RECURSIVE COMPUTATION OF THE SCORE AND OBSERVED INFORMATION MATRIX IN HIDDEN MARKOV MODELS

Olivier Cappé & Eric Moulines

Centre National de la Recherche Scientifique & Ecole Nationale Supérieure des Télécommunications
46 rue Barrault, 75634 Paris cedex 13, France
cappe at enst.fr, moulines at enst.fr

ABSTRACT

Hidden Markov Models (henceforth abbreviated to HMMs), taken in their most general acception, that is, including models in which the state space of the hidden chain is continuous, have become a widely used class of statistical models with applications in diverse areas such as communications, engineering, bioinformatics, econometrics and many more. This contribution focus on the computation of derivatives of the log-likelihood and proposes a (comparatively!) simple and general framework, based on the use of Fisher and Louis identities, to obtain recursive equations for computing the score and observed information matrix. This approach is thought to be simpler than (although equivalent to) the solution provided by the so-called sensitivity equations. It is based on the original remark that recursive smoothers for HMMs are also available for some functionals of the hidden states which do not reduce to sum functionals. This view of the problem also suggests ways in which these exact equations could be approximated using sequential Monte Carlo methods.

KEYWORDS: Hidden Markov Models, Score, Information Matrix, Smoothing, Recursive Computation

1. INTRODUCTION

Hidden Markov models constitute a very important class of probabilistic signal models. They have been used, sometimes with great success, to model finite-valued data with finite-valued underlying states (digital communications, bioinformatics), (multivariate) continuous sequences with finite-valued underlying states (speech recognition) or continuous sequences with continuous hidden spaces (tracking, econometrics, etc.) Interestingly, it is only recently, with a renewed interest in the theory of general HMMs starting from the early 1990s, that the solidarity between the different variants of HMMs considered in pioneering works such as [1] (finite-valued HMMs) and [2] (Gaussian linear state-space models) has become clearer.

We focus here on the computation of derivatives of the log-likelihood in general HMMs with emphasis on the score (first derivative) and observed information matrix (opposite of the second derivative). Although, parameter estimation obviously is a strong motivation for computing such quantities (the score in particular) we would like to stress that it is in no way the only motivation. First, there exists a well-known likelihood optimization technique which does not require the computation of the score, namely the Expectation-Maximization (EM) algorithm [3], which is *de facto* standard for training HMMs parameters. Second, the score and observed information are important for other tasks such as hypothesis testing – "does a pre-specified model fits my data correctly?", change detection, etc.

In the following we discuss recursive techniques for evaluating the score and observed information matrix, where "recursive" means that each new observation may be taken into account with a non-increasing (with the observation index) computational effort, both in terms of actual computations and memory requirements. It is precisely one of the strongest appeal of HMMs that, due to the assumed Markov dependence of the hidden states, all quantities of interest may be evaluated recursively using efficient algorithms. Recursivity is also essential because it opens an opportunity for on-line (as opposed to batch) processing which is of prime importance in many applications, particularly in signal processing. Finally, defining quantities of interest recursively is also required in order to use sequential Monte Carlo techniques [4] which have recently emerged as a powerful tool for handling general HMMs for which exact computation is not feasible.

2. NOTATIONS AND BASIC DEFINITIONS

Although we limit our discussion to the case of HMMs, the techniques discussed here also apply for more general models such as Markov switching autoregressive models [5]. A hidden Markov model is such that

1. $\{X_k\}_{k\geq 0}$ is Markovian with initial distribution ν and transition density function q such that, for any function f,

$$\mathrm{E}[f(X_0)] = \int f(x)\,\nu(x)dx$$

and

$$\mathrm{E}[f(X_n)|X_{0:n-1}] = \int f(x)\,q(X_{n-1},x)dx$$

where $X_{0:n-1}$ is a concise notation for the collection of variables $X_0,\ldots,X_{n-1}$. The state space of the hidden chain will be denoted by X. In the following, we assume that all probability distributions admits densities with respect to a measure on X which we simply denote by dx. This convention is used only for notational simplicity and the technique described here are valid in fairly general state spaces [5].

