2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

A Recursive Learning Algorithm for Model Reduction of Hidden Markov Models

Kun Deng, Prashant G. Mehta, Sean P. Meyn, and Mathukumalli Vidyasagar

Abstract— **This paper is concerned with a recursive learning algorithm for model reduction of Hidden Markov Models (HMMs) with finite state space and finite observation space. The state space is aggregated/partitioned to reduce the complexity of the HMM. The optimal aggregation is obtained by minimizing the Kullback-Leibler divergence rate *between the laws of the observation process*. The optimal aggregated HMM is given as a function of the partition function of the state space. The optimal partition is obtained by using a recursive stochastic approximation learning algorithm, which can be implemented through a single sample path of the HMM. Convergence of the algorithm is established using ergodicity of the filtering process and standard stochastic approximation arguments.**

I. Introduction

A fundamental problem for Hidden Markov Models (HMMs) that arise in applications is the large size of the underlying state space [1]. Aggregation of state space represents perhaps the most straightforward approach to the model reduction. It can be justified using a singular perturbation framework (see [2]) for *nearly completely decomposable* Markov chains (NCDMC). Recently, we proposed to employ the Kullback-Leibler divergence rate (K-L rate) for model reduction of Markov chains via aggregation of the state space [3]. By using the fact that the joint state and observation process of HMM is Markovian, we also extended this aggregation framework for the model reduction of HMMs based on the *K-L rate between laws of the joint process* [4], [5].

The problem with an aggregation based on joint process is that two HMMs may have very similar laws of the observation process, while the K-L rate of their joint laws might be very large or even unbounded. In this paper, we propose to use the *K-L rate between laws of the observation process* as the "probability distance" to compare two HMMs. If this distance is zero, then the two HMMs are equivalent in the probability sense up to a permutation of the state space (see Section II-A for more details). This K-L rate pseudo-metric has been studied in statistics [6], speech recognition [7], bioinformatics [8], and control theory [9], [10].

Kun Deng, Prashant G. Mehta, and Sean P. Meyn are with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, 1308 West Main Street, Urbana, IL 61801. Email: kundeng2@illinois.edu, mehtapg@illinois.edu, and meyn@illinois.edu

Mathukumalli Vidyasagar is with the Department of Bioengineering, University of Texas at Dallas, 800 W. Campbell Road, Richardson, TX 75080. Email: m.vidyasagar@utdallas.edu

Financial support from the National Science Foundation (grants CNS 0931416, ECCS 0523620, and ECCS 1001643) is gratefully acknowledged.

The goal of this paper is to find a reduced model of HMM via aggregation of the state space by minimizing the K-L rate between original and reduced laws of the observation process. There are two main ideas in this paper. One, we use a optimal representation of the aggregated HMM derived in our earlier work [4], [5] as a structured model for optimization. We take advantage of the optimal representation to overcome some of the complexity issues in computing the K-L rate. The second idea is to generate observations from the original HMM with large state space, but to recursively evaluate the filter only for the aggregated HMM with much smaller state space. The aggregated HMM is represented in terms of parameters from the original HMM and the partition function which needs to be optimized.

Since the partition function space is exponentially large (M^N for the M-partition of N-state space), we first parameterize the discrete partition function space into a smaller real parameter space [4] and then convert the optimal partition problem to an optimal estimation problem, in fact, the Maximum Likelihood Estimation (MLE) problem of the HMM [6], [11]. We employ a gradient-based simulation algorithm to solve the MLE problem. The algorithm is recursively updated based on the stochastic gradient of the nonlinear filter evaluated using the aggregated HMM model. The convergence of the algorithm is established based on the stochastic approximation arguments as well as the the ergodicity of the filtering process.

II. Preliminaries and Notations

A. Hidden Markov Model

In this paper we consider a discrete-time HMM $\{X_n, Y_n\}_{n\geq 0}$ defined on a probability space $(\Omega, \mathcal{F}, \mathsf{P})$. Without loss of generality, we assume that $(\Omega, \mathcal{F}, \mathsf{P})$ is a canonical probability space and the $\{X_n, Y_n\}_{n\geq 0}$ is a coordinate process taking values on the product space $\mathcal{N}\times\mathcal{O}$, where finite sets $\mathcal{N} = \{1, \ldots, N\}$ and $\mathcal{O} = \{1, \ldots, O\}$ denote the state space and observation space, respectively.

The unobserved *state process* $\{X_n\}_{n\geq 0}$ is a time-homogeneous Markov chain with the initial distribution μ and the transition matrix A. For any time $n \geq 0$ and any $i, j \in \mathcal{N}$,

$$\mathsf{P}(X_0 = i) = \mu_i, \quad \mathsf{P}(X_{n+1} = j | X_n = i) = A_{ij}.$$

The n-step distribution of the chain is then given by $\mathsf{P}(X_n = i) = (\mu A^n)_i$.

