Learning Parametric-Output HMMs with Two Aliased States

Roi Weiss ROIWEI@CS.BGU.AC.IL
Department of Computer Science, Ben-Gurion University, Beer Sheva, 84105, Israel.

Boaz Nadler BOAZ.NADLER@WEIZMANN.AC.IL
Department of Computer Science and Applied Mathematics, The Weizmann Institute of Science, Rehovot, 76100, Israel.

Abstract

In various applications involving hidden Markov models (HMMs), some of the hidden states are *aliased*, having identical output distributions. The minimality, identifiability and learnability of such aliased HMMs have been long standing problems, with only partial solutions provided thus far. In this paper we focus on parametric-output HMMs, whose output distributions come from a parametric family, and that have exactly two aliased states. For this class, we present a complete characterization of their minimality and identifiability. Furthermore, for a large family of parametric output distributions, we derive computationally efficient and statistically consistent algorithms to detect the presence of aliasing and learn the aliased HMM transition and emission parameters. We illustrate our theoretical analysis by several simulations.

1. Introduction

HMMs are a fundamental tool in the analysis of time series. A discrete time HMM with n hidden states is characterized by a $n \times n$ transition matrix and by the emissions probabilities from these n states. In several applications, the HMMs, or more general processes such as partially observable Markov decision processes, are *aliased*, with some states having identical output distributions. In modeling of ion channel gating, for example, a common assumption is that at any given time an ion channel can be in only one of a finite number of hidden states, some of which are open and conducting current while others are closed, see e.g. Fredkin & Rice (1992). Fitting an aliased HMM to electric current measurements, allows biologists to gain important insights regarding the gating process. Other examples appear in the fields of reinforcement learning (Chrisman, 1992; McCallum, 1995; Brafman & Shani, 2004; Shani et al., 2005) and robot navigation (Jefferies & Yeap, 2008; Zatuchna & Bagnall, 2009). In the latter case, aliasing occurs whenever different spatial locations appear (statistically) identical to the robot, given its limited sensing devices. As a last example, HMMs with several silent states that do not emit any output (Leggetter & Woodland, 1994; Stanke & Waack, 2003; Brejova et al., 2007), can also be viewed as aliased.

Key notions related to the study of HMMs, be them aliased or not, are their minimality, identifiability and learnability:

> *Minimality.* Is there an HMM with fewer states that induces the same distribution over all output sequences?
>
> *Identifiability.* Does the distribution over all output sequences uniquely determines the HMM's parameters, up to a permutation of its hidden states?
>
> *Learning.* Given a long output sequence from a minimal and identifiable HMM, efficiently learn its parameters.

For non-aliased HMMs, these notions have been intensively studied and by now are relatively well understood, see for example Petrie (1969); Finesso (1990); Leroux (1992); Allman et al. (2009) and Cappé et al. (2005). The most common approach to learn the parameters of an HMM is via the Baum-Welch iterative algorithm (Baum et al., 1970). Recently, tensor decompositions and other computationally efficient spectral methods have been developed to learn non-aliased HMMs (Hsu et al., 2009; Siddiqi et al., 2010; Anandkumar et al., 2012; Kontorovich et al., 2013).

In contrast, the minimality, identifiability and learnability of aliased HMMs have been long standing problems, with only partial solutions provided thus far. For example, Blackwell & Koopmans (1957) characterized the identifiability of a specific aliased HMM with 4 states. The identifiability of deterministic output HMMs, where each hidden state outputs a deterministic symbol, was partially resolved by Ito et al. (1992). To the best of our knowledge, precise characterizations of the minimality, identifiability and learnability of probabilistic output HMMs with aliased

Proceedings of the 32^{nd} *International Conference on Machine Learning*, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

states are still open problems. In particular, the recently developed tensor and spectral methods mentioned above, explicitly require the HMM to be non-aliasing, and are not directly applicable to learning aliased HMMs.

Main results. In this paper we study the minimality, identifiability and learnability of parametric-output HMMs that have *exactly two* aliased states. This is the simplest possible class of aliased HMMs, and as shown below, even its analysis is far from trivial. Our main contributions are as follows: First, we provide a complete characterization of their minimality and identifiability, deriving necessary and sufficient conditions for each of these notions to hold. Our identifiability conditions are easy to check for any given 2-aliased HMM, and extend those derived by Ito et al. (1992) for deterministic outputs. Second, we address the problem of learning a possibly aliased HMM from a long sequence of its outputs. To this end, we first derive an algorithm to *detect* whether an observed output sequence corresponds to a non-aliased HMM or to an aliased one. In the former case, the HMM can be learned by various methods, such as Anandkumar et al. (2012); Kontorovich et al. (2013). In the latter case we show how the aliased states can be identified and present a method to recover the HMM parameters. Our approach is applicable to any family of output distributions whose mixtures are efficiently learnable. Examples include high dimensional Gaussians and products distributions, see Feldman et al. (2008); Belkin & Sinha (2010); Anandkumar et al. (2012) and references therein. After learning the output mixture parameters, our moment-based algorithm requires only a single pass over the data. As far as we know, it is the first statistically consistent and computationally efficient scheme to handle 2-aliased HMMs. While our approach may be extended to more complicated aliasing, such cases are beyond the scope of this paper. We conclude with some simulations illustrating the performance of our proposed algorithms.

2. Definitions & Problem Setup

Notation. We denote by I_n the $n \times n$ identity matrix and $\mathbf{1}_n = (1, \dots, 1)^\intercal \in \mathbb{R}^n$. For $\boldsymbol{v} \in \mathbb{R}^n$, $\mathrm{diag}(\boldsymbol{v})$ is the $n \times n$ diagonal matrix with entries v_i on its diagonal. The i-th row and column of a matrix $A \in \mathbb{R}^{n \times n}$ are denoted by $A_{[i,\cdot]}$ and $A_{[\cdot,i]}$, respectively. We also denote $[n] = \{1, 2, \dots, n\}$. For a discrete random variable X we abbreviate $P(x)$ for $\Pr(X = x)$. For a second random variable Z, the quantity $P(z \,|\, x)$ denotes either $\Pr(Z = z \,|\, X = x)$, or the conditional density $p(Z = z | X = x)$, depending on whether Z is discrete or continuous.

Hidden Markov Models. Consider a discrete-time HMM with n hidden states $\{1, \dots, n\}$, whose output alphabet $\mathcal{Y}$ is either discrete or continuous. Let $\mathcal{F}_\theta = \{f_\theta : \mathcal{Y} \to \mathbb{R} \,|\, \theta \in \Theta\}$ be a family of *parametric* probability density functions where Θ is a suitable parameter space. A *parametric-output* HMM is defined by a tuple $H = (A, \boldsymbol{\theta}, \boldsymbol{\pi}^0)$ where A is the $n \times n$ transition matrix between the hidden states

$$A_{ij} = \Pr(X_{t+1} = i \,|\, X_t = j) = P(i \,|\, j),$$

$\boldsymbol{\pi}^0 \in \mathbb{R}^n$ is the distribution of the initial state, and the vector of parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n) \in \Theta^n$ determines the n probability density functions $(f_{\theta_1}, f_{\theta_2}, \dots, f_{\theta_n})$.

