Asymptotically Efficient Identification of Known-Sensor Hidden Markov Models

Robert Mattila, Cristian R. Rojas, *Member, IEEE*,
Vikram Krishnamurthy, *Fellow, IEEE* and Bo Wahlberg, *Fellow, IEEE*

***Abstract*—We consider estimating the transition probability matrix of a finite-state finite-observation alphabet hidden Markov model with known observation probabilities. We propose a two-step algorithm: a method of moments estimator (formulated as a convex optimization problem) followed by a single iteration of a Newton-Raphson maximum likelihood estimator. The two-fold contribution of this letter is, firstly, to theoretically show that the proposed estimator is consistent and asymptotically efficient, and secondly, to numerically show that the method is computationally less demanding than conventional methods – in particular for large data sets.**

***Index Terms*—Hidden Markov models, method of moments, maximum likelihood, system identification**

I. Introduction

THE *hidden Markov model* (HMM) has been applied in a diverse range of fields, e.g., signal processing [1], gene sequencing [2], [3] and speech recognition [4]. The standard way of estimating the parameters of an HMM is by employing a *maximum likelihood* (ML) criterion. However, numerical "hill climbing" algorithms for computing the ML estimate, such as direct maximization using Newton-Raphson (and variants, e.g., [5]) and the *expectation-maximization* (EM, e.g, [4], [6]) algorithm are, in general, only guaranteed to converge to local stationary points in the likelihood surface. It is also known that these schemes can, depending on the initial starting point of the algorithms, the shape of the likelihood surface and the size of the data set, exhibit long run-times.

An alternative to ML criterion is to match moments of an HMM, resulting in a *method of moments* estimator (see, e.g. [7] for details). In such a method, observable correlations in the HMM data are related to the parameters of the system. The correlations are empirically estimated and used in the inverted relations to recover parameter estimates. A number of methods of moments for HMMs have been proposed in the recent years; e.g., [8]–[14]. The main benefits over iterative ML schemes are usually consistency and a shorter run-time, however, since typically only low-order moments are considered, there is a loss of efficiency in the resulting estimate.

In the present letter, the problem of estimating the transition probabilities of a finite discrete-time HMM with known sensor uncertainties, i.e., observation matrix, is considered. This setup can be motivated in two ways: firstly, it can be seen as the second step in a *decoupling approach* to learning the HMM parameters (see [11]), or alternatively, by any application where the sensor used to measure the system is designed/known to the user.

The main idea in this letter is a hybrid two-step algorithm based on combining the advantages of the two aforementioned approaches. The first step uses a method of moments estimator which requires a single pass over the data set (compared to iterative algorithms, such as EM, that require multiple iterations over the data set). The second step uses the method of moments estimate to initialize a non-iterative second-order direct likelihood maximization procedure. This allows us to avoid resorting to ad hoc heuristics for localizing a good starting point. More importantly, we show that it is *sufficient to perform only a single iteration* of the ML procedure to obtain an asymptotically efficient estimate. Put differently, only two passes through the data set are necessary in order to obtain an asymptotically efficient estimate.

To summarize, the main contributions of this letter are:

- a proposed two-step identification algorithm that exploits the benefits of both the method of moments approach (low computational burden and consistency) and direct likelihood maximization (high accuracy);
- we prove the consistency and asymptotic efficiency of the proposed estimator. Hence, the problem of only local convergence that may haunt iterative ML algorithms, such as EM, is shown to be avoided;
- numerical studies that show that the proposed method is up to an order of magnitude faster than the standard EM algorithm – with comparable accuracy (when the EM iterations approach the global optimum of the likelihood function). Moreover, the run-time is, roughly, constant for a fixed data size, whereas the run-time of EM is highly dependent on the data (due to the number of iterations needed for convergence).

The outline of the remaining part of this letter is as follows. We first present a brief overview of related work below. Section II then poses the problem formally and Section III presents the algorithm. In Section IV asymptotic efficiency is proven, and Section V presents numerical studies.