The *observation process* $\{Y_n\}_{n\geq 0}$ are mutually independent conditioned on the state process of the Markov chain,

i.e., for any time $n \geq 0$, any $i_0, \ldots, i_n \in \mathcal{N}$, and any $r_0, \ldots, r_n \in \mathcal{O}$,

$$\mathsf{P}(Y_n = r_n, \ldots, Y_0 = r_0 | X_n = i_n, \ldots, X_0 = i_0) = \prod_{k=0}^{n} \mathsf{P}(Y_k = r_k | X_k = i_k).$$

The conditional distribution of Y_n only depends on X_n, which can be described by the transition matrix C. For any time $n \geq 0$, any $i \in \mathcal{N}$, and any $r \in \mathcal{O}$,

$$\mathsf{P}(Y_n = r | X_n = i) = C_{ir}.$$

For any $r \in \mathcal{O}$, denote the diagonal matrix $B(r) := \mathrm{diag}(b_i(r))$, where the vector $b(r) = [C_{1r}, C_{2r}, \ldots, C_{Nr}]^T$.

The complete statistics of the HMM $\{X_n, Y_n\}_{n\geq 0}$ are fully characterized by a model, denoted by $\xi = (\mu, A, C)$. For an HMM with the parameter set ξ, we denote the probability measure and associated expectation as P_ξ and E_ξ, respectively. We make following two assumptions:

Assumption 1 (Ergodicity) *All Markov chains are assumed to be* irreducible *and* aperiodic.

Under Assumption 1, there exists a unique *invariant distribution* π such that $\pi = \pi A$. In fact, the chain is *geometrically ergodic*, i.e., the n-step distribution of the chain converges geometrically fast to the invariant distribution π in total variation sense [1].

Assumption 2 (Nondegeneracy) *The transition matrix C is strictly positive, i.e., $C_{ir} > 0$ for any $i \in \mathcal{N}$ and $r \in \mathcal{O}$.*

Under Assumption 2, the unobserved state process $\{X_n\}_{n\geq 0}$ can be statistically inferred from any sample path of observations of the observed process $\{Y_n\}_{n\geq 0}$.

B. Filter recursion and its stability

For an HMM, an important problem is to compute the *prediction filter*: For any time $n \geq 0$ and any $i \in \mathcal{N}$,

$$p_n(i) := \mathsf{P}(X_n = i | Y_{n-1}, \ldots, Y_0)$$

where we take $p_0 = \mu$. The prediction filter is used to obtain the *predictive distribution* of the observations: For any $n \geq 0$,

$$\mathsf{P}(Y_n | Y_{n-1}, \ldots, Y_0) = b^T(Y_n) p_n. \tag{1}$$

The solution to the HMM filtering problem is *recursive* in nature. For any time $n \geq 0$,

$$p_{n+1} = \frac{A^T B(Y_n) p_n}{b^T(Y_n) p_n}. \tag{2}$$

The recursive nature of the filter is inherited from the Markovian nature of the state process, and is computationally very convenient for on-line estimation.

The recursive filter defined in (2) is *exponentially stable* for general HMMs [1], [12] under egodicity and nondegeneracy assumptions. Here we state the results of [1] for HMMs defined on finite state and observation spaces.

Proposition 1 *Suppose Assumption 1 and Assumption 2 hold. Then, for any two distributions μ and ν, there exists constants $0 < C_1 < \infty$, $0 < C_2 < \infty$, and $0 < \rho < 1$ such that*

(i) For any $n \geq 0$,

$$\|p^\mu_{n+1} - p^\nu_{n+1}\|_{\mathrm{TV}} \leq C_1 (1-\rho)^n \|\mu - \nu\|_{\mathrm{TV}}$$

where p^μ_{n+1} and p^ν_{n+1} denote two filter recursions defined in (2) starting with initial distributions μ and ν, respectively.

(ii) For any $0 \leq k \leq n$,

$$\|\mathsf{P}^\mu(X_{n+1} | Y_0^n) - \mathsf{P}^\mu(X_{n+1} | Y_k^n)\|_{\mathrm{TV}} \leq C_2 (1-\rho)^{n-k}$$

where P^μ denotes the probability measure with the initial distribution μ.

The stability of the filter implies that the extended Markov chain $\{X_n, Y_n, p_n\}_{n\geq 0}$ is geometrically ergodic [12]. Thus the initial distributions are forgotten exponentially fast and are hence asymptotically not important in the analysis of the filtering process.

C. Probability distance between HMMs

In this section, we define the probability distance between two HMMs using the Kullback-Leibler divergence rate. For two HMMs $\xi = (\mu, A, C)$ and $\bar{\xi} = (\bar{\mu}, \bar{A}, \bar{C})$ defined on the same observation space $\mathcal{O}$ (but not necessarily on the same state space), we consider the K-L rate between laws of the observations [7]:

$$\begin{aligned} R(\xi \| \bar{\xi}) &:= \lim_{n\to\infty} \frac{1}{n} D\left(\mathsf{P}_\xi(Y_0^n) \| \mathsf{P}_{\bar{\xi}}(Y_0^n)\right) \\ &= \lim_{n\to\infty} \frac{1}{n} \mathsf{E}_\xi \left[\log \frac{\mathsf{P}_\xi(Y_0^n)}{\mathsf{P}_{\bar{\xi}}(Y_0^n)}\right]. \end{aligned}$$