To produce the HMM's output sequence, first a Markov sequence of hidden states $x = (x_t)_{t=0}^{T-1}$ is generated according to the distribution

$$P(x) = \pi^0_{x_0} \prod_{t=1}^{T-1} P(x_t \,|\, x_{t-1}).$$

Next, the output sequence $y = (y_t)_{t=0}^{T-1}$, where the output y_t at time t depends only on x_t, is generated according to

$$P(y \,|\, x) = \prod_{t=0}^{T-1} P(y_t \,|\, x_t) = \prod_{t=0}^{T-1} f_{\theta_{x_t}}(y_t).$$

We denote by $P_{H,k} : \mathcal{Y}^k \to \mathbb{R}$ the joint distribution of the first k consecutive outputs of the HMM H. For $y = (y_0, \dots, y_{k-1}) \in \mathcal{Y}^k$ this distribution is given by

$$P_{H,k}(y) = \sum_{x \in [n]^k} P(y \,|\, x) P(x).$$

Further we denote by $\mathcal{P}_H = \{P_{H,k} \,|\, k \geq 1\}$ the set of all these distributions.

2-Aliased HMMs. For an HMM H with output parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_n) \in \Theta^n$ we say that states i and j are *aliased* if $\theta_i = \theta_j$. In this paper we consider the special case where H has *exactly two* aliased states, denoted as 2A-HMM. Without loss of generality, we assume the aliased states are the two last ones, $n-1$ and n. Thus, $\theta_i \neq \theta_j$ for all $1 \leq i < j \leq n-1$, whereas $\theta_{n-1} = \theta_n$.

We denote the vector of the $n-1$ *unique* output parameters of H by $\bar{\boldsymbol{\theta}} = (\theta_1, \theta_2, \dots, \theta_{n-2}, \theta_{\overline{n-1}}) \in \Theta^{n-1}$. For future use, we define the *aliased kernel* $\bar{K} \in \mathbb{R}^{(n-1) \times (n-1)}$ as the matrix of inner products between the $n-1$ different f_{θ_i}'s,

$$\bar{K}_{ij} \equiv \langle f_{\theta_i}, f_{\theta_j} \rangle = \int_{\mathcal{Y}} f_{\theta_i}(y) f_{\theta_j}(y) \mathrm{d}y, \quad i, j \in [n-1]. \quad (1)$$

Assumptions. As in previous works (Leroux, 1992; Kontorovich et al., 2013), we make the following standard assumptions:

(**A1**) The parametric family $\mathcal{F}_\theta$ of the output distributions is linearly independent of order n: for any distinct $\{\theta_i\}_{i=1}^n$, $\sum_{i=1}^n a_i f_{\theta_i} \equiv 0$ iff $a_i = 0$ for all $i \in [n]$.

(**A2**) The transition matrix A is ergodic and its unique stationary distribution $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$ is positive.

Note that assumption (A1) implies that the parametric family $\mathcal{F}_\theta$ is *identifiable*, namely $f_\theta = f_{\theta'}$ iff $\theta = \theta'$. It also implies that the kernel matrix $\bar{K}$ of (1) is full rank $n{-}1$.

3. Decomposing the transition matrix A

The main tool in our analysis is a novel decomposition of the 2A-HMM's transition matrix into its non-aliased and aliased parts. As shown in Lemma 1 below, the aliased part consists of three rank-one matrices, that correspond to the exit from, entrance to, and dynamics within the two aliased states. This decomposition is used to derive the conditions for minimality and identifiability (Section 4), and plays a key role in learning the HMM (Section 5).

To this end, we introduce a *pseudo-state* $\bar{n}$, combining the two aliased states $n{-}1$ and n. We define

$$\pi_{\bar{n}} = \pi_{n-1} + \pi_n \quad \text{and} \quad \beta = \pi_{n-1}/\pi_{\bar{n}}. \tag{2}$$

We shall make extensive use of the following two matrices:

$$B = \left(\begin{array}{ccc|cc} & & & 0 & 0 \\ & I_{n-2} & & \vdots & \vdots \\ & & & 0 & 0 \\ \hline 0 & \ldots & 0 & 1 & 1 \end{array}\right) \in \mathbb{R}^{(n-1)\times n},$$

$$C_\beta = \left(\begin{array}{ccc|c} & & & 0 \\ & I_{n-2} & & \vdots \\ & & & 0 \\ \hline 0 & \ldots & 0 & \beta \\ 0 & \ldots & 0 & 1{-}\beta \end{array}\right) \in \mathbb{R}^{n\times(n-1)}.$$

As explained below, these matrices can be viewed as projection and lifting operators, mapping between non-aliased and aliased quantities.

Non-aliased part. The non-aliased part of A is a stochastic matrix $\bar{A} \in \mathbb{R}^{(n-1)\times(n-1)}$, obtained by *merging* the two aliased states $n{-}1$ and n into the pseudo-state $\bar{n}$. Its entries are given by

$$\bar{A} = \left(\begin{array}{ccc|c} & & & P(1\,|\,\bar{n}) \\ & A_{[1:n-2]\times[1:n-2]} & & \vdots \\ & & & P(n{-}2\,|\,\bar{n}) \\ \hline P(\bar{n}\,|\,1) & \ldots & P(\bar{n}\,|\,n{-}2) & P(\bar{n}\,|\,\bar{n}) \end{array}\right), \tag{3}$$

where the transition probabilities *into* the pseudo-state are

$$P(\bar{n}\,|\,j) \;=\; P(n{-}1\,|\,j) + P(n\,|\,j), \quad \forall j \in [n],$$

the transition probabilities *out of* the pseudo-state are defined with respect to the stationary distribution by

$$P(i\,|\,\bar{n}) \;=\; \beta P(i\,|\,n{-}1) + (1{-}\beta)P(i\,|\,n), \quad \forall i \in [n]$$

and lastly, the probability to *stay* in the pseudo-state is

$$P(\bar{n}\,|\,\bar{n}) \;=\; \beta P(\bar{n}\,|\,n{-}1) + (1{-}\beta)P(\bar{n}\,|\,n).$$

It is easy to check that the unique stationary distribution of $\bar{A}$ is $\bar{\boldsymbol{\pi}} = (\pi_1, \pi_2, \ldots, \pi_{n-2}, \pi_{\bar{n}}) \in \mathbb{R}^{n-1}$. Finally, note that $\bar{A} = BAC_\beta$, $\bar{\boldsymbol{\pi}} = B\boldsymbol{\pi}$ and $\boldsymbol{\pi} = C_\beta\bar{\boldsymbol{\pi}}$, justifying the lifting and projection interpretation of the matrices B, C_β.