2. $\{Y_k\}_{k\geq 0}$ is conditionally independent given $\{X_k\}_{k\geq 0}$ with (marginal) transition density function g such that, for arbitrary functions $f_0,\ldots,f_n$,

$$\mathrm{E}\left[\prod_{k=0}^{n} f_k(Y_k)\middle| X_{0:n}\right] = \prod_{k=0}^{n}\int f_k(y)\,g(X_k,y)dy$$

where g is sometimes referred to as the *(conditional) likelihood function.* In the following, we always consider g as a function of its first argument only and write $g_k(x) = g(x, Y_k)$ for the conditional likelihood function evaluated in Y_k.

3. Since we consider models with an unknown parameter vector θ, quantities that do depend on the parameter will be marked using θ as a superscript.

It is well-known that[1] the log-likelihood $\ell_n^\theta \stackrel{\text{def}}{=} \log \mathrm{L}_n^\theta$ of the observations $Y_{0:n}$ in this model may be evaluated as

$$\begin{aligned}\ell_n^\theta &= \log \mathrm{L}_0^\theta + \sum_{k=1}^{n} \log \frac{\mathrm{L}_k^\theta}{\mathrm{L}_{k-1}^\theta} \qquad (1)\\ &= \log \int \nu^\theta(x) g_0^\theta(x)\\ &\quad + \sum_{k=1}^{n} \log \left[\iint \phi_{k-1|k-1}^\theta(x) q^\theta(x, x') g_k^\theta(x') dx dx' \right]\end{aligned}$$

where $\phi_{k|k}^\theta$ is the *filtering* probability density function associated to the distribution of the hidden state X_k given the observations $Y_{0:k}$ up to index k (and for the parameter value θ). The filtering probability density function $\phi_{k|k}^\theta$ may be evaluated recursively using the equation

$$\phi_{k|k}^\theta(x_k) \propto \int \phi_{k|k}^\theta(x_{k-1}) q^\theta(x_{k-1}, x_k) g_k^\theta(x_k) dx_{k-1} \qquad (2)$$

where the normalization factor is precisely the term that appears between brackets in (1), which may also be interpreted as the likelihood ratio $\mathrm{L}_k^\theta / \mathrm{L}_{k-1}^\theta$ [5, 6, 7].

To compute the derivatives of ℓ_n^θ, a traditional approach, usually known as the *sensitivity equations method* consists in differentiating formally (2) with respect to θ [8]. This approach, which was originally developed for Gaussian linear state-space models, applies to HMMs in general as well [7] and has been used successfully for gradient-based parameter estimation [9, 10, 11]. It does however give rise to somewhat complicated expressions since everything in (2), including the normalization constant, does depend on θ.

A (in our view) simpler approach, advocated by [12], consists in using Fisher and Louis identities [3, 13] which state that

$$\nabla \ell_n^\theta = \sum_{k=0}^{n-1} \mathrm{E}^\theta [\nabla \log r_k^\theta(X_k, X_{k+1}) \,|\, Y_{0:n}] \qquad (3)$$

and

$$\begin{aligned}\nabla^2 \ell_n^\theta + \nabla \ell_n^\theta \left(\nabla \ell_n^\theta \right)^t &= \sum_{k=0}^{n-1} \mathrm{E}^\theta \left[\nabla^2 \log r_k^\theta(X_k) \middle| Y_{0:n} \right] \qquad (4)\\ &+ \sum_{k=0}^{n-1} \sum_{j=0}^{n-1} \mathrm{E}^\theta \Bigg[\nabla \log r_k^\theta(X_k, X_{k+1})\\ &\qquad \left(\nabla \log r_j^\theta(X_j, X_{j+1}) \right)^t \Bigg| Y_{0:n} \Bigg]\end{aligned}$$

respectively, where

$$r_k^\theta(x, x') \stackrel{\text{def}}{=} q^\theta(x, x') g_{k+1}^\theta(x') \qquad (5)$$

for $k \geq 1$ and

$$r_0^\theta(x, x') \stackrel{\text{def}}{=} \nu^\theta(x) g_0^\theta(x') q^\theta(x, x') g_1^\theta(x')$$

The notations ∇ and ∇^2 stand for the gradient and the Hessian, respectively, and the superscript t denotes matrix transposition.

At first glance, (3) and (4) hardly seem reconcilable with the principle behind the sensitivity equations since (3) and (4) involve the joint smoothing distribution. It has been shown however by [14] – see also [15] – that the conditional expectation of expressions of the form taken by (3), that is, sum functionals of the hidden states, may be computed recursively. Note that this remark also applies to the quantities that are needed to implement the EM algorithm [5, 15]. The observed information matrix in (4) however implies a quantity (on the second and third lines) which is obviously not a sum functional but the square of a sum functional.