As shown in [9], [11], the following asymptotic results can be established under Assumption 1 and Assumption 2: There exist finite constants $H(\xi, \xi)$ and $H(\xi, \bar{\xi})$ such that the following limits exist in P_ξ-a.s. sense:

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{P}_\xi(Y_0^n) = \lim_{n\to\infty} \frac{1}{n} \mathsf{E}_\xi \left[\log \mathsf{P}_\xi(Y_0^n)\right] = H(\xi, \xi) \tag{3}$$

$$\lim_{n\to\infty} \frac{1}{n} \log \mathsf{P}_{\bar{\xi}}(Y_0^n) = \lim_{n\to\infty} \frac{1}{n} \mathsf{E}_\xi \left[\log \mathsf{P}_{\bar{\xi}}(Y_0^n)\right] = H(\xi, \bar{\xi}). \tag{4}$$

The convergence of (3) follows directly from the Shannon-McMillan-Breiman theorem for finite-valued stationary ergodic process [9] and the limit $H(\xi, \xi)$ is equal to the entropy rate of the observation process $\{Y_n\}_{n\geq 0}$. The convergence of (4) was first established in [13] for finite-valued stationary ergodic HMMs.

Thus, the probability distance between two HMMs is well-defined through the K-L rate between laws of the observations:

$$R(\xi \| \bar{\xi}) = H(\xi, \xi) - H(\xi, \bar{\xi}). \tag{5}$$

In general, we do not have an explicit expression for $R(\xi \| \bar{\xi})$ in terms of parameters of HMMs ξ and $\bar{\xi}$. The prediction filter is usually employed to approximate the K-L rate given a sufficient number of observations [6], [11].

III. Model Reduction of HMM

A. Reduction via aggregation of state space

Consider the HMM $\xi = (\mu, A, C)$ defined on the state space $\mathcal{N}$ and the observation space $\mathcal{O}$. We want to find another HMM $\bar{\xi} = (\bar{\mu}, \bar{A}, \bar{C})$ defined on the state space $\mathcal{M} = \{1, \ldots, M\}$ with cardinality $M \leq N$ and the observation space $\mathcal{O}$ such that the probability distance $R(\xi \| \bar{\xi})$ is minimized. Additionally, we want the reduced HMM $\bar{\xi}$ to be obtained by aggregating the state space of the HMM ξ. The relationship between $\mathcal{N}$ and $\mathcal{M}$ is described by a partition function ϕ.

Definition 1 *Let $\mathcal{N} = \{1, 2, \ldots, n\}$ and $\mathcal{M} = \{1, 2, \ldots, m\}$ be two finite state spaces with $m \leq n$. A* partition function $\phi : \mathcal{N} \mapsto \mathcal{M}$ *is a surjective function from $\mathcal{N}$ onto $\mathcal{M}$, and $\phi^{-1}(k)$ denotes the k^{th} group in $\mathcal{N}$.*

Let Φ denote all M-partition functions from $\mathcal{N}$ to $\mathcal{M}$.

As shown in our prior work [4], [5], an optimal representation of the aggregated HMM, (6)–(8) below, is obtained by minimizing the *K-L rate between joint laws of the states and observations* together. Given the focus vision to K-L rate between laws of the observations, it would have been ideal to construct an optimal model based on the observations alone. This however is a difficult problem. Instead, we use the representation (6)–(8) for the aggregated HMM. The problem of optimal partition selection is based on the K-L rate between laws of the observations.

For any fixed partition function $\phi \in \Phi$, the aggregated HMM $\bar{\xi}(\phi) = (\bar{\mu}(\phi), \bar{A}(\phi), \bar{C}(\phi))$ is represented as a function of ϕ (see e.g. Theorem 2 of [4]).

$$\bar{\mu}_k(\phi) = \sum_{i \in \phi^{-1}(k)} \mu_i, \quad k \in \mathcal{M} \tag{6}$$

$$\bar{A}_{kl}(\phi) = \frac{\sum_{i \in \phi^{-1}(k)} \pi_i \sum_{j \in \phi^{-1}(l)} A_{ij}}{\sum_{i \in \phi^{-1}(k)} \pi_i}, \quad k, l \in \mathcal{M} \tag{7}$$

$$\bar{C}_{kr}(\phi) = \frac{\sum_{i \in \phi^{-1}(k)} \pi_i C_{ir}}{\sum_{i \in \phi^{-1}(k)} \pi_i}, \quad k \in \mathcal{M},\ r \in \mathcal{O}. \tag{8}$$

For any fixed $\phi \in \Phi$, we observe that the aggregated HMM $\bar{\xi}(\phi)$ satisfies both Assumption 1 and Assumption 2, i.e., the underlying aggregated Markov chain with the transition matrix $\bar{A}(\phi)$ is ergodic and the transition matrix $\bar{C}(\phi)$ is non-degenerate. Thus the probability distance $R(\xi \| \bar{\xi}(\phi))$ is well-defined for any $\phi \in \Phi$. We also observe that:

$$\begin{aligned}
\bar{\mu}_k(\phi) &= \mathsf{P}^{\mu}_{\xi}(X_0 \in \phi^{-1}(k)) \\
\bar{A}_{kl}(\phi) &= \mathsf{P}^{\pi}_{\xi}(X_{n+1} \in \phi^{-1}(l) | X_n \in \phi^{-1}(k)) \\
\bar{C}_{kr}(\phi) &= \mathsf{P}^{\pi}_{\xi}(Y_n = r | X_n \in \phi^{-1}(k))
\end{aligned}$$

where P^{μ}_{ξ} and P^{π}_{ξ} denote probability measures with initial distributions μ and π, respectively. This result is consistent with the optimal prediction theory from the statistical mechanics [14].