Aliased part. Next we introduce some key quantities that distinguish between the two aliased states. Let $\text{supp}_{\text{in}} = \{j \in [n]\,|\,P(\bar{n}\,|\,j) > 0\}$ be the set of states that can move into at least one of the aliased states. We define

$$\alpha_j \;=\; \begin{cases} \frac{P(n-1\,|\,j)}{P(\bar{n}\,|\,j)} & j \in \text{supp}_{\text{in}} \\ 0 & \text{otherwise,} \end{cases} \tag{4}$$

as the *relative probability* of moving from state j to state $n{-}1$, conditional on moving to either $n{-}1$ or n. We define the two vectors $\boldsymbol{\delta}^{\text{out}}, \boldsymbol{\delta}^{\text{in}} \in \mathbb{R}^{n-1}$ as follows: $\forall i, j \in [n{-}1]$,

$$\delta_i^{\text{out}} \;=\; \begin{cases} P(i\,|\,n{-}1) - P(i\,|\,n) & i < n-1 \\ P(\bar{n}\,|\,n{-}1) - P(\bar{n}\,|\,n) & i = n-1 \end{cases} \tag{5}$$

$$\delta_j^{\text{in}} \;=\; \begin{cases} (\alpha_j - \beta)P(\bar{n}\,|\,j) & j < n{-}1 \\ \beta(\alpha_{n-1} - \beta)P(\bar{n}\,|\,n{-}1) & \\ \quad + (1{-}\beta)(\alpha_n - \beta)P(\bar{n}\,|\,n) & j = n{-}1. \end{cases} \tag{6}$$

In other words, $\boldsymbol{\delta}^{\text{out}}$ captures differences in the transition probabilities *out of* the aliased states. In particular, if $\boldsymbol{\delta}^{\text{out}} = \mathbf{0}$ then starting from either one of the two aliased states, the Markov chain evolution is identical. As proven in Theorem 1 below, such an HMM is not minimal as its two aliased states can be lumped together,

Similarly, $\boldsymbol{\delta}^{\text{in}}$ compares the relative probabilities *into* the aliased states α_j, to the stationary one $\beta = \pi_{n-1}/\pi_{\bar{n}}$. This quantity also plays a role in the minimality of the HMM.

Lastly, for our decomposition, we define the scalar

$$\kappa = (\alpha_{n-1} - \beta)P(\bar{n}\,|\,n{-}1) - (\alpha_n - \beta)P(\bar{n}\,|\,n). \tag{7}$$

Decomposing A. The following lemma provides a decomposition of the transition matrix in terms of $\bar{A}$, $\boldsymbol{\delta}^{\text{out}}$, $\boldsymbol{\delta}^{\text{in}}$, κ and β (all proofs are given in the Appendix).

Lemma 1. *The transition matrix A of a 2A-HMM can be decomposed as*

$$A = C_\beta\bar{A}B + C_\beta\boldsymbol{\delta}^{\text{out}}\boldsymbol{c}_\beta^\intercal + \boldsymbol{b}(\boldsymbol{\delta}^{\text{in}})^\intercal B + \kappa\,\boldsymbol{b}\boldsymbol{c}_\beta^\intercal, \tag{8}$$

where $\boldsymbol{c}_\beta{}^\intercal = (0, \ldots, 0, 1{-}\beta, -\beta) \in \mathbb{R}^n$ *and* $\boldsymbol{b} = (0, \ldots, 0, 1, -1)^\intercal \in \mathbb{R}^n$.

In (8), the first term is the merged transition matrix $\bar{A} \in \mathbb{R}^{(n-1)\times(n-1)}$ lifted back into $\mathbb{R}^{n\times n}$. This term captures all of the non-aliased transitions. The second matrix is zero except in the last two columns, accounting for the exit transition probabilities from the two aliased states. Similarly, the third matrix is zero except in the last two rows, differentiating the entry probabilities. The fourth term is non-zero only on the lower right 2×2 block involving the aliased states $n-1$, n. This term corresponds to the internal dynamics between them. Note that each of the last three terms is at most a rank-1 matrix, which together can be seen as a perturbation due to the presence of aliasing.

In Section 4 we will show the importance of Eq. (8) for the minimality and identifiability of two-aliased HMMs. In section 5 we shall see that given a long output sequence from the HMM, the presence of aliasing can be detected and the quantities $\bar{A}$, $\boldsymbol{\delta}^{\text{out}}$, $\boldsymbol{\delta}^{\text{in}}$, κ and β can all be estimated from it. An estimate for A is then obtained via Eq. (8).

4. Minimality and Identifiability

Two HMMs H and H' are said to be *equivalent* if their observed output sequences are statistically indistinguishable, namely $\mathcal{P}_{H'} = \mathcal{P}_H$. Similarly, an HMM H is *minimal* if there is no equivalent HMM with fewer number of states. Note that if H is non-aliased then Assumptions (A1-A2) readily imply that it is also minimal (Leroux, 1992). In this section we present necessary and sufficient conditions for a 2A-HMM to be minimal, and for two minimal 2A-HMMs to be equivalent. Finally, we derive necessary and sufficient conditions for a minimal 2A-HMM to be identifiable.

4.1. Minimality

The minimality of an HMM is closely related to the notion of *lumpability*: can hidden states be merged without changing the distribution $\mathcal{P}_H$ (Fredkin & Rice, 1986; White et al., 2000; Huang et al., 2014). Obviously, an HMM is minimal iff no subset of hidden states can be merged. The following theorem gives precise conditions for the minimality of a 2A-HMM.

Theorem 1. *Let H be a 2A-HMM satisfying Assumptions (A1-A2) whose initial state X_0 is distributed according to $\boldsymbol{\pi}^0 = (\pi_1^0, \pi_2^0, \ldots, \beta^0\pi_{\bar{n}}^0, (1-\beta^0)\pi_{\bar{n}}^0)$. Then,*

(i) If $\pi_{\bar{n}}^0 \neq 0$ and $\beta^0 \neq \beta$ then H is minimal iff $\boldsymbol{\delta}^{\text{out}} \neq \mathbf{0}$.

(ii) If $\pi_{\bar{n}}^0 = 0$ or $\beta^0 = \beta$ then H is minimal iff both $\boldsymbol{\delta}^{\text{out}} \neq \mathbf{0}$ and $\boldsymbol{\delta}^{\text{in}} \neq \mathbf{0}$.

By Theorem 1, a necessary condition for minimality of a 2A-HMM is that the two aliased states have different exit probabilities, $\boldsymbol{\delta}^{\text{out}} \neq 0$. Namely, there exists a non-aliased state $i \in [n-2]$ such that $P(i \,|\, n-1) \neq P(i \,|\, n)$. Otherwise the two aliased states can be merged. If the 2A-HMM is started from its stationary distribution, then an additional necessary condition is $\boldsymbol{\delta}^{\text{in}} \neq 0$. This last condition implies that there is a non-aliased state $j \in \text{supp}_{\text{in}} \setminus \{n-1, n\}$ with relative entrance probability $\alpha_j \neq \beta$.