The purpose of this contribution is to show that there exist recursive smoothing relations for any functional $\{t_n(x_{0:n})\}_{n \geq 0}$ of the hidden states which is such that

$$t_{n+1}(x_{0:n+1}) = m_n(x_n, x_{n+1}) t_n(x_{0:n}) + s_n(x_n, x_{n+1}) \qquad (6)$$

where $\{m_n\}_{n \geq 0}$ and $\{s_n\}_{n \geq 0}$ are two sequences of, possibly vector- or matrix-valued (with suitable dimensions), functions on $\mathsf{X} \times \mathsf{X}$ and t_0 is a function on X.

3. A GENERAL RECURSIVE SMOOTHING FORMULA

In this section, we establish our main result temporarily omitting the dependence with respect to the parameter (superscript θ) from our notations.

Proposition 1. *Let* $(t_n)_{n \geq 0}$ *be a sequence of integrable functions defined by* (6). *The sequence of auxiliary functions* $\{\tau_n\}_{n \geq 0}$ *on* X *such that*

$$\int f(x)\, \tau_n(x) dx \stackrel{\text{def}}{=} \mathrm{E}\left[f(X_n)\; t_n(X_{0:n}) \middle| Y_{0:n} \right] \qquad (7)$$

for integrable functions f*, may be updated recursively according to*

$$\begin{aligned}\tau_{n+1}(x_{n+1}) = c_{n+1}^{-1} \int \Bigg[&\tau_n(x_n)\, m_n(x_n, x_{n+1}) \qquad (8)\\ &+ \phi_{n|n}(x_n)\, s_n(x_n, x_{n+1}) \Bigg] r_n(x_n, x_{n+1}) dx_n\end{aligned}$$

for $n \geq 1$[2]*. At any index* n*,* $\mathrm{E}[t_n(X_{0:n}) \,|\, Y_{0:n}]$ *may be evaluated by computing* $\int \tau_n(x) dx$*.*

In order to use (8) *it is required that the standard filtering recursion be carried out in parallel where the notation*

$$c_k \stackrel{\text{def}}{=} \iint \phi_{k-1|k-1}(x_{k-1}) r_{k-1}(x_{k-1}, x_k) dx_{k-1} dx_k$$

stands for the normalizing constant which is computed when normalizing (2)*.*

[1] These relations will be established below in Section 3.

[2] The initialization of the recursion, that is, the computation of τ_0 and τ_1, is discussed below at the end of the proof of Proposition 1.

The auxiliary function $\tau_n(x)$ as defined in Proposition 1 should be understood as the minimal summary of the joint distribution of the hidden states $X_{0:n}$ given the observations $Y_{0:n}$ which still allows for recursive computation of $\mathrm{E}\left[t_n(X_{0:n})\middle|\,Y_{0:n}\right]$. Proposition 1 is a simple consequence of the structure of the joint smoothing distribution and of the particular choice of the smoothing functional in (6).

Proof. Let $\phi_{0:n|n}(x_{0:n})$ denote the joint probability density function of the hidden states $X_{0:n}$ *conditioned on the observations* $Y_{0:n}$ (which we call the "joint-smoothing" density). From the modelling assumptions given at the beginning of Section 2, the joint probability density function of the states and observations up to time n is given by

$$\prod_{k=0}^{n-1} r_k(x_k, dx_{k+1})$$

where the notation introduced in (5) has been used. Since the likelihood L_n is precisely obtained by marginalizing the above expression with respect to the state variables, Bayes rule implies that

$$\phi_{0:n|n}(x_{0:n}) = \mathrm{L}_n^{-1} \prod_{k=0}^{n-1} r_k(x_k, dx_{k+1}) \tag{9}$$

Comparing the above expression for consecutive indices $n \geq 1$ and $n+1$ gives the following update equation for the joint smoothing probability density function:

$$\phi_{0:n+1|n+1}(x_{0:n+1}) = \left(\frac{\mathrm{L}_{n+1}}{\mathrm{L}_n}\right)^{-1} \phi_{0:n|n}(x_{0:n})\, r_n(x_n, x_{n+1}) \tag{10}$$