B. Maximum likelihood estimation formulation

For a fixed partition function $\phi \in \Phi$, the aggregated HMM is represented as $\bar{\xi}(\phi) = (\bar{\mu}(\phi), \bar{A}(\phi), \bar{C}(\phi))$. The problem then is to find the optimal ϕ^* such that

$$\phi^* \in \arg\min_{\phi \in \Phi} R(\xi \| \bar{\xi}(\phi))$$

which, after using (5), is equivalent to the following maximization problem:

$$\phi^* \in \arg\max_{\phi \in \Phi} H(\xi, \bar{\xi}(\phi)). \tag{9}$$

Due to the almost sure convergence of log-likelihood function to the limit $H(\xi, \bar{\xi}(\phi))$ (see (4)), we instead consider the following stochastic counterpart of (9):

$$\widehat{\phi}_n \in \arg\max_{\phi \in \Phi} l_n(\phi) \tag{10}$$

where the *log-likelihood rate* is defined as

$$l_n(\phi) := \frac{1}{n} \log \mathsf{P}_{\bar{\xi}(\phi)}(y_0^n) \tag{11}$$

with observations $\{y_0, \ldots, y_n\}$ generated from the HMM ξ.

The optimization problem (10) is the *maximum likelihood estimation* in statistics: In effect, we select the partition function which gives the highest probability of the observations generated from the true model. Note that the objective function (10) converges to the objective function of (9) in P_{ξ}-a.s. sense. One may wonder whether $\widehat{\phi}_n \to \phi^*$ P_{ξ}-a.s. as $n \to \infty$. The answer to this question is affirmative due to the fact that the partition function space Φ is a finite set.

Proposition 2 *Let Φ denote a finite partition function space and consider an equivalent class in Φ*

$$\Phi^e := \{\phi \in \Phi : \mathsf{P}_{\bar{\xi}(\phi)} = \mathsf{P}_{\bar{\xi}(\phi^*)} \text{ for almost all } \{Y_n\}_{n \geq 0}\}.$$

Then P_{ξ}-a.s.,

(i) For any $\phi \in \Phi$, we have $H(\xi, \bar{\xi}(\phi)) \leq H(\xi, \bar{\xi}(\phi^*))$ where the equality holds if and only if $\phi \in \Phi^e$.

(ii) Maximum likelihood estimation is consistent: $\widehat{\phi}_n \to \phi^e$ as $n \to \infty$ for some $\phi^e \in \Phi^e$.

C. Hypothesis testing-based approach for optimal partition selection

Since the partition function space Φ is a finite set, the optimization problem (10) can in practice be approached through the *hypothesis testing*: We are given $|\Phi|$ different hypotheses (or $|\Phi|$ different aggregated HMMs), and our goal is to decide on the basis of observations alone which of the hypotheses holds true (or which of the aggregated HMM is with the maximum log-likelihood rate). If the set Φ is of moderate size, then the maximum log-likelihood rate hypothesis can be found efficiently. All we need to do is to compute $|\Phi|$ different filters, one for each partition function. For any fixed-length observations $\{y_0, y_1, \ldots, y_n\}$, we choose the n-step hypothesis $\widehat{\phi}_n$ as the one with the largest log-likelihood rate. Then $\widehat{\phi}_n$ asymptotically converges to the global maximum ϕ^* as $n \to \infty$ (see Proposition 2).

IV. Recursive Learning Algorithm

In general, the optimization problem (10) is intractable because of the curse of dimensionality. The curse here arises due to the large size of the partition function space, e.g., $L = |\Phi| = M^N$ for the M-partition of the N-state space. To confront this complexity issue, a parametric representation is used to represent the partition function in terms of a small number of parameters. A recursive learning algorithm is described to adaptively update the parameters based on a sample path of the HMM.

A. Parameterization of the partition function space via randomization

The randomization of the partition function gives us greater flexibility to solve the optimization problem (10). A *randomized partition policy* is defined as a mapping,

$$\eta : \mathcal{N} \to [0,1]^L$$

with the component $\eta_\phi(i)$ such that $\sum_{\phi\in\Phi} \eta_\phi(i) = 1$ for every $i \in \mathcal{N}$. Under a policy η, the partition function ϕ is assigned to the state i with the probability $\eta_\phi(i)$, independent of everything else.

