) both involve the state predictor filter $\phi_t(i)$ at time index $t-1$ and t as well as its gradient at time index $t-1$. The recursions should be started with $\boldsymbol{\nabla}_{\boldsymbol{\theta}} \phi_1(i) = \boldsymbol{\nabla}_{\boldsymbol{\eta}} \phi_1(i) = \mathbf{0}$ in the case where the initial state distribution is known.

The above recursions, when used for computing the gradient of the log-likelihood have a numerical complexity proportional to $T \times N^2$, which is comparable with that of the forward-backward [6]. Whatever the length of the sequence, the storage space needed to compute the gradient of the log-likelihood is constant and proportional to $N \times p$ where p is the total number of parameters, whereas the storage space needed for the forward-backward is proportional to $N \times T$. This feature is most useful in situations where dealing with long observation sequences is unavoidable as for the analysis of DNA sequences [1] or blind equalization for communication channels [3].

The proposed method for estimating the HMM parameters simply consist in using a general purpose quasi-Newton optimization algorithm [8], where the log-likelihood and its gradient are evaluated recursively using the equations given above. Such algorithms are based on the Newton approximation formula, and the "quasi" prefix corresponds to the fact that one has to use embedded line searches to adjust the step size at each iteration [8].

4. COMPARISON WITH THE EM BASED APPROACH

We first briefly review the main convergence properties of the EM algorithm in a general context: Under suitable regularity conditions, the main numerical characteristics of the EM algorithm are:

1. Global convergence (in Zangwill's sense) in the set of stationary points of the likelihood [12].

2. When the algorithm converges (pointwise) to $\boldsymbol{\theta}^*$, a first order approximation yields the following update formula [4]

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \mathbf{I}_c(\boldsymbol{\theta}^*) \; \boldsymbol{\nabla}_{\boldsymbol{\theta}} \log p(\mathbf{X}_1^T; \boldsymbol{\theta}) \big|_{\boldsymbol{\theta}_n} \qquad (10)$$

where $\mathbf{I}_c(\boldsymbol{\theta}^*)$ is the so-called *complete data information matrix* (see [2] for its precise definition).

Eq. (10) indicates that at convergence, each iteration of the EM algorithm closely resembles a step of the Newton algorithm for maximizing the log-likelihood, except for the fact that $\mathbf{I}_c(\boldsymbol{\theta}^*)$ is the "complete data information matrix" and not the observed data information matrix $\mathbf{I}(\boldsymbol{\theta}^*)$, defined as the inverse of the Hessian of the log-likelihood. Because of this mismatch in the descent direction, in practice the EM performs like a gradient descent algorithm (linear convergence) with a low speed of convergence [4], [10].

In sharp contrast, the proposed method reaches superlinear convergence when it falls in the attraction domain of one of the local maxima of the likelihood [8] (see next section). One drawback of the proposed quasi-Newton approach is that it makes it necessary to deal with possible constraint on the HMM parameters. The most obvious constraint is that $\mathbf{\Pi}$ must be a stochastic matrix, where each line contains positive elements which sum to one. Usual solutions to this problem include the reparameterization of each line of the transition matrix as a point on the $\mathbb{R}^N$ hypersphere represented by $N-1$ angles [11]. In the following section, we use the computationally simpler solution which consists in using the natural transition matrix parameterization and *(i)* projecting the gradient on the subspace orthogonal to the linear constraints $\sum_{j=1}^{N} \Pi_{ij} = 1$ $(1 \leq i \leq N)$, *(ii)* take into account the parameters bounds $(0 \leq \Pi_{ij} \leq 1)$ when determining the step size. This second method has the disadvantage that it may exhibit poor convergence if the true parameter value lies close to the inequality constraints boundaries (if one of the transition probabilities is very small for instance). For the distribution parameters, it is possible in most cases to get rid of the constraints by a mere reparameterization of the model as is demonstrated in next section.

5. RESULTS

We illustrate the behavior of the proposed method in the case where the state-conditional distributions are Gaussian. For reason of simplicity, the observation are assumed to be scalar, but the same computation scheme could be used straightforwardly if the state-conditional distributions were multivariate normal with diagonal covariance matrices. The distribution parameters are $\boldsymbol{\theta} = (\mu_1, \ldots, \mu_N, \kappa_1, \ldots, \kappa_N)'$ where μ_i is the mean value corresponding to state i and $\kappa_i = \log \sigma_i$ is the logarithm of the standard deviation for the same state. The use of κ_i rather than σ_i makes it possible to treat the optimization with respect to $\boldsymbol{\theta}$ as an unconstrained optimization problem. For this model, the term $\boldsymbol{\nabla}_{\boldsymbol{\theta}} f_i(x_t)$ featured in (6) and (7) only has two non-null components corresponding to:

$$\frac{\partial f_i(x_t)}{\partial \mu_i} = \frac{(x_t - \mu_i)}{\sigma_i^2} f_i(x_t) \qquad (11)$$

and

$$\frac{\partial f_i(x_t)}{\partial \kappa_i} = \frac{1}{\sigma_i} \left[\left(\frac{x_t - \mu_i}{\sigma_i} \right)^2 - 1 \right] f_i(x_t) \qquad (12)$$

To evaluate the proposed method, we use data simulated from a 3 states HMM with known parameters:

$$\boldsymbol{\mu} = \begin{bmatrix} -2 \\ 1 \\ 5 \end{bmatrix} \quad \boldsymbol{\sigma} = \begin{bmatrix} 1 \\ 1 \\ 3.3 \end{bmatrix}$$

$$\mathbf{\Pi} = \begin{bmatrix} 0.7 & 0.1 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.2 & 0.5 \end{bmatrix}$$

The stationary probability distribution (marginal probability distribution of the observations for large value of the time index t) corresponding to the actual parameters is plotted as the solid curve on fig. 1 together with those corresponding to the initial value of the parameters (dotted line) and to the MLE (dashed line). Note that the initial guess of the HMM parameters was chosen so that both iterative procedures converge to the MLE (and not to a local maxima).