Related Work

HMM parameter estimation is by now a classical area (with more than 50 years of literature). There has recently been interest in the machine learning community for employing

This work was partially supported by the Swedish Research Council and the National Science Foundation Grant 1714180. Robert Mattila, Cristian R. Rojas and Bo Wahlberg are with the Department of Automatic Control, School of Electrical Engineering, KTH Royal Institute of Technology. Stockholm, Sweden. (e-mails: {rmattila,crro,bo}@kth.se). Vikram Krishnamurthy is with the Department of Electrical and Computer Engineering, Cornell University. Cornell Tech, NY, USA. (e-mail: vikramk@cornell.edu).

methods of moments for HMMs. The method presented in [10, Appendix A] demonstrates how to recover explicit estimates of the transition and observation matrices by exploiting the special structure of the moments of an HMM. This method has been further generalized and put in a tensor framework; see, e.g., [9], [12] and references therein. The appealing attribute of these methods is that they generate non-iterative estimates using simple linear algebra operations (eigen and singular-value decompositions). However, the non-negativity and sum-to-one properties of the estimated probabilities cannot be guaranteed.

There are a number of proposed methods of moments for HMMs formulated as optimization problems (which allow constraints to be forced on the estimates), e.g., [8], [11] and [14]. The identification problem is *decoupled* in [11] into two stages: first an estimation of the output parameters, and then a moment matching optimization problem. The resulting optimization problem is related to the one in [8] and to the problem in the present work. The method we propose in this letter could be seen as a possible improvement of the second step in the setting of [11].

In the general setting, hybrid approaches, such as the combination of EM and direct likelihood maximization, and other attempts to accelerate EM has been studied in, e.g., [15], [16]. Iterative direct likelihood maximization for HMMs, as well as methods for obtaining the necessary gradient and Hessian expressions, are treated in, e.g., [5], [17]–[21]. The combination of a method of moments and EM has, in the case of HMMs, been considered in [11].

II. Preliminaries and Problem Formulation

All vectors are column vectors unless transposed, $\mathbb{1}$ denotes the vector of all ones. The vector operator $\text{diag} : \mathbb{R}^n \to \mathbb{R}^{n\times n}$ gives the matrix where the vector has been put on the diagonal, and all other elements are zero. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The element at row i and column j of a matrix is $[\cdot]_{ij}$, and the element at position i of a vector is $[\cdot]_i$. Inequalities $(>, \geq, \leq, <)$ between vectors or matrices should be interpreted elementwise. The indicator function $\mathrm{I}\{\cdot\}$ takes the value 1 if the expression $\cdot$ is fulfilled and 0 otherwise. Let $\to_p$ and $\to_d$ denote convergence in probability and in distribution, respectively, and let $\mathcal{O}_p$ and o_p be stochastic-order symbols. $\sim$ denotes "distributed according to".

A. Problem Formulation

Consider a discrete-time finite-state *hidden Markov model* (HMM) on the state space $\mathcal{X} = \{1, 2, \ldots, X\}$ with the transition probability matrix

$$[P]_{ij} = \Pr[x_{k+1} = j | x_k = i]. \tag{1}$$

Observations are made from the set $\mathcal{Y} = \{1, 2, \ldots, Y\}$ according to the observation probability matrix

$$[B]_{ij} = \Pr[y_k = j | x_k = i]. \tag{2}$$

These matrices are row-stochastic, i.e., the elements in each row sum to one. Denote the initial distribution as π_0 and the stationary distribution as π_∞.

The HMM moments are joint probabilities of tuples of observations. The second order moments can be represented by $Y \times Y$ matrices M_k with elements

$$[M_k]_{ij} = \Pr[y_k = i, y_{k+1} = j]. \tag{3}$$

The following equation (see [22] for a derivation) relates the second order moments and the system parameters,

$$M_k = B^T \operatorname{diag}((P^T)^k \pi_0) P B, \tag{4}$$

and is the key to the method of moments formulation of the problem.