The policy is said to be deterministic if for every state i, there is a single partition function $\phi^{(i)}$ such that $\eta_{\phi^{(i)}}(i) = 1$. If the function $\phi^{(i)}$ is the same for all i then the policy η yields a consistent partition of the space $\mathcal{N}$. If $\eta(\cdot)$ is a degenerate probability distribution (i.e., a dirac delta in the probability simplex of Φ), then a partition function can be uniquely obtained from $\eta(\cdot)$. In practice, a numerical method will in general only lead to a partition function with high probability determined by $\eta(\cdot)$.

The combinatorial optimization problem (10) involves a very large partition space Φ. Following the consideration of [4], we consider the randomized policies $\eta(\cdot;\theta)$ which are described in terms of a parameter vector $\theta = (\theta(1), \ldots, \theta(K))^T$, where the dimension K is chosen much smaller than L, the dimension of Φ.

The following assumption is made for the ease of the optimization over the parameter θ:

Assumption 3 *The parameter space Θ is a compact subset of a K-dimensional real vector space $\mathbb{R}^K$. For any $i \in \mathcal{N}$, the randomized and parameterized policy $\eta(i;\theta)$ is twice differentiable with respect to θ, and has bounded first and second derivatives for all $\theta \in \Theta$.*

B. Parametric representation of the MLE problem

For any $\theta \in \Theta$, we consider a randomized partition policy $\eta(\cdot;\theta)$ such that for every $i \in \mathcal{N}$, $\eta(i;\theta)$ depends smoothly on θ, $\eta_\phi(i;\theta) \geq 0$, and $\sum_{\phi\in\Phi} \eta_\phi(i;\theta) = 1$. We associate a probability measure $\mathsf{P}_{\eta(\cdot;\theta)}$ and the corresponding expectation $\mathsf{E}_{\eta(\cdot;\theta)}$ with the policy $\eta(\cdot;\theta)$. For any measurable function $f(\phi)$, we define

$$\mathsf{E}_{\eta(\cdot;\theta)}[f(\phi)] := \sum_{\phi\in\Phi} \eta_\phi(\cdot;\theta) f(\phi).$$

The parameterized one-step log-likelihood can also be defined: For any $n \geq 0$,

$$g_n(\theta) := \mathsf{E}_{\eta(X_n;\theta)} \left[\log \left(\mathsf{P}_{\bar{\xi}(\phi)}(Y_n | Y_0^{n-1}) \right) \right]$$

where X_n is the hidden state associated with the observation Y_n generated from the HMM ξ.

The parameterized maximization problem is defined as

$$\theta^* \in \arg\max_{\theta\in\Theta} \widetilde{H}(\theta) \tag{12}$$

where the parameterized average cost is given by

$$\widetilde{H}(\theta) = \lim_{n\to\infty} \frac{1}{n} \mathsf{E}_\xi \left[\sum_{k=0}^{n} g_k(\theta) \right].$$

The parameterized maximum likelihood estimation (MLE) is the stochastic counterpart of (12):

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} \tilde{l}_n(\theta) \tag{13}$$

where the parameterized log-likelihood rate is defined as

$$\tilde{l}_n(\theta) = \frac{1}{n} \sum_{k=0}^{n} g_k(\theta).$$

C. Recursive learning algorithm and its convergence

Under Assumption 1–3, one can show that the MLE $\widehat{\theta}_n$ converge to θ^* P_ξ-a.s as $n \to \infty$. However, the maximum of (12) or (13) with respect to θ is typically very difficult to compute. Instead, we describe a recursive learning algorithm that searches for a maximum along the gradient-ascent direction of the log-likelihood rate $\tilde{l}_n(\theta)$.

In order to compute the gradient of $\tilde{l}_n(\theta)$, we employ the simulation to produce a sample-based estimate $\nabla h_n(\theta)$ of $\nabla \tilde{l}_n(\theta)$ (we denote $\nabla := \nabla_\theta$ for short). At every time step n, the estimate h_n is computed using the current observation as well as finite length of past observations: For any time $n \geq 0$,

$$h_n(\theta) := \frac{1}{\lfloor m_n \rfloor + 1} \left(\sum_{k=n-\lfloor m_n \rfloor}^{n} \widetilde{g}_k(\theta) \right) \tag{14}$$

where the finite-length log-likelihood

$$\widetilde{g}_k(\theta) := \mathsf{E}_{\eta(X_k;\theta)} \left[\log \left(\mathsf{P}_{\bar{\xi}(\phi)}(Y_k | Y_{n-\lfloor m_n \rfloor}^{k-1}) \right) \right],$$

and the averaging sequence $\{m_n\}_{n\geq 0}$ satisfies the following assumption:

Assumption 4 *For any $n \geq 0$,*

$$0 \leq m_0 \leq m_1 \leq \ldots \leq m_{n-1} \leq m_n \leq n,$$

and as $n \to \infty$, $m_n \to \infty$.

Given any partition function ϕ and the observations $\{Y_{n-\lfloor m_n \rfloor}, \ldots, Y_n\}$, the estimate h_n can be computed through the filter recursion (2) of $\{\mathsf{P}_{\bar{\xi}(\phi)}(Y_{n-\lfloor m_n \rfloor}), \ldots, \mathsf{P}_{\bar{\xi}(\phi)}(Y_n | Y_{n-\lfloor m_n \rfloor}^{n-1})\}$. Due to the ergodicity of the filter (see Proposition 1 (ii)), the recursion can be started with an arbitrary initial distribution $\bar{\mu}$ on $\mathcal{M}$.