4.2. Identifiability

Recall that an HMM H is (strictly) *identifiable* if $\mathcal{P}_H$ uniquely determines the transition matrix A and the output parameters $\boldsymbol{\theta}$, up to a permutation of the hidden states. We establish the conditions for identifiability of a 2A-HMM in two steps. First we derive a novel geometric characterization of the set of all minimal HMMs that are equivalent to H, up to a permutation of the hidden states (Theorem 2). Then we give necessary and sufficient conditions for H to be identifiable, namely for this set to be the singleton set, consisting of only H itself (Appendix C). In the process, we provide a simple procedure (Algorithm 1) to determine whether a given minimal 2A-HMM is identifiable or not.

Equivalence between minimal 2A-HMMs. Necessary and sufficient conditions for the equivalence of two minimal HMMs were studied in several works (Finesso, 1990; Ito et al., 1992; Vanluyten et al., 2008). We now provide analogous conditions for parametric output 2A-HMMs. Toward this end, we define the following 2-dimensional family of matrices $S(\tau_{n-1}, \tau_n) \in \mathbb{R}^{n \times n}$ given by

$$S(\tau_{n-1}, \tau_n) = \left(\begin{array}{ccc|cc} & & & 0 & 0 \\ & I_{n-2} & & \vdots & \vdots \\ & & & 0 & 0 \\ \hline 0 & \ldots & 0 & \tau_{n-1} & \tau_n \\ 0 & \ldots & 0 & 1-\tau_{n-1} & 1-\tau_n \end{array} \right).$$

Clearly, for $\tau_{n-1} \neq \tau_n$, S is invertible. As in (Ito et al., 1992), consider then the following similarity transformation of the transition matrix A,

$$A_H(\tau_{n-1}, \tau_n) = S(\tau_{n-1}, \tau_n)^{-1} A S(\tau_{n-1}, \tau_n). \quad (9)$$

It is easy to verify that $\mathbf{1}_n^\intercal A_H = \mathbf{1}_n^\intercal$. However, A_H is not necessarily stochastic, as depending on τ_{n-1}, τ_n it may have negative entries. The following lemma resolves the equivalence of 2A-HMMs, in terms of this transformation.

Lemma 2. *Let $H = (A, \boldsymbol{\theta}, \boldsymbol{\pi})$ be a minimal 2A-HMM satisfying Assumptions (A1-A2). Then a minimal HMM $H' = (A', \boldsymbol{\theta}', \boldsymbol{\pi}')$ with n' states is equivalent to H iff $n' = n$ and there exists a permutation matrix $\Pi \in \mathbb{R}^{n \times n}$ and $\tau_{n-1} > \tau_n$ such that $\boldsymbol{\theta}' = \Pi\,\boldsymbol{\theta}$ and*

$$\boldsymbol{\pi}' = \Pi\, S(\tau_{n-1}, \tau_n)^{-1} \boldsymbol{\pi}, \;\; A' = \Pi\, A_H(\tau_{n-1}, \tau_n)\, \Pi^{-1} \geq 0.$$

The feasible region. By Lemma 2, any matrix $A_H(\tau_{n-1}, \tau_n)$ whose entries are all non-negative yields an HMM equivalent to the original one. We thus define the *feasible region* of H by

$$\Gamma_H = \{(\tau_{n-1}, \tau_n) \in \mathbb{R}^2 \,|\, A_H(\tau_{n-1}, \tau_n) \geq 0,\, \tau_{n-1} > \tau_n\}. \quad (10)$$

By definition, Γ_H is non-empty, since $(\tau_{n-1}, \tau_n) = (1, 0)$ recovers the original matrix A. As we show below, Γ_H is determined by three simpler regions $\Gamma_1, \Gamma_2, \Gamma_3 \subset \mathbb{R}^2$. The region Γ_1 ensures that all entries of A_H are non-negative except possibly in the lower right 2×2 block corresponding to the two aliased states. The regions Γ_2 and Γ_3 ensure non-negativity of the latter, depending on whether the aliased relative probabilities of (4) satisfy $\alpha_{n-1} \geq \alpha_n$ or $\alpha_{n-1} < \alpha_n$, respectively. For ease of exposition we assume as a convention that $P(\bar{n} \,|\, n-1) \geq P(\bar{n} \,|\, n)$.

Theorem 2. *Let H be a minimal 2A-HMM satisfying Assumptions (A1-A2). There exist $(\tau_{n-1}^{\min}, \tau_n^{\min})$, $(\tau_{n-1}^{\max}, \tau_n^{\max})$, $(\tau^-, \tau^+) \in \mathbb{R}^2$, and convex monotonic decreasing functions $f, g : \mathbb{R} \to \mathbb{R}$ such that*

$$\Gamma_H = \begin{cases} \Gamma_1 \cap \Gamma_2 & \alpha_{n-1} \geq \alpha_n \\ \Gamma_1 \cap \Gamma_3 & \alpha_{n-1} < \alpha_n, \end{cases}$$

where the regions $\Gamma_1, \Gamma_2, \Gamma_3 \subset \mathbb{R}^2$ are given by

$$\begin{aligned} \Gamma_1 &= [\tau_{n-1}^{\min}, \tau_{n-1}^{\max}] \times [\tau_n^{\max}, \tau_n^{\min}] \\ \Gamma_2 &= [\tau^+, \infty) \times [\tau^-, \tau^+] \\ \Gamma_3 &= \{(\tau_{n-1}, \tau_n) \in \Gamma_1 \,|\, f(\tau_{n-1}) \leq \tau_n \leq g(\tau_{n-1})\,\}. \end{aligned}$$

In addition, the set Γ_H is connected.

The feasible regions in the two possible cases ($\alpha_{n-1} \geq \alpha_n$ or $\alpha_{n-1} < \alpha_n$) are illustrated in Appendix C, Fig.4.

Strict Identifiability. By Lemma 2, for strict identifiability of H, Γ_H should be the singleton set $\Gamma_H = \{(1, 0)\}$. Due to lack of space, sufficient and necessary conditions for this to hold, as well as a corresponding simple procedure to determine whether a 2A-HMM is identifiable, are given in Appendix C.2.

Remark. While beyond the scope of this paper, we note that instead of strict identifiability of a given HMM, several works studied a different concept of *generic* identifiability (Allman et al., 2009), proving that under mild conditions the class of HMMs is generically identifiable. In contrast, if we restrict ourselves to the class of 2A-HMMs, then our Theorem 2 implies that this class is generically *non-identifiable*. The reason is that by Theorem 2, for any 2A-HMM whose matrix A has all its entries positive, there are an infinite number of equivalent 2A-HMMs, implying its non-identifiability.

5. Learning a 2A-HMM

Let $(Y_t)_{t=0}^{T-1}$ be an output sequence generated by a parametric-output HMM that satisfies Assumptions (A1-A2) and initialized with its stationary distribution, $X_0 \sim \boldsymbol{\pi}$. We assume the HMM is either non-aliasing, with $n-1$ states, or 2-aliasing with n states. We further assume that the HMM is minimal and identifiable, as otherwise its parameters cannot be uniquely determined.