Marginalizing the previous relation with respect to all variables but x_{n+1} yields the marginal filtering update

$$\phi_{n+1|n+1}(x_{n+1}) = \left(\frac{\mathrm{L}_{n+1}}{\mathrm{L}_n}\right)^{-1} \int \phi_{n|n}(x_n)\, r_n(x_n, x_{n+1}) dx_n \tag{11}$$

which also shows that the normalizing constant $c_{n+1} = \mathrm{L}_{n+1}/\mathrm{L}_n$ may be computed as

$$c_{n+1} = \iint \phi_{n|n}(x_n)\, r_n(x_n, x_{n+1}) dx_n dx_{n+1} \tag{12}$$

By definition,

$$\tau_{n+1}(x_{n+1}) = \int \cdots \int t_{n+1}(x_{0:n+1})\, \phi_{0:n+1|n+1}(x_{0:n+1}) dx_{0:n}$$

Using the recursive structures of t_{n+1} and $\phi_{0:n+1|n+1}$ given by (6) and (10), respectively, yields (for $n \geq 1$)

$$\tau^\theta_{n+1}(x_{n+1}) = \int \cdots \int \left[m^\theta_n(x_n, x_{n+1})\, t^\theta_n(x_{0:n}) + s^\theta_n(x_n, x_{n+1})\right] \left(c^\theta_{n+1}\right)^{-1} \phi^\theta_{0:n|n}(x_{0:n})\, r^\theta_n(x_n, x_{n+1}) dx_{0:n}$$

Equation (8) then follows by evaluating the integrals with respect to all the variables but x_n.

The case of the first update (computation of τ_0 and τ_1), which is slightly different in form due to our definition of r_0, obviously can be handled similarly starting with

$$c_0 = \int \nu(x_0) g_0(x_0) dx_0$$
$$\phi_{0|0}(x_0) = c_0^{-1} \nu(x_0) g_0(x_0)$$
$$\tau_0(x_0) = t_0(x_0)\, \phi_{0|0}(x_0)$$

It is then easy to check that $\tau_1(x_1)$ can be obtained by applying (8) with $\tilde{r}_0(x_0, x_1) = q(x_0, x_1) g_1(x_1)$ rather than with r_0 as defined in (5). □

Retrospectively, it is obvious that the choice of the functional in (6) has been mainly guided by the objective of mimicking the structure of the joint smoothing distributions $\{\tau_n\}_{n\geq 0}$. More precisely, $t_{n+1}(x_{0:n+1})$ should be easily expressed as a function of $t_n(x_{0:n})$ and terms that only involve the last two variables x_n and x_{n+1}. It is clear that (6) is not the most general structure for which a recursive smoothing relation similar to (8) holds. This type of functional is sufficient, however, to handle our two main objects of interest in the context of the present contribution, that is, the score in (3) and the observed information matrix in (4).

4. APPLICATION TO THE SCORE AND OBSERVED INFORMATION MATRIX

To apply Proposition 1 to the expression of the score given in (3), one simply needs to remark that Fisher's identity corresponds to a sum functional which is such that

$$t^\theta_n(x_{0:n}) = \sum_{k=0}^{n-1} \nabla \log r^\theta_k(x_k, x_{k+1})$$

Hence, we may use Proposition 1 with $m^\theta_n \equiv 1$ and

$$s^\theta_n(x_n, x_{n+1}) = \nabla \log r^\theta_n(x_n, x_{n+1})$$

In this case, (8) may be rewritten as

$$\tau^\theta_{n+1}(x_{n+1}) = \left(c^\theta_{n+1}\right)^{-1} \int \Big[\tau^\theta_n(x_n) + \phi^\theta_{n|n}(x_n)\, \nabla \log r^\theta_k(x_n, x_{n+1})\Big] r^\theta_n(x_n, x_{n+1}) dx_n \tag{13}$$