The estimate $\nabla h_n(\theta)$ asymptotically converges to $\nabla \tilde{l}_n(\theta)$ as $n \to \infty$ and the convergence is geometrically fast due to the ergodicity of the filter. By choosing the sequence $\{m_n\}_{n\geq 0}$ alternatively, we compute $h_n(\theta)$ efficiently: e.g., one can take $m_n = n^\alpha$ where selecting $\alpha \in (0, 1]$ allows one to tradeoff between the computation efficiency and the estimation performance.

A recursive learning algorithm is employed to approach the optimization problem (12). Let $\{x_n, y_n\}_{n\geq 0}$ denote a sample path generated from the HMM ξ. The recursive learning algorithm for updating the parameter vector is given by: For any $n \geq 0$,

$$\bar{\theta}_{n+1} = \bar{\theta}_n + \gamma_n \nabla h_n(\bar{\theta}_n) \tag{15}$$

where $\bar{\theta}_0$ is taken to be an arbitrary point in Θ, the value $\bar{\theta}_n$ is assumed to be available from the previous iteration, and $h_n(\theta)$ is computed using (14). In addition, another adaptive algorithm for updating the log-likelihood rate is run in parallel

$$\bar{l}_{n+1} = \bar{l}_n + \gamma_n (h_n(\bar{\theta}_n) - \bar{l}_n) \tag{16}$$

where $\bar{l}_n$ is the estimated log-likelihood rate and parameter $\bar{\theta}_n$ comes from (15). The diminishing stepsize γ_n satisfies the standard stochastic approximation conditions:

Assumption 5 *The stepsize values $\{\gamma_n\}_{n\geq 0}$ are non-negative and satisfy*

$$\sum_{n=0}^{\infty} \gamma_n = \infty, \quad \sum_{n=0}^{\infty} \gamma_n^2 < \infty.$$

The convergence of the simulation-based algorithm is established using the ODE method and ergodicity of the filtering process:

Proposition 3 *Suppose,*

(i) The sample path $\{x_n, y_n\}_{n\geq 0}$ are generated from the HMM ξ, which satisfies Assumption 1 and Assumption 2.

(ii) The randomized and parameterized policy $\eta(\cdot;\theta)$ satisfies Assumption 3.

(iii) The averaging sequence $\{m_n\}_{n\geq 0}$ and the stepsize sequence $\{\gamma_n\}_{n\geq 0}$ satisfy Assumption 4 and Assumption 5, respectively.

(iv) The parameter vector sequence $\{\bar{\theta}_n\}_{n\geq 0}$ and the log-likelihood rate sequence $\{\bar{l}_n\}_{n\geq 0}$ are updated according to the recursive learning algorithm (15) and (16), respectively.

Then, as $n \to \infty$, the sequence $\tilde{l}_n(\bar{\theta}_n)$ converges to a non-positive limit,

$$\nabla \tilde{l}_n(\bar{\theta}_n) \to 0 \quad \text{and} \quad \bar{l}_n \to \tilde{l}_n(\bar{\theta}_n),$$

all in P_ξ-a.s. sense.

D. A simple bi-partition parameterization

Let $\mathcal{M} = \{1, 2\}$ denote the reduced aggregated state space with two superstates. For the bi-partition problem, a partition function ϕ takes only two values, either $\phi(i) = 1$ or $\phi(i) = 2$ for any state $i \in \mathcal{N}$. Let Θ be a sufficiently large compact subset of $\mathbb{R}^N$. We consider a real-valued parameter vector $\theta := (\theta(1), \ldots, \theta(N))^T \in \Theta$, where $\theta(i)$ decides the group assignment for the state $i \in \mathcal{N}$. In particular, we use $\zeta(\theta(i)) := \frac{1}{1+\exp(M\theta(i))}$ to reflect the probability that $\phi(i) = 1$, where $M > 0$ is some positive constant.

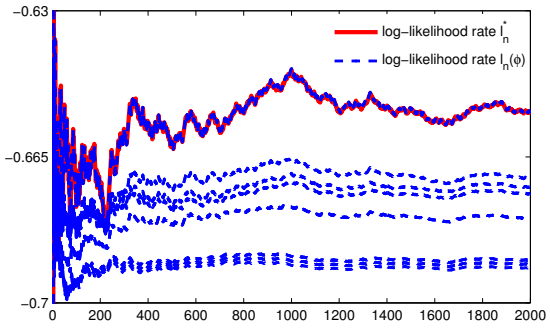


Fig. 1. The original log-likelihood rate l_n^* is compared with the 8 different aggregated log-likelihood rates $l_n(\phi)$.