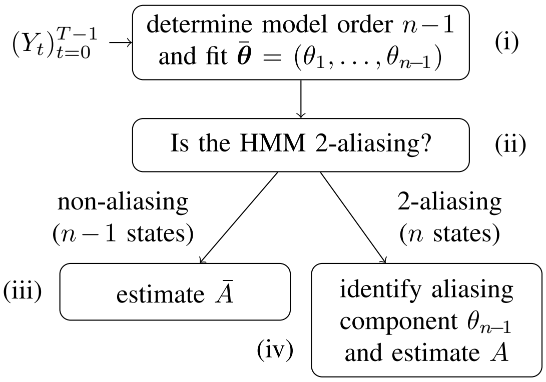
$(Y_t)_{t=0}^{T-1}$
determine model order $n-1$ and fit $\bar{\boldsymbol{\theta}} = (\theta_1, \ldots, \theta_{n-1})$
(i)
Is the HMM 2-aliasing?
(ii)
non-aliasing ($n-1$ states)
2-aliasing (n states)
(iii)
estimate $\bar{A}$
(iv)
identify aliasing component θ_{n-1} and estimate A

Figure 1. Learning a 2A-HMM.

In this section we study the problems of detecting whether the HMM is aliasing and recovering its output parameters $\boldsymbol{\theta}$ and transition matrix A, all in terms of $(Y_t)_{t=0}^{T-1}$. As outlined in Fig.1, the proposed learning procedure consists of the following steps:

(i) Determine the number of output components $n-1$ and estimate the $n-1$ unique output distribution parameters $\bar{\boldsymbol{\theta}}$ and the projected stationary distribution $\bar{\boldsymbol{\pi}}$.

(ii) Detect if the HMM is 2-aliasing.

(iii) In case of a non-aliased HMM, estimate the $(n-1) \times (n-1)$ transition matrix $\bar{A}$, as for example in Kontorovich et al. (2013) or Anandkumar et al. (2012).

(iv) In case of a 2-aliased HMM, identify the component θ_{n-1} corresponding to the two aliased states, and estimate the $n \times n$ transition matrix A.

We now describe in detail each of these steps. As far as we know, our learning procedure is the first to consistently learn a 2A-HMM in a computationally efficient way. In particular, the solutions for problems (ii) and (iv) are new.

Estimating the output distribution parameters. As the HMM is stationary, each observable Y_t is a random realization from the following *parametric mixture model*,

$$Y \sim \sum_{i=1}^{n-1} \bar{\pi}_i f_{\bar{\theta}_i}(y). \tag{11}$$

Hence, the number of unique output components $n-1$, the corresponding output parameters $\bar{\boldsymbol{\theta}}$ and the projected stationary distribution $\bar{\boldsymbol{\pi}}$ can be estimated by fitting a mixture model (11) to the observed output sequence $(Y_t)_{t=0}^{T-1}$.

Consistent methods to determine the number of components in a mixture are well known in the literature (Titterington et al., 1985). Estimating $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\pi}}$ can be done by either the EM algorithm, or any recently developed spectral method (Dasgupta, 1999; Achlioptas & McSherry, 2005; Anandkumar et al., 2012). As our focus is on the aliasing aspects of the HMM, in what follows we assume that the number of unique output components $n-1$, the output parameters $\bar{\boldsymbol{\theta}}$ and the projected stationary distribution $\bar{\boldsymbol{\pi}}$ are *exactly known*. As in Kontorovich et al. (2013), it is possible to show that our method is robust to small errors in these quantities (not presented).

5.1. Moments

To solve problems (ii), (iii) and (iv) above, we first introduce the moment-based quantities we shall make use of. Given $\bar{\boldsymbol{\theta}}$ and $\bar{\boldsymbol{\pi}}$ or estimates of them, for any $i, j \in [n-1]$, we define the *second order moments with time lag t* by

$$\mathcal{M}^{(t)}_{ij} = \mathbf{E}[f_{\theta_i}(Y_0) f_{\theta_j}(Y_t)], \quad t \in \{1,2,3\}. \tag{12}$$

The consecutive in time *third order moments* are defined by

$$\mathcal{G}^{(c)}_{ij} = \mathbf{E}[f_{\theta_i}(Y_0) f_{\theta_c}(Y_1) f_{\theta_j}(Y_2)], \quad \forall c \in [n-1]. \tag{13}$$

We also define the *lifted kernel*, $\mathcal{K} = B^{\mathsf{T}} \bar{K} B \in \mathbb{R}^{n \times n}$. One can easily verify that for a 2A-HMM,

$$\begin{aligned} \mathcal{M}^{(t)} &= \bar{K} B A^t C_\beta \operatorname{diag}(\bar{\boldsymbol{\pi}}) \bar{K} & (14)\\ \mathcal{G}^{(c)} &= \bar{K} B A \operatorname{diag}(\mathcal{K}_{[\cdot,c]}) A C_\beta \operatorname{diag}(\bar{\boldsymbol{\pi}}) \bar{K}. & (15) \end{aligned}$$

Next we define the *kernel free* moments $M^{(t)}, G^{(c)} \in \mathbb{R}^{(n-1)\times(n-1)}$ as follows:

$$\begin{aligned} M^{(t)} &= \bar{K}^{-1} \mathcal{M}^{(t)} \bar{K}^{-1} \operatorname{diag}(\bar{\boldsymbol{\pi}})^{-1} & (16)\\ G^{(c)} &= \bar{K}^{-1} \mathcal{G}^{(c)} \bar{K}^{-1} \operatorname{diag}(\bar{\boldsymbol{\pi}})^{-1}. & (17) \end{aligned}$$

Note that by Assumption (A1), the kernel $\bar{K}$ is full rank and thus $\bar{K}^{-1}$ exists. Similarly, by (A2) $\bar{\boldsymbol{\pi}} > 0$, so $\operatorname{diag}(\bar{\boldsymbol{\pi}})^{-1}$ also exists. Thus, (16,17) are well defined.

Let $R^{(2)}, R^{(3)}, F^{(c)} \in \mathbb{R}^{(n-1)\times(n-1)}$ be given by

$$R^{(2)} = M^{(2)} - (M^{(1)})^2 \tag{18}$$

$$R^{(3)} = M^{(3)} - M^{(2)} M^{(1)} - M^{(1)} M^{(2)} + (M^{(1)})^3 \tag{19}$$

$$F^{(c)} = G^{(c)} - M^{(1)} \operatorname{diag}(\bar{K}_{[\cdot,c]}) M^{(1)}. \tag{20}$$

The following key lemma relates the moments (18, 19, 20) to the decomposition (8) of the transition matrix A.