As we are interested in the asymptotic behaviour, we make the assumption that the initial distribution π_0 is known to us – its influence will anyway diminish over time. The most important assumption we make is that the observation probabilities B are known. There are three motivations for this assumption: i) it admits the problem to a convex formulation, ii) it holds in any real-world application where the sensor is designed by the user, and iii) our method can be seen as an intermediate step of the *decoupling* approach in [11]. The identification problem we consider is, hence,

Problem 1. *Consider an HMM with known initial distribution π_0 and known observation matrix B. The HMM is initialized according to π_0 and a sequence of observations $y_0, y_1, \ldots, y_N$ is obtained. Given the sequence of $N+1$ observations $\{y_k\}_{k=0}^N$, estimate the transition matrix P.*

III. Asymptotically Efficient Two-Step Algorithm

In this section, we outline the two-step algorithm which is the main contribution of this letter.

Step 1. Initial Method of Moments Estimate

In light of (3), use the empirical moments estimate

$$[\hat{M}_\infty]_{ij} = \frac{1}{N} \sum_{k=0}^{N-1} \mathrm{I}\{y_k = i, y_{k+1} = j\}, \tag{5}$$

for the (stationary) second order moments.

In the moment matching optimization problem, we need to impose a number of constraints. Firstly, that the transition matrix is a valid stochastic matrix, that is: the non-negativity and sum-to-one properties of its rows. We will require that the transition matrix of the HMM is ergodic (aperiodic and irreducible). This implies, first of all, that π_∞ is the right eigenvector of P^T corresponding to the eigenvalue 1 and therefore satisfies the condition $\pi_\infty = P^T \pi_\infty$, and secondly, that π_∞ has strictly positive entries. We therefore also include in the optimization problem a polytopic bound $\underline{\Pi}$ on π_∞ such that for a vector $x \in \underline{\Pi} \Rightarrow x > 0$.[1]

To summarize, estimating the transition matrix P involves solving the optimization problem (as the limit is taken in equation (4) towards stationarity):

$$\min_{\pi_\infty, P} \quad \|\hat{M}_\infty - B^T \operatorname{diag}(\pi_\infty) P B\|_F^2$$

[1]This polyhedron can, for example, be obtained if it is possible to *a priori* lower bound the elements of the transition matrix P using another matrix L. In particular, this is possible since then the stationary distribution π_∞ lies in a polyhedron $\underline{\Pi}$ spanned by the normalized (i.e., non-negative and with elements that sum to one) columns of the matrix $(I - L^T)^{-1}$ – see [23] for details.

$$\begin{aligned} \text{s.t.} \quad & P \geq 0, \qquad \pi_\infty \geq 0, \\ & P\mathbb{1} = \mathbb{1}, \quad \mathbb{1}^T \pi_\infty = 1, \\ & \pi_\infty \in \underline{\Pi}, \quad \pi_\infty = P^T \pi_\infty. \end{aligned} \tag{6}$$

This is, in general, a non-convex optimization problem. The lemma below shows that convex optimization techniques can be used to solve the problem.

Lemma 1. *The solution of problem (6) is obtainable by solving the convex problem*

$$\begin{aligned} \min_{A} \quad & \|\hat{M}_\infty - B^T A B\|_F^2 \\ \textit{s.t.} \quad & A \geq 0, \mathbb{1}^T A \mathbb{1} = 1, \\ & A\mathbb{1} \in \underline{\Pi}, A\mathbb{1} = A^T \mathbb{1}, \end{aligned} \tag{7}$$

and using (9) and (10), see below, to recover π_∞ and P from the variable A.