For the observed information matrix, (4) shows that in order to evaluate $\nabla^2 \ell^\theta_n$ on needs to compute three different terms. The first, $\nabla \ell^\theta_n \left(\nabla \ell^\theta_n\right)^t$, is a simple function of the score which we already know how to compute. The second term,

$$\sum_{k=0}^{n-1} \mathrm{E}^\theta \left[\nabla^2 \log r^\theta_k(X_k) \middle| Y_{0:n}\right]$$

may be handled exactly as in the case of the score, upon defining $m^\theta_n \equiv 1$ and

$$s^\theta_n(x_n, x_{n+1}) = \nabla^2 \log r^\theta_n(x_n, x_{n+1})$$

The third term is the most problematic as it involves a functional of the form

$$t_n^\theta(x_{0:n}) = \\ \sum_{k=0}^{n-1}\sum_{j=0}^{n-1} \nabla \log r_k^\theta(x_k, x_{k+1}) \left(\nabla \log r_j^\theta(x_j, x_{j+1})\right)^t$$

If we define $\tilde{t}_n^\theta(x_{0:n})$ by

$$\tilde{t}_n^\theta(x_{0:n}) = \sum_{k=0}^{n-1} \nabla \log r_k^\theta(x_k, x_{k+1})$$

that is, the functional which appeared above in the case of the score, we have the simple recursion

$$\begin{aligned} t_{n+1}^\theta(x_{0:n+1}) = t_n^\theta(x_{0:n}) + \tilde{t}_n^\theta(x_{0:n}) \left(\nabla \log r_n^\theta(x_n, x_{n+1})\right)^t \\ + \nabla \log r_n^\theta(x_n, x_{n+1}) \left(\tilde{t}_n^\theta(x_{0:n})\right)^t \\ + \nabla \log r_n^\theta(x_n, x_{n+1}) \left(\nabla \log r_n^\theta(x_n, x_{n+1})\right)^t \end{aligned}$$

Hence, if $\tilde{\tau}_n^\theta$ denotes the auxiliary function associated to the score by (7), which may be updated recursively according to (13), we have the following update formula

$$\begin{aligned} \tau_{n+1}^\theta(x_{n+1}) = \left(c_{n+1}^\theta\right)^{-1} \int \Big[\tau_n^\theta(x_n) \qquad (14) \\ + \tilde{\tau}_n^\theta(x_n) \left(\nabla \log r_n^\theta(x_n, x_{n+1})\right)^t \\ + \nabla \log r_n^\theta(x_n, x_{n+1}) \left(\tilde{\tau}_n^\theta(x_n)\right)^t \\ + \phi_{n|n}^\theta(x_n) \nabla \log r_n^\theta(x_n, x_{n+1}) \left(\nabla \log r_n^\theta(x_n, x_{n+1})\right)^t \\ \Big] r_n^\theta(x_n, x_{n+1}) dx_n \end{aligned}$$

where τ_n^θ is the auxiliary function corresponding to the Hessian functional t_{n+1}^θ.

Of course, the practical usefulness of the above abstract recursive formulas crucially depends on our ability to carry out the integrations that appear in (13), (14) and similar equations. In practice, exact computation is mostly feasible in case where the state space X consists of a finite set of point (so-called discrete HMMs) or in the case of Gaussian linear state-space models. We refer to Chapter 10 of [5] for some detailed examples of both situations. In general HMMs, neither the above recursive smoothing equations nor the simpler filtering recursion – Equations (11) and (12) – are feasible and approximate computations must be used. In this context, however, the above recursive smoothing equations may nonetheless be used to derive numerical approximations schemes based on the sequential Monte Carlo (or "particle filtering") approach [16].

5. CONNECTION WITH THE SENSITIVITY EQUATIONS APPROACH

There has been some debate as to whether the above framework really is equivalent to the sensitivity equations approach briefly outlined in Section 2 [7, 17, 18]. The apparent difference between both methods is that when using the Fisher and Louis identities one fundamentally computes expectations with respect to the *joint smoothing distribution*, that is, the distribution of all state variables $X_{0:n}$ given the corresponding observations $Y_{0:n}$; in contrast, the sensitivity equations are obtained by differentiating, with respect to the parameter θ, a decomposition – Equation (1) – which only involves the sequence of filtering distributions (distributions of X_k given the observations up to time k, $Y_{0:k}$). Even if Proposition 1 shows that the computations can also be carried out recursively in the first approach, the gap with the sensitivity equations seems hard to bridge. It turn outs, however, that differentiation with respect to θ, in particular of the normalization factors c_n^θ, does give rise to terms that cannot be expressed as function of the filtering distributions anymore. In effect, both approaches are closely related, which we show below for the score function.