Lemma 3. *Let H be a minimal 2A-HMM with aliased states $n-1$ and n. Let $\bar{A}$, $\boldsymbol{\delta}^{\text{out}}$, $\boldsymbol{\delta}^{\text{in}}$ and κ be defined in (3,5,6,7) respectively. Then the following relations hold:*

$$\begin{aligned} M^{(1)} &= \bar{A} & (21)\\ R^{(2)} &= \boldsymbol{\delta}^{\text{out}} (\boldsymbol{\delta}^{\text{in}})^{\mathsf{T}} & (22)\\ R^{(3)} &= \kappa R^{(2)} & (23)\\ F^{(c)} &= \bar{K}_{n-1,c} R^{(2)}, \quad \forall c \in [n-1]. & (24) \end{aligned}$$

In the following, these relations will be used to detect aliasing, identify the aliased states and recover the aliased transition matrix A.

Empirical moments. In practice, the unknown moments (12,13) are estimated from the output sequence $(Y_t)_{t=0}^{T-1}$ by

$$\begin{aligned} \hat{\mathcal{M}}^{(t)}_{ij} &= \frac{1}{T-t} \sum_{l=0}^{T-t-1} f_{\theta_i}(Y_l) f_{\theta_j}(Y_{l+t}),\\ \hat{\mathcal{G}}^{(c)}_{ij} &= \frac{1}{T-2} \sum_{l=0}^{T-3} f_{\theta_i}(Y_l) f_{\theta_c}(Y_{l+1}) f_{\theta_j}(Y_{l+2}). \end{aligned}$$

With known (or estimated) $\bar{K}, \bar{\boldsymbol{\pi}}$ the corresponding empirical kernel free moments are given by

$$\begin{aligned} \hat{M}^{(t)} &= \bar{K}^{-1} \hat{\mathcal{M}}^{(t)} \bar{K}^{-1} \operatorname{diag}(\bar{\boldsymbol{\pi}})^{-1} & (25)\\ \hat{G}^{(c)} &= \bar{K}^{-1} \hat{\mathcal{G}}^{(c)} \bar{K}^{-1} \operatorname{diag}(\bar{\boldsymbol{\pi}})^{-1}. & (26) \end{aligned}$$

The empirical estimates for (18,19,20) similarly follow.

To analyze the error between the empirical and population quantities, we make the following additional assumption:

(**A3**) The output distributions are bounded. Namely there exists $L > 0$ such that $\forall i \in [n]$ and $\forall y \in \mathcal{Y}$, $f_{\theta_i}(y) \leq L$.

Lemma 4. *Let $(Y_t)_{t=0}^{T-1}$ be an output sequence generated by an HMM satisfying Assumptions (A1-A3). Then, as $T \to \infty$, for any $t \in \{1,2,3\}$ and $c \in [n-1]$, all error terms $\hat{M}^{(t)} - M^{(t)}$, $\hat{R}^{(t)} - R^{(t)}$ and $\hat{F}^{(c)} - F^{(c)}$ are $O_P(T^{-\frac{1}{2}})$.*

In fact, due to strong mixing, all of the above quantities are asymptotically normally distributed (Bradley, 2005).

5.2. Detection of aliasing

We now proceed to detect if the HMM is aliased (step (ii) in Fig.1). We pose this as a hypothesis testing problem:

$$\begin{aligned} &\mathcal{H}_0 : \ H \text{ is non-aliased with } n-1 \text{ states}\\ &\qquad\qquad\text{vs.}\\ &\mathcal{H}_1 : \ H \text{ is 2-aliased with } n \text{ states.} \end{aligned}$$

We begin with the following simple observation:

Lemma 5. *Let H be a minimal non-aliased HMM with $n-1$ states, satisfying Assumptions (A1-A3). Then $R^{(2)} = 0$.*

In contrast, if H is 2-aliasing then according to (22) we have $R^{(2)} = \boldsymbol{\delta}^{\text{out}} (\boldsymbol{\delta}^{\text{in}})^{\mathsf{T}}$. In addition, since the HMM is assumed to be minimal and started from the stationary distribution, Theorem 1 implies that both $\boldsymbol{\delta}^{\text{out}} \neq 0$ and $\boldsymbol{\delta}^{\text{in}} \neq 0$. Thus $R^{(2)}$ is exactly a rank-1 matrix, which we write as

$$R^{(2)} = \sigma \boldsymbol{u} \boldsymbol{v}^{\mathsf{T}} \quad \text{with} \quad \|\boldsymbol{u}\|_2 = \|\boldsymbol{v}\|_2 = 1, \quad \sigma > 0, \tag{27}$$

where σ is the unique non-zero singular value of $R^{(2)}$. Hence, our hypothesis testing problem takes the form:

$$\mathcal{H}_0 : \ R^{(2)} = 0 \quad \text{vs.} \quad \mathcal{H}_1 : \ R^{(2)} = \sigma \boldsymbol{u} \boldsymbol{v}^{\mathsf{T}} \text{ with } \sigma > 0.$$

In practice, we only have the empirical estimate $\hat{R}^{(2)}$. Even if $\sigma = 0$, this matrix is typically full rank with $n-1$ nonzero singular values. Our problem is thus detecting the rank of a matrix from a noisy version of it. There are multiple methods to do so. In this paper, motivated by Kritchman & Nadler (2009), we adopt the largest singular value $\hat{\sigma}_1$ of $\hat{R}^{(2)}$ as our test statistic. The resulting test is

$$\text{if } \hat{\sigma}_1 \geq h_T \text{ return } \mathcal{H}_1, \text{ otherwise return } \mathcal{H}_0, \qquad (28)$$

where h_T is a predefined threshold. By Lemma 4, as $T \to \infty$ the singular values of $\hat{R}^{(2)}$ converge to those of $R^{(2)}$. Thus, as the following lemma shows, with a suitable threshold this test is asymptotically consistent.

Lemma 6. *Let H be a minimal HMM satisfying Assumptions (A1-A3) which is either non-aliased or 2-aliased. Then for any $0 < \epsilon < \frac{1}{2}$, the test (28) with $h_T = \Omega(T^{-\frac{1}{2}+\epsilon})$ is consistent. Namely, as $T \to \infty$*

$$P(\textit{reject } \mathcal{H}_1 \,|\, \mathcal{H}_1 \textit{ holds}) + P(\textit{reject } \mathcal{H}_0 \,|\, \mathcal{H}_0 \textit{ holds}) \to 0$$

and asymptotically the test correctly detects whether the HMM is non-aliased or 2-aliased.

If the HMM was detected as non-aliasing, then its $(n-1) \times (n-1)$ transition matrix can be estimated, for example, by the spectral methods of Kontorovich et al. (2013) or Anandkumar et al. (2012). We now turn our attention to the case where the HMM was detected as an aliased one.