Proof. In problem (7), we identify the product $\mathrm{diag}(\pi_\infty)P$ in problem (6) as a new parameter A, i.e.,

$$A = \mathrm{diag}(\pi_\infty)P, \tag{8}$$

and optimize over its elements instead of over π_∞ and P jointly. Notice that it is possible to recover π_∞ and P from A as follows: Firstly, recover π_∞ from

$$A\mathbb{1} = \mathrm{diag}(\pi_\infty)P\mathbb{1} = \pi_\infty, \tag{9}$$

employing the fact that $P\mathbb{1} = \mathbb{1}$. Secondly, recover P from

$$\mathrm{diag}(\pi_\infty)^{-1} A = \mathrm{diag}(\pi_\infty)^{-1} \mathrm{diag}(\pi_\infty) P = P. \tag{10}$$

The lemma follows by noting that the cost functions in problems (6) and (7) are the same, and then mapping feasible solutions between the two problems. □

Solving problem (7) requires only a single pass over the data to obtain $\hat{M}_\infty$, and then solving a data-size independent convex (quadratic) optimization problem to compute an estimate of the transition matrix P. The trade-off compared to ML estimation, which requires multiple iterations over the observation data set, is of course between estimation accuracy and computational cost: the method of moments outlined above employs only the second order moments and will hence have disregarded some of the information in the observed data.

Step 2. Single Newton-Raphson Step

We propose to exploit the trade-off by first obtaining an estimate of P using the convex method of moments (7), and then taking *a single* Newton-Raphson step on the likelihood function to increase the accuracy of the estimate.

The (log-)likelihood function of the observed data is

$$l_N(\theta) = \log \Pr[\,\{y_k\}_{k=0}^N \,|\, x_0 \sim \pi_0;\, \theta\,], \tag{11}$$

where θ is a parametrization of the transition matrix P. Denote the estimate resulting from the method of moments (7) as $\hat{\theta}_{\text{MM}}$. Then a single Newton-Raphson step is performed as follows:[2]

$$\hat{\theta}_{\text{NR}} = \hat{\theta}_{\text{MM}} - \left[\nabla_\theta^2 l_N(\hat{\theta}_{\text{MM}})\right]^{-1} \nabla_\theta l_N(\hat{\theta}_{\text{MM}}), \tag{12}$$

[2]We assume that parametrization handles the constraints, if not, then the Newton-Raphson step can be formulated as a constrained quadratic program.

where the gradient $\nabla_\theta l_N(\hat{\theta})$ and Hessian $\nabla_\theta^2 l_N(\hat{\theta})$ can be computed recursively – see e.g., [5], [17]–[21].

Compared to direct maximization of the likelihood function using the Newton-Raphson method (see, e.g., [5], [19]), this procedure is non-iterative and hence, the gradient and Hessian *need only to be computed once*.

IV. Analysis

In this section we analyze the properties of the proposed algorithm. First we state the assumptions.

Assumption 1. *The transition matrix P has positive elements. The observation matrix B is given, has full rank and is positive. There is a polytopic bound on π_∞ such that all components of π_∞ are strictly greater than zero.*

The following lemma establishes (strong) consistency of the method of moments procedure.

Lemma 2. *The estimates of P and π_∞ obtained using (9) and (10) from problem* (7) *with the estimator $\hat{M}_\infty$ in (5), converge to their true values as the number of observations $N \to \infty$ with probability one.*

Proof (outline). The lemma follows by showing

1) that the estimate $\hat{M}_\infty$ converges to M_∞ (using a law of large numbers, [5, Theorem 14.2.53]);
2) that the solution $\hat{A}$ of the optimization problem converges to A (follows by the fundamental theorem of statistical learning [24, Lemma 1.1] and the convexity of the cost function [25, Theorem 10.8]);
3) that the solution of the optimization problem $\hat{A}$ can be uniquely mapped to P and π_∞.

Full details are available in the supplementary material, [22]. □

Next, we provide the main theorem of this letter.