Recall that the log-likelihood may be written according to (1) as the sum

$$\ell_n^\theta = \sum_{k=0}^{n} \log c_k^\theta \qquad (15)$$

where $c_k^\theta = \mathrm{L}_k^\theta / \mathrm{L}_{k-1}^\theta$ is also the normalizing constant that appears in the filtering recursion (11). To differentiate (11) with respect to θ, we assume that c_{k+1}^θ does not vanish and we use the identity

$$\nabla_\theta \frac{u(\theta)}{v(\theta)} = v^{-1}(\theta)\nabla_\theta u(\theta) - \frac{u(\theta)}{v(\theta)} \nabla_\theta \log v(\theta)$$

to obtain

$$\begin{aligned} \nabla \phi_{n+1|n+1}^\theta(x_{n+1}) = \\ \rho_{n+1}^\theta(x_{n+1}) - \phi_{n+1|n+1}^\theta(x_{n+1}) \nabla \log c_{n+1}^\theta \qquad (16) \end{aligned}$$

where

$$\rho_{n+1}^\theta(x_{n+1}) \overset{\text{def}}{=} \left(c_{n+1}^\theta\right)^{-1} \nabla \int \phi_{n|n}^\theta(x_n)\, r_n^\theta(x_n, x_{n+1}) dx_n \qquad (17)$$

We further assume that we may interchange integration and differentiation with respect to θ. Thus, as $\phi_{n+1|n+1}^\theta$ is a probability density function, $\nabla \int \phi_{n+1|n+1}^\theta(x_{n+1}) dx_{n+1} = 0$. Therefore, integration of both sides of (16) with respect to x_{n+1} yields

$$0 = \int \rho_{n+1}^\theta(x_{n+1}) dx_{n+1} - \nabla \log c_{n+1}^\theta$$

Hence, the gradient of the log-likelihood increment $c_{n+1}^\theta = \ell_{n+1}^\theta - \ell_n^\theta$ may be expressed from ρ_{n+1}^θ as

$$\nabla \log c_{n+1}^\theta = \int \rho_{n+1}^\theta(x_{n+1}) dx_{n+1} \qquad (18)$$

Now, we evaluate the derivative in (17) assuming also that r_n is non-zero to obtain

$$\begin{aligned} \rho_{n+1}^\theta(x_{n+1}) = \left(c_{n+1}^\theta\right)^{-1} \int \Big[\nabla \phi_{n|n}^\theta(x_n) \\ + \nabla \log r_n^\theta(x_n, x_{n+1})\, \phi_{n|n}^\theta(x_n) \Big] r_n^\theta(x_n, x_{n+1}) dx_n \end{aligned}$$

Plugging (16) into the above equation yields an update formula for ρ_{n+1}^θ:

$$\begin{aligned} \rho_{n+1}^\theta(x_{n+1}) = \left(c_{n+1}^\theta\right)^{-1} \int \Big[\nabla \phi_{n|n}^\theta(x_n) \\ + \nabla \log r_n^\theta(x_n, x_{n+1})\, \phi_{n|n}^\theta(x_n) \Big] r_n^\theta(x_n, x_{n+1}) dx_n \\ - \phi_{n+1|n+1}^\theta(x_{n+1}) \nabla \log c_n^\theta \qquad (19) \end{aligned}$$

where (11) has been used for the last term on the right-hand side.

In the above derivations, we used a new auxiliary function ρ_n^θ, defined in (17), whose integral is the quantity of interest $\nabla \log c_n^\theta$. Obviously, one can equivalently use as auxiliary function the derivative of the filtering probability density function $\nabla \phi_{n|n}^\theta$, which is directly related to ρ_n^θ by (16). The quantity $\nabla \phi_{n|n}(\cdot)$, which is referred to as the *tangent filter* by [9], is also known as the *filter sensitivity* and may be of interest in its own right. Using $\nabla \phi_{n|n}^\theta$ instead of ρ_n^θ does not however modify the nature of algorithm.