At time n, we only need to consider the randomized and parameterized partition policy for the state X_n. Suppose the current stat is $X_n = i \in \mathcal{N}$, and partition function at time $n-1$ is $\widetilde{\phi}$. The policy is defined for all $\phi \in \Phi$:

- If $\phi(j) = \widetilde{\phi}(j)$ for every $j \in \mathcal{N}/\{i\}$, then
$$\eta_\phi(i;\theta) = \zeta(\theta(i))\mathbb{1}_{\{\phi(i)=1\}} + (1 - \zeta(\theta(i)))\mathbb{1}_{\{\phi(i)=2\}}.$$
- Otherwise, $\eta_\phi(i;\theta) = 0$.

One can easily verify that the policy satisfies the Assumption 3. At each time step, the policy only affects or changes the probability of the group assignment for the state X_n and keep others unchanged. Thus this policy can save a lot of computations at each time-step, which makes it more suitable for on-line estimation.

V. Simulation and Discussion

In this section, we use a simple HMM ξ to illustrate the theoretical results and algorithms described in this paper. The HMM $\xi = (\mu, A, C)$ has 4 states and 2 observations. The transition matrices

$$A = \begin{bmatrix} 0.500 & 0.200 & 0.225 & 0.075 \\ 0.200 & 0.500 & 0.135 & 0.165 \\ 0.030 & 0.270 & 0.500 & 0.200 \\ 0.150 & 0.165 & 0.185 & 0.500 \end{bmatrix}, C = \begin{bmatrix} 0.15 & 0.85 \\ 0.05 & 0.95 \\ 0.89 & 0.11 \\ 0.88 & 0.12 \end{bmatrix}$$

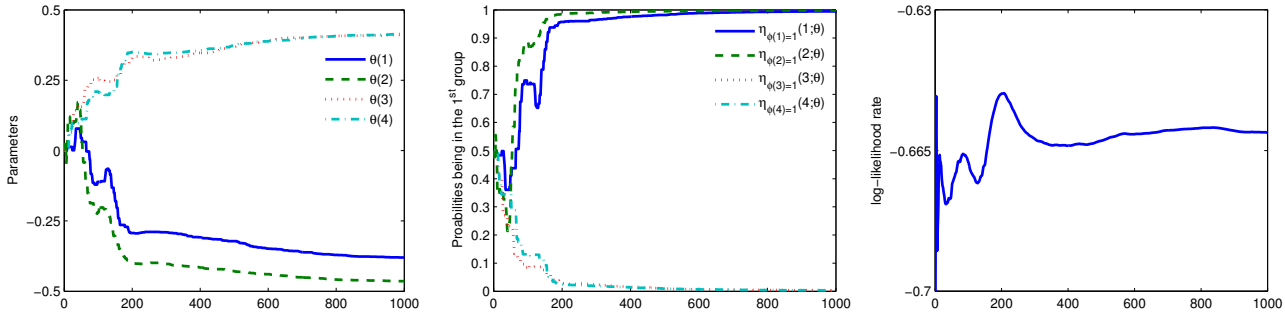
with the initial distribution $\mu = \pi$, the invariant distribution of A.

We consider the bi-partition problem of the HMM ξ here, i.e., the state space $\mathcal{N} = \{1, 2, 3, 4\}$ is aggregated into the state space $\mathcal{M} = \{1, 2\}$.

A. Hypothesis testing approach for a simple HMM

Note that the partition function space Φ is of a moderate size ($|\Phi| = 2^4 = 16$). Thus the hypothesis testing method is employed in this subsection to find the optimal partition function as described in Section III-C.

First, a sample path of $n = 2000$ observations $\{y_0, \ldots, y_n\}$ is generated according to the HMM ξ. The



(a) (b) (c)

Fig. 2. Plots of (a) the estimated parameter vector $\bar{\theta}_n$, (b) probabilities of the states being in the first group $\eta_{\phi=[1,1,1,1]}(\cdot;\bar{\theta}_n)$, and (c) the estimated log-likelihood rate $\bar{l}_n$ for the HMM ξ with the recursive learning algorithm (15) and (16).

original log-likelihood rate $l_n^* = n^{-1}\log \mathsf{P}_\xi(y_0^n)$ is computed based on the recursive filer of the HMM ξ (see Section II-B for more details).

Second, for any fixed $\phi \in \Phi$, the aggregated HMM $\bar{\xi}(\phi)$ is obtained using the representation (6)–(8). Then we compute the aggregated log-likelihood rate $l_n(\phi) = n^{-1}\log \mathsf{P}_{\bar{\xi}(\phi)}(y_0^n)$ (11) for every aggregated HMM $\bar{\xi}(\phi)$ based on the recursive filter of $\bar{\xi}(\phi)$. Note that if the partition functions ϕ_1 and ϕ_2 are symmetric (e.g., $\phi_1 = [1,2,2,2]$ and $\phi_2 = [2,1,1,1]$ are symmetric), then the probability laws $\mathsf{P}_{\bar{\xi}(\phi_1)} = \mathsf{P}_{\bar{\xi}(\phi_2)}$ for almost all observations. Based on the symmetry of the problem, we only need to consider 8 partition functions for the hypothesis testing. In Fig. 1, we depict the original log-likelihood rate l_n^* as well as 8 different aggregated log-likelihood rates $l_n(\phi)$ (two symmetric partition functions correspond to the same log-likelihood rate).