Theorem 1. *The estimate $\hat{\theta}_{NR}$ obtained by the two-step algorithm (7)-(12) is asymptotically efficient, i.e., as $N \to \infty$,*

$$\sqrt{N}(\hat{\theta}_{NR} - \theta^*) \to_d \mathcal{N}(0, I_F^{-1}(\theta^*)), \tag{13}$$

where $\mathcal{N}$ is a normal distribution, θ^* corresponds to the true parameters and I_F is the Fisher information matrix.

Proof (outline). The theorem follows by showing that

1) the estimate $\hat{M}_\infty$ follows a central limit theorem [26, Corollary 5], and using this, concluding that $\hat{M}_\infty = M_\infty + \mathcal{O}_p(N^{-1/2})$ [27, Appendix A];
2) this order in probability can be propagated through the optimization problem (7) to obtain a similar order on $\hat{P}$ and $\hat{\pi}_\infty$ [28, Theorem 2.1];
3) verifying that certain regularity conditions hold to ensure that we have a central limit theorem for the gradient and a law of large numbers for the Hessian matrix of the log-likelihood function [5, Theorems 12.5.5 and 12.5.6];
4) verifying by explicit computation that the single Newton-Raphson step yields an asymptotically efficient estimator.

Again, full details are available in the supplementary material, [22].

□

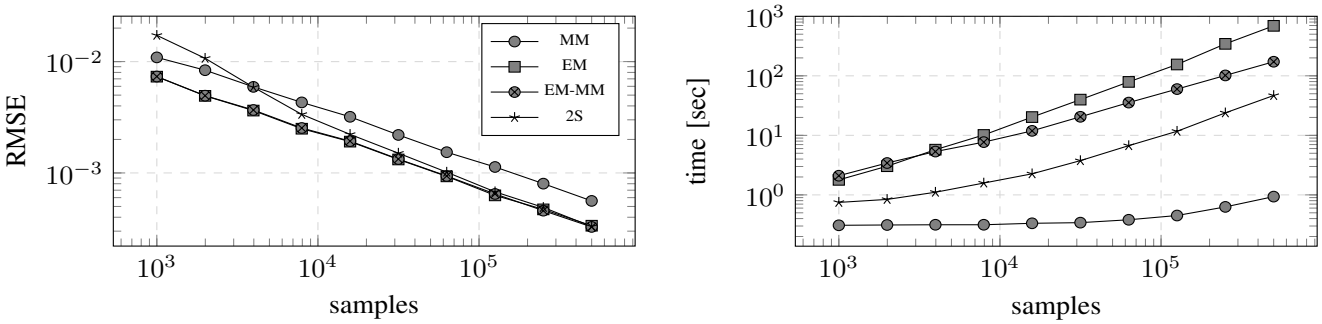

Figure 1: RMSE and run-time simulation data for a varying number of samples.

V. Numerical Evaluation

In this section, we evaluate the performance of the proposed two-step algorithm and compare it to the standard EM algorithm for ML estimation. The EM implementation of Matlab R2015a was employed (modified as to account for the fact that the observation matrix is assumed known). The first step of the proposed algorithm, i.e., solving the convex optimization problem (7), was performed using the CVX package [29]. The second step, i.e., the single Newton-Raphson update (12), can be implemented in (at least) two ways. The first is to recursively compute the gradient and Hessian as explained in, e.g., [5], [17]–[21]. The second, and the one we opted for, is to use *automatic differentiation* (AD, e.g., [30]). We interfaced Matlab to the ForwardDiff.jl-package in Julia [31] in our implementation. A small regularization term was added to the Hessian. Each simulation was run on an Intel Xeon CPU at 3.1 GHz.

We sampled observations from randomly generated systems of size $X = Y = 5$. Notice that there are a total of 20 unknown parameters (i.e., elements of P) to estimate for such systems. We used an elementwise lower bound $\underline{\Pi}$ of one tenth of the minimum element of the true stationary distribution of each system. We compared the performance of the proposed two-step algorithm (2S) to the estimate resulting from the method of moments (MM), as well as, the EM algorithm started in two different initial points: a random point (EM) and the method of moments estimate (EM-MM).