Estimating the non-aliased transition matrix $\bar{A}$. It is easy to show that in the 2-aliased case, the $(n-1) \times (n-1)$ transition matrix most consistent with the first two moments is nothing but the non-aliased transition matrix $\bar{A}$. Hence, applying for example the spectral method of Kontorovich et al. (2013) yields an estimate $\hat{\bar{A}}$, which is not only strongly consistent, but also satisfies that as $T \to \infty$,

$$\hat{\bar{A}} = \bar{A} + O_P(T^{-\frac{1}{2}}). \qquad (29)$$

5.3. Identifying the aliased component θ_{n-1}

Assuming the HMM was detected as 2-aliasing, our next task, step (iv), is to identify the aliased component. Recall that if the aliased component is θ_{n-1}, then by (24)

$$F^{(c)} \quad = \quad \bar{K}_{n-1,c} R^{(2)}, \quad \forall c \in [n-1].$$

We thus estimate the index $i \in [n-1]$ of the aliased component by solving the following least squares problem:

$$\hat{i} = \operatorname*{argmin}_{i \in [n-1]} \sum_{c \in [n-1]} \left\| \hat{F}^{(c)} - \bar{K}_{i,c} \hat{R}^{(2)} \right\|_{\mathrm{F}}^2. \qquad (30)$$

The following result shows this method is consistent.

Lemma 7. *For a minimal 2A-HMM satisfying Assumptions (A1-A3) with aliased states $n-1$ and n,*

$$\lim_{T \to \infty} \Pr(\, \hat{i} \neq n-1) = 0.$$

5.4. Learning the aliased transition matrix A

Given the aliased component, we estimate the $n \times n$ transition matrix A using the decomposition (8). First, recall that by (22), $R^{(2)} = \boldsymbol{\delta}^{\text{out}} (\boldsymbol{\delta}^{\text{in}})^{\mathsf{T}} = \sigma \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$. As singular vectors are determined only up to scaling, we have that

$$\boldsymbol{\delta}^{\text{out}} = \gamma \boldsymbol{u} \qquad \text{and} \qquad \boldsymbol{\delta}^{\text{in}} = \frac{\sigma}{\gamma} \boldsymbol{v},$$

where $\gamma \in \mathbb{R}$ is a yet undetermined constant. Thus, the decomposition (8) of A takes the form:

$$A = C_\beta \bar{A} B + \gamma C_\beta \boldsymbol{u} \boldsymbol{c}_\beta^{\mathsf{T}} + \frac{\sigma}{\gamma} \boldsymbol{b} \boldsymbol{v}^{\mathsf{T}} B + \kappa\, \boldsymbol{b} \boldsymbol{c}_\beta^{\mathsf{T}}. \qquad (31)$$

Since $\bar{A}, \sigma, \boldsymbol{u}$ and $\boldsymbol{v}$ were estimated in previous steps, we are left to determine the scalars γ, β and κ of Eq. (7).

As for κ, according to (23) we have $R^{(3)} = \kappa R^{(2)}$. Thus, plugging the empirical versions, $\hat{\kappa}$ is estimated by

$$\hat{\kappa} = \operatorname*{argmin}_{r \in \mathbb{R}} \left\| \hat{R}^{(3)} - r \hat{R}^{(2)} \right\|_{\mathrm{F}}^2. \qquad (32)$$

To determine γ and β we turn to the similarity transformation $A_H(\tau_{n-1}, \tau_n)$, given in (9). As shown in Section 3, this transformation characterizes all transition matrices equivalent to A. To relate A_H to the form of the decomposition (31), we reparametrize τ_{n-1} and τ_n as follows:

$$\gamma' = \gamma(\tau_{n-1} - \tau_n), \qquad \beta' = \frac{\beta - \tau_n}{\tau_{n-1} - \tau_n}.$$

Replacing τ_{n-1}, τ_n with γ', β' we find that A_H is given by

$$A_H = C_{\beta'} \bar{A} B + \gamma' C_{\beta'} \boldsymbol{u} \boldsymbol{c}_{\beta'}^{\mathsf{T}} + \frac{\sigma}{\gamma'} \boldsymbol{b} \boldsymbol{v}^{\mathsf{T}} B + \kappa\, \boldsymbol{b} \boldsymbol{c}_{\beta'}^{\mathsf{T}}. \qquad (33)$$

Note that putting $\gamma' = \gamma$ and $\beta' = \beta$ recovers the decomposition (31) for the original transition matrix A.

Now, since H is assumed identifiable, the constraint $A_H(\tau_{n-1}, \tau_n) \geq 0$ has the unique solution $(\tau_{n-1}, \tau_n) = (1, 0)$, or equivalently $(\gamma', \beta') = (\gamma, \beta)$. Thus, with exact knowledge of the various moments, only a single pair of values (γ', β') will yield a non-negative matrix (33). This perfectly recovers γ, β and the original transition matrix A.

In practice we plug into (33) the empirical versions $\hat{\bar{A}}$, $\hat{\kappa}$, $\hat{\sigma}_1$, $\hat{\boldsymbol{u}}_1$ and $\hat{\boldsymbol{v}}_1$, where $\hat{\boldsymbol{u}}_1$, $\hat{\boldsymbol{v}}_1$ are the left and right singular vectors of $\hat{R}^{(2)}$, corresponding to the singular value $\hat{\sigma}_1$. As described in Appendix D.5, the values $(\hat{\gamma}, \hat{\beta})$ are found by maximizing a simple two dimensional smooth function. The resulting estimate for the aliased transition matrix is

$$\hat{A} \quad = \quad C_{\hat{\beta}} \hat{\bar{A}} B + \hat{\gamma} C_{\hat{\beta}} \hat{\boldsymbol{u}}_1 \boldsymbol{c}_{\hat{\beta}}^{\mathsf{T}} + \frac{\hat{\sigma}_1}{\hat{\gamma}} \boldsymbol{b} \hat{\boldsymbol{v}}_1^{\mathsf{T}} B + \hat{\kappa}\, \boldsymbol{b} \boldsymbol{c}_{\hat{\beta}}^{\mathsf{T}}.$$

The following theorem proves our method is consistent.

Theorem 3. *Let H be a 2A-HMM satisfying assumption (A1-A3) with aliased states $n-1$ and n. Then as $T \to \infty$,*

$$\hat{A} = A + o_P(1).$$

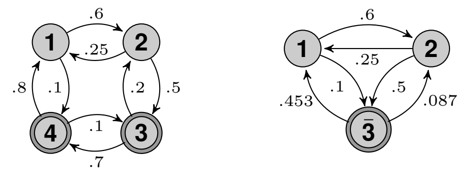

Figure 2. The aliased HMM (left) and its corresponding non-aliased version with states 3 and 4 merged (right).

6. Numerical simulations

The following simulation results illustrate the consistency of our methods to detect aliasing, identify the aliased component and learn the transition matrix A. As our focus is on the aliasing, we assume for simplicity that the output parameters $\bar{\boldsymbol{\theta}}$ and the projected stationary distributions $\bar{\boldsymbol{\pi}}$ are exactly known.