Finally, we choose the optimal partition function corresponding to the largest log-likelihood rate. For this example, the optimal partition functions is $\phi^* = [1,1,2,2]$ or $\phi^* = [2,2,1,1]$. The two corresponding aggregated HMMs are equivalent up to the permutation of the state space. We also note that for this special example the optimal aggregated log-likelihood rate is almost the same as the original one.

B. Recursive learning approach

From the hypothesis testing of all partition functions, we know that $\phi = [1,1,2,2]$ is the optimal bi-partition of the HMM ξ. In this subsection, we apply the recursive learning algorithm (15) and (16) to find the optimal partition function based on a single sample-path $\{x_n, y_n\}_{n\geq 0}$ of the HMM ξ.

The randomized and parameterized bi-partition policy, with the constant $M = 15$, is chosen for the recursive learning algorithm as described in Section IV-D. The averaging sequence is taken as $m_n = n^{0.8}$ and the stepsize sequence is taken as $\gamma_n = \frac{1}{n+1}$ for $n \geq 0$. The parameter space Θ is a sufficiently large compact subset of $\mathbb{R}^N$ and the algorithm is initialized with the parameter vector $\bar{\theta}_0 = [0,0,0,0]$.

In Fig. 2, we depict a typical run of the recursive learning algorithm for the 1000 iterations. After $n = 1000$ iterations, the estimated parameter vector $\bar{\theta}_n = [-0.3802, -0.4643, 0.4117, 0.4154]$, and the probabilities of states being in the first group are $\eta_{\phi=[1,1,1,1]}(\cdot;\bar{\theta}_n) = [0.9948, 0.9984, 0.0034, 0.0032]$. From this, the optimal partition function $\phi = [1,1,2,2]$ can be determined with high probability. The corresponding estimated log-likelihood rate is equal to $\bar{l}_n = -0.6605$, which is close to maximum log-likelihood rate depicted in Fig. 1. The recursive learning algorithm thus recovers the optimal partition function for this example.

REFERENCES

[1] O. Cappé, E. Moulines, and T. Ryden, *Inference in Hidden Markov Models (Springer Series in Statistics)*. Secaucus, NJ: Springer-Verlag New York, Inc., 2005.

[2] R. G. Phillips and P. V. Kokotovic, "A singular perturbation approach to modeling and control of Markov chains," *IEEE Trans. Automat. Contr.*, vol. 26, no. 5, pp. 1087–1094, 1981.

[3] K. Deng, Y. Sun, P. G. Mehta, and S. P. Meyn, "An information-theoretic framework to aggregate a Markov chain," in *Proceedings of American Control Conference*, St. Louis, MO, 2009, pp. 731–736.

[4] K. Deng, P. Mehta, and S. Meyn, "Aggregation-based model reduction of a Hidden Markov Model," in *Proceedings of IEEE Conference of Decision and Control*, Atlanta, GA, 2010, pp. 6183–6188.

[5] M. Vidyasagar, "Reduced-order modeling of Markov and Hidden Markov Processes via aggregation," in *Proceedings of IEEE Conference of Decision and Control*, Atlanta, GA, 2010, pp. 1810–1815.

[6] B. G. Leroux, "Maximum-likelihood estimation for hidden Markov models," *Stochastic Process. Applic.*, vol. 40, no. 1, pp. 127–143, 1992.

[7] B. H. Juang and L. R. Rabiner, "A probabilistic distance measure for Hidden Markov Models," *AT&T Tech. J.*, vol. 64, no. 2, pp. 391–408, 1985.

[8] M. Vidyasagar, S. S. Mande, C. V. S. K. Reddy, and V. V. R. Rao, "The 4M algorithm for finding genes in prokaryotic genomes," *IEEE Trans. Automat. Contr.*, vol. 53, pp. 26–37, 2008, Special Issue.

[9] L. Xie, V. Ugrinovskii, and I. R. Petersen, "Probabilistic distances between finite-state finite-alphabet Hidden Markov Models," *IEEE Trans. Automat. Contr.*, vol. 50, no. 4, pp. 505–511, 2005.

[10] Y. Sun and P. G. Mehta, "The Kullback-Leiber vrate pseudo-metric for comparing dynamical systems," *IEEE Trans. Automat. Contr.*, vol. 55, no. 7, pp. 1585–1598, 2010.

[11] R. Douc, E. Moulines, J. Olsson, and R. van Handel, "Consistency of the maximum likelihood estimator for general hidden markov models," *Ann. Statist.*, vol. 39, no. 1, pp. 474–513, 2011.

[12] F. L. Gland and L. Mevel, "Exponential forgetting and geometric ergodicity in hidden Markov models," *Math. of Control Signals and Systems*, vol. 13, pp. 63–93, 2000.

[13] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Ann. Math. Stat.*, vol. 37, no. 6, pp. 1559–1563, 1966.

[14] W. E, T. Li, and E. Vanden-Eijnden, "Optimal partition and effective dynamics of complex networks," *PNAS*, vol. 105, no. 23, pp. 7907–7912, 2008.