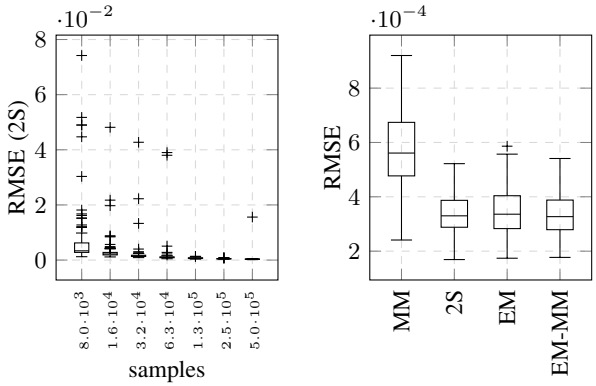

Figure 2: Each box contains 100 simulations. Left: RMSE of the proposed algorithm at different data sizes; Right: RMSE at $5 \cdot 10^5$ samples (one outlier not seen).

Fig. 1 presents the median over 100 simulations for each batch size of, left, the *root mean squared errors* (RMSEs) and, right, the run-times. Fig. 2 presents box plots of, left, the RMSEs of the proposed algorithm at various data sizes and, right, the RMSEs of the compared algorithms for $5 \cdot 10^5$ samples. All boxes contain 100 simulations. Three things can be noted from the figures.

Firstly, in the left plot of Fig. 1, the loss of accuracy resulting from only using the second order moments (compared to all moments in EM) is apparent from the distance between the MM-curve and the EM-curves. This can also be seen in the right plot of Fig. 2.

Secondly, also in the left plot of Fig. 1, we see that the asymptotic results become valid at around 10^5 samples when the accuracy of the 2S estimate becomes comparable to that of EM. In other words, the initial estimate is now sufficiently close to the optimum for the second order approximation to be appropriate. That the number of observed outliers drop in the left plot of Fig. 2 indicates the same conclusion: The outliers occurred when the Hessian was not negative definite – a result of the Newton-Raphson step leading the iterate towards a local stationary point. Note that this can be detected prior to employing the method.

Thirdly, in the right plot of Fig. 1, it can be see that the run-times of the compared algorithms differ by up to an order of magnitude. It should moreover be noted that the run-time of the proposed algorithm is more or less constant for a fixed data size (i.e., independent of the system and the observations), whereas the run-time of EM is highly dependent on the data (due to the number of iterations needed to converge): The maximum run-times for $5 \cdot 10^5$ observations were 1083 and 480 seconds for EM and EM-MM, respectively, whereas for the proposed method it was 54 seconds.

VI. Conclusion

This letter has proposed and analyzed a two-step algorithm for identification of HMMs with known sensor uncertainties. A method of moments was combined with direct likelihood maximization to exploit the benefits of both approaches: lower computational cost and consistency in the former, and accuracy in the later. Theoretical guarantees were given for asymptotic efficiency and numerical simulations showed that the algorithm can yield an accuracy comparable to that of the standard EM algorithm, but in up to an order of magnitude less time.

REFERENCES

[1] V. Krishnamurthy, *Partially Observed Markov Decision Processes*. Cambridge University Press, 2016.

[2] R. Durbin, *Biological sequence analysis: Probabalistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[3] M. Vidyasagar, *Hidden Markov Processes: Theory and Applications to Biology*. Princeton University Press, 2014.

[4] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

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

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, pp. 1–38, 1977.

[7] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, 1993.

[8] B. Lakshminarayanan and R. Raich, "Non-negative matrix factorization for parameter estimation in hidden Markov models," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP'10)*, 2010, pp. 89–94.

[9] A. Anandkumar, D. Hsu, and S. M. Kakade, "A method of moments for mixture models and hidden Markov models," in *Proceedings of the 25th Conference on Learning Theory (COLT'12)*, 2012, pp. 33.1–33.34.