Recall from Sections 3 – 4 that Proposition 1 asserts that the score $\nabla \ell_n^\theta$ may be computed as $\int \tau_n^\theta(x_n)(dx_n)$ for an auxiliary function τ_n^θ which is updated by (13). Comparing (13) with (19), it is easily established by recurrence on n that

$$\rho_n^\theta = \tau_n^\theta - \left(\sum_{l=0}^{n-1} \nabla_\theta \log c_l^\theta \right) \phi_{n|n}^\theta \tag{20}$$

for $n \geq 1$. Hence, whereas $\int \tau_n^\theta(x_n) dx_n$ evaluates to $\nabla \ell_n^\theta$, the gradient of the log-likelihood up to index n, $\int \rho_n^\theta(x_n) dx_n$ equals the gradient of the *increment* of the log-likelihood, $\ell_n^\theta - \ell_{n-1}^\theta$. But in the latter case, the term ℓ_{n-1}^θ is indeed decomposed into the telescoping sum $\ell_{n-1}^\theta = \sum_{k=0}^{n-1} \nabla \log c_k^\theta$ of log-likelihood increments. Thus, the sensitivity equations and the use of Fisher's identity combined with the recursive smoothing algorithm of Proposition 1 are equivalent.

6. CONCLUSIONS

We have shown that it is possible to obtain generic and relatively simple recursive smoothing relations for a large class of functionals of the hidden state variables. This class of functionals includes in particular the forms taken by the score and the observed information matrix when decomposed according to the Fisher and Louis identities, respectively. Although the score and observed information equations can also be equivalently derived following the sensitivity approach, we feel that the framework described in this contribution offers a clearer point of view on the problem of computing log-likelihood derivatives as well as a connection with the more general task of recursively evaluating smoothed estimates.

7. REFERENCES

[1] L. E. Baum, T. P. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, 1970.

[2] R. E. Kalman and R. Bucy, "New results in linear filtering and prediction theory," *J. Basic Eng., Trans. ASME, Series D*, vol. 83, no. 3, pp. 95–108, 1961.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 39, no. 1, pp. 1–38 (with discussion), 1977.

[4] A. Doucet, N. De Freitas, and N. Gordon, Eds., *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

[5] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, Springer, 2005.

[6] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte-Carlo sampling methods for Bayesian filtering," *Stat. Comput.*, vol. 10, pp. 197–208, 2000.

[7] F. Campillo and F. Le Gland, "MLE for patially observed diffusions: Direct maximization vs. the EM algorithm," *Stoch. Proc. App.*, vol. 33, pp. 245–274, 1989.

[8] N. Gupta and R. Mehra, "Computational aspects of maximum likelihood estimation and reduction in sensitivity function calculations," *IEEE Trans. Automat. Control*, vol. 19, no. 6, pp. 774–783, 1974.

[9] F. Le Gland and L. Mevel, "Recursive estimation in HMMs," in *Proc. IEEE Conf. Decis. Control*, 1997, pp. 3468–3473.

[10] O. Cappé, V. Buchoux, and E. Moulines, "Quasi-Newton method for maximum likelihood estimation of hidden Markov models," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, vol. 4, pp. 2265–2268.

[11] I. B. Collings and T. Rydén, "A new maximum likelihood gradient algorithm for on-line hidden Markov model identification," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1998, vol. 4, pp. 2261–2264.

[12] M. Segal and E. Weinstein, "A new method for evaluating the log-likelihood gradient, the Hessian, and the Fisher information matrix for linear dynamic systems," *IEEE Trans. Inform. Theory*, vol. 35, pp. 682–687, 1989.

[13] T. A. Louis, "Finding the observed information matrix when using the EM algorithm," *J. Roy. Statist. Soc. Ser. B*, vol. 44, pp. 226–233, 1982.

[14] O. Zeitouni and A. Dembo, "Exact filters for the estimation of the number of transitions of finite-state continuous-time Markov processes," *IEEE Trans. Inform. Theory*, vol. 34, no. 4, July 1988.

[15] R. J. Elliott, L. Aggoun, and J. B. Moore, *Hidden Markov models: Estimation and Control*, Springer, 1995.

[16] O. Cappé, "Recursive computation of smoothed functionals of hidden Markovian processes using a particle approximation," *Monte Carlo Methods Appl.*, vol. 7, no. 1-2, pp. 81–92, 2001.

[17] J. Fichou, F. Le Gland, and L. Mevel, "Particle based methods for parameter estimation and tracking : Numerical experiments," Tech. Rep. PI-1604, INRIA, 2004.

[18] A. Doucet and V. B. Tadić, "Parameter estimation in general state-space models using particle methods," *Ann. Inst. Statist. Math.*, vol. 55, no. 2, pp. 409–422, 2003.