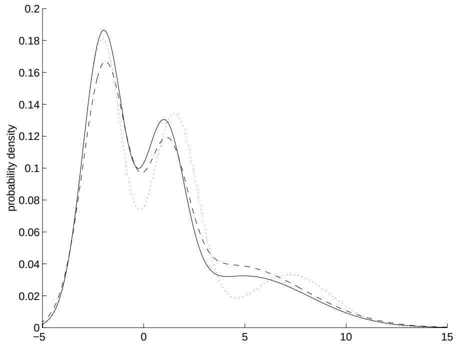

Figure 1: Stationary probability distributions corresponding to true parameters (solid line), initial parameters (dotted line), MLE (dashed line).

Fig. 2 displays the relative distance (measured using L^2 norm) between the estimated parameters and the MLE as a function of the iteration index. The curve corresponding to EM (star signs) is approximately linear for large iterations indexes, which denotes linear convergence of the algorithm. The slope of this curve (in the rightmost part of the plot) is 0.94 which confirms that EM does converge very slowly even for small-size HMMs. As expected, the quasi-Newton optimization scheme (circles) converges much faster. Of course, each iteration of the proposed algorithm is more complex than a single step of EM because it requires the computation of the gradient through (6)-(9) as well as several evaluations (usually 3 when quadratic interpolation line-search algorithms are used [8]) of the log-likelihood. On the other-hand, fig. 2 shows that it is virtually impossible to approach

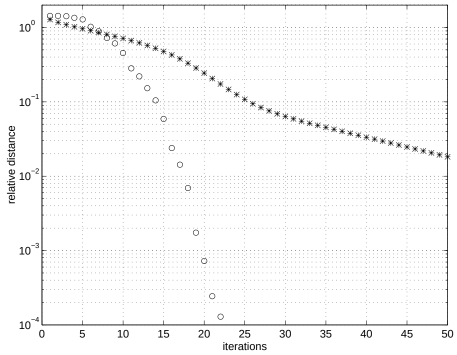

Figure 2: Relative distance to the MLE for EM (stars) and quasi-Newton optimization (circles).

the MLE with a reasonable precision when using EM (25 iterations for a 10% error and approximately 65 for a 1% error).

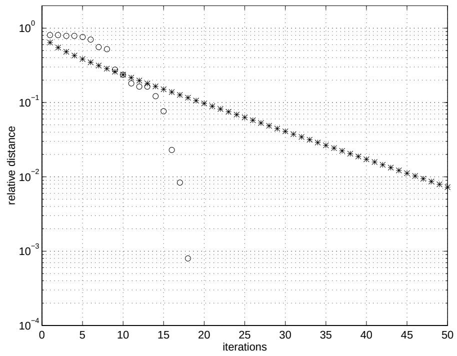

Figure 3: Relative distance to the MLE for EM (stars) and quasi-Newton optimization (circles), for an observation sequence of length 2000.

Fig. 3 correspond to the same model with an observation sequence 10 times longer (2000 observations). With this data length, the speed of convergence of EM is slightly better but still stays close to 1 (measured value of 0.91). Looking at both figs. 2 and 3, it is clear that even for large sample sizes, the proposed quasi-Newton optimization scheme largely outperforms EM in terms of convergence speed and accuracy.

6. CONCLUSION

The HMM training method proposed in this paper is based on a standard well-documented class of algorithms which are readily available in most numerical optimization packages. For HMMs, the use of such fast-converging algorithm appears to be a promising alternative to standard EM training. The proposed approach also makes it possible to carry out all computations using a single forward scan through the data at each iteration which is preferable in cases where the training sequence is long. Finally, this approach can be extended, more naturally than EM, to adaptive estimation tasks (see [5]).

7. REFERENCES

[1] G. Churchill. Hidden markov chains and the analysis of genome structure. *Computers & chemistry*, 16(2):107–1115, 1992.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 39(1):1–38 (with discussion), 1977.

[3] G. K. Kaleh and R. Vallet. Joint parameter estimation and symbol detection for linear or non-linear unknown channels. *IEEE Trans. Communications*, 42(7), 1994.

[4] K. Lange. A gradient algorithm locally equivalent to the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 57(2):425–437, 1995.

[5] F. LeGland and L. Mevel. Recursive estimation in HMMs. In *36th IEEE Conf. on decision and control*, San Diego, 1997.

[6] I. L. MacDonald and W. Zucchini. *Hidden Markov models and other models for discrete-valued time series*. Chapman & Hall, 1997.

[7] L. Mevel. *Statistique asymptotique pour les modèles de Markov caché.* PhD thesis, Université de Rennes 1, 1997.

[8] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical recipes in C : the art of scientific computing*. Cambridge University Press, second edition, 1992.

[9] L. R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–285, February 1989.

[10] R. A. Redner and H. F. Walker. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239, April 1984.

[11] J.B. Moore R.J. Elliot, L. Aggoun. *Hidden Markov models: Estimation and control*. Springer-Verlag, New York, 1994.

[12] C. F. J. Wu. On the convergence properties of the EM algorithm. *Annals of Statistics*, 11(1):95–103, 1983.