[10] D. Hsu, S. M. Kakade, and T. Zhang, "A spectral algorithm for learning hidden Markov models," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1460–1480, 2012.

[11] A. Kontorovich, B. Nadler, and R. Weiss, "On learning parametric-output HMMs," in *Proceedings of the 30th International Conference on Machine Learning (ICML'13)*, vol. 28, 2013, pp. 702–710.

[12] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky, "Tensor decompositions for learning latent variable models," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2773–2832, 2014.

[13] R. Mattila, V. Krishnamurthy, and B. Wahlberg, "Recursive identification of chain dynamics in hidden Markov models using non-negative matrix factorization," in *Proceedings of the 54th IEEE Conference on Decision and Control (CDC'15)*, 2015, pp. 4011–4016.

[14] C. Subakan, J. Traa, P. Smaragdis, and D. Hsu, "Method of moments learning for left-to-right hidden Markov models," in *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'15)*, 2015, pp. 1–5.

[15] I. Meilijson, "A fast improvement to the EM algorithm on its own terms," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 127–138, 1989.

[16] J. A. Fessler and A. O. Hero, "Space-alternating generalized expectation-maximization algorithm," *IEEE Transactions on Signal Processing*, vol. 42, no. 10, pp. 2664–2677, 1994.

[17] T. C. Lystig and J. P. Hughes, "Exact computation of the observed information matrix for hidden Markov models," *Journal of Computational and Graphical Statistics*, vol. 11, no. 3, pp. 678–689, 2002.

[18] O. Cappé and E. Moulines, "Recursive computation of the score and observed information matrix in hidden Markov models," in *Proceedings of the 13th IEEE Workshop on Statistical Signal Processing*, 2005, pp. 703–708.

[19] R. Turner, "Direct maximization of the likelihood of a hidden Markov model," *Computational Statistics & Data Analysis*, vol. 52, no. 9, pp. 4147–4160, 2008.

[20] W. Khreich, E. Granger, A. Miri, and R. Sabourin, "A survey of techniques for incremental learning of HMM parameters," *Information Sciences*, vol. 197, pp. 105–130, 2012.

[21] I. L. MacDonald, "Numerical maximisation of likelihood: A neglected alternative to EM?" *International Statistical Review*, vol. 82, no. 2, pp. 296–308, 2014.

[22] R. Mattila, C. R. Rojas, V. Krishnamurthy, and B. Wahlberg, "Asymptotically efficient identification of known-sensor hidden Markov models," *arXiv:1702.00155 [cs.SY]*, 2017. [Online]. Available: https://arxiv.org/abs/1702.00155

[23] P.-J. Courtois and P. Semal, "On polyhedra of Perron-Frobenius eigenvectors," *Linear algebra and its applications*, vol. 65, pp. 157–170, 1985.

[24] M. Campi, "System identification and the limits of learning from data," Available at: http://marco-campi.unibs.it/pdf-pszip/sys-id-and-limits-learning.pdf, 2006. [Online]. Available: http://marco-campi.unibs.it/pdf-pszip/sys-id-and-limits-learning.pdf

[25] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1970.

[26] G. L. Jones, "On the Markov chain central limit theorem," *Probability Surveys*, vol. 1, pp. 299–320, 2004.

[27] D. Pollard, *Convergence of Stochastic Processes*. Springer, 1984.

[28] J. W. Daniel, "Stability of the solution of definite quadratic programs," *Mathematical Programming*, vol. 5, no. 1, pp. 41–53, 1973.

[29] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[30] A. Griewank and A. Walther, *Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation*, 2nd ed. SIAM, 2008.

[31] J. Revels, M. Lubin, and T. Papamarkou, "Forward-mode automatic differentiation in Julia," *arXiv:1607.07892 [cs.MS]*, 2016. [Online]. Available: https://arxiv.org/abs/1607.07892