Motivated by ion channel gating (Crouzy & Sigworth, 1990; Rosales et al., 2001; Witkoskie & Cao, 2004), we consider the following HMM H with $n = 4$ hidden states (Fig.2, left). The output distributions are univariate Gaussians $\mathcal{N}(\mu_i, \sigma_i^2)$, the matrix A and $(f_{\theta_i})_{i=1}^4$ are given by

$$A = \begin{pmatrix} 0.3 & 0.25 & 0.0 & 0.8 \\ 0.6 & 0.25 & 0.2 & 0.0 \\ 0.0 & 0.5 & 0.1 & 0.1 \\ 0.1 & 0.0 & 0.7 & 0.1 \end{pmatrix}, \quad \begin{aligned} f_{\theta_1} &= \mathcal{N}(3,1) \\ f_{\theta_2} &= \mathcal{N}(6,1) \\ f_{\theta_3} &= \mathcal{N}(0,1) \\ f_{\theta_4} &= \mathcal{N}(0,1). \end{aligned}$$

States 3 and 4 are aliased and by Procedure 1 in Appendix C.3 this 2A-HMM is identifiable. The rank-1 matrix $R^{(2)}$ has a singular value $\sigma = 0.33$. Fig.2 (right) shows its non-aliased version $\bar{H}$ with states 3 and 4 merged.

To test our aliasing detection algorithm, we generated T outputs from the original aliased HMM and from its non-aliased version $\bar{H}$. Fig.3 (top left) shows the empirical densities (averaged over 1000 independent runs) of the largest singular value of $\hat{R}^{(2)}$, for both H and $\bar{H}$. Fig.3 (top right) shows similar results for a 2A-HMM with $\sigma = 0.22$. When $\sigma = 0.33$, already $T = 1000$ outputs suffice for essentially perfect detection of aliasing. For the smaller $\sigma = 0.22$, more samples are required. Fig.3 (middle left) shows the false alarm and misdetection probabilities vs. sample size T of the aliasing detection test (28) with threshold $h_T = 2T^{-\frac{1}{3}}$. The consistency of our method is evident.

Fig.3 (middle right) shows the probability of misidentifying the aliased component $\theta_{\bar{3}}$. We considered the same 2A-HMM H but with different means for the Gaussian output distribution of the aliased states, $\mu_{\bar{3}} = \{0, 1, 2\}$. As expected, when $f_{\theta_{\bar{3}}}$ is closer to the output distribution of the non-aliased state f_{θ_1} (with mean $\mu_1 = 3$), identifying the aliased component is more difficult.

Finally, we considered the following methods to estimate A: The Baum-Welch algorithm with random initial guess

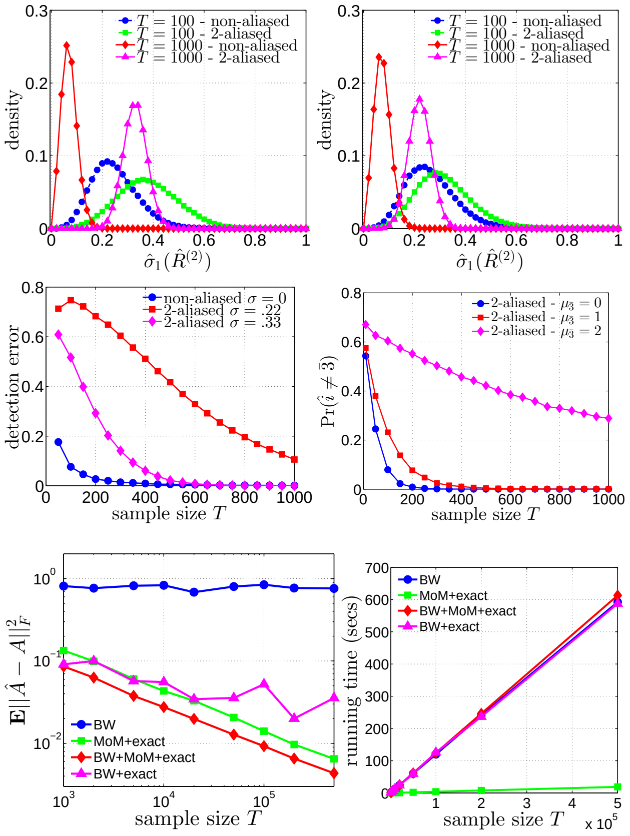

Figure 3. **Top**: Empirical density of the largest singular value of $\hat{R}^{(2)}$ with $\sigma = 0.33$ (left) and $\sigma = 0.22$ (right). **Middle**: Misdetection probability of aliasing/non-aliasing (left) and probability of misidentifying the correct aliased component (right). **Bottom**: Average error $\mathbf{E}\|\hat{A} - A\|_F^2$ and runtime comparison of different algorithms vs. sample size T.

of the HMM parameters (BW); our method of moments with exactly known $\bar{\boldsymbol{\theta}}$ (MoM+Exact); BW initialized with the output of our method (BW+MoM+Exact); and BW with exactly known output distributions but random initial guess of the transition matrix (BW+Exact). Fig.3 (bottom left) shows on a logarithmic scale the mean square error $\mathbf{E}\|\hat{A} - A\|_F^2$ vs. sample size T, averaged over 100 independent realizations. Fig.3 (bottom right) shows the running time as a function of T. In both figures, the number of iterations of the BW was set to 20.

These results show that with a random initial guess of the HMM parameters, BW requires far more than 20 iterations to converge. Even with exact knowledge of the output distributions but a random initial guess of the matrix A, BW still fails to converge after 20 iterations. In contrast, our method yields a relatively accurate estimator in only a fraction of run-time. Furthermore, using this estimator as an initial guess for BW yields even better accuracy.

Acknowledgments

We thank Aryeh Kontorovich for the helpful discussions. This research was supported by the Frankel Center for Computer Science.

References

Achlioptas, Dimitris and McSherry, Frank. On spectral learning of mixtures of distributions. In *Learning Theory*, pp. 458–469. Springer, 2005.

Allman, Elizabeth S, Matias, Catherine, and Rhodes, John A. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, pp. 3099–3132, 2009.

Anandkumar, Animashree, Hsu, Daniel, and Kakade, Sham M. A method of moments for mixture models and hidden Markov models. In *COLT*, 2012.

Baum, L.E., Petrie, T., Soules, G., and Weiss, N. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):pp. 164–171, 1970.

Belkin, M. and Sinha, K. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pp. 103–112, 2010.

Blackwell, David and Koopmans, Lambert. On the identifiability problem for functions of finite Markov chains. *The Annals of Mathematical Statistics*, pp. 1011–1015, 1957.

Bradley, Richard C. Basic properties of strong mixing conditions. a survey and some open questions. *Probab. Surveys*, 2:107–144, 2005.

Brafman, R. I. and Shani, G. Resolving perceptual aliasing in the presence of noisy sensors. In *NIPS*, pp. 1249–1256, 2004.

Brejova, B., Brown, D. G., and Vinar, T. The most probable annotation problem in HMMs and its application to bioinformatics. *Journal of Computer and System Sciences*, 73(7):1060–1077, 2007.

Cappé, Olivier, Moulines, Eric, and Rydén, Tobias. *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York, 2005.

Chrisman, L. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *AAAI*, pp. 183–188. Citeseer, 1992.

Crouzy, Serge C and Sigworth, Frederick J. Yet another approach to the dwell-time omission problem of single-channel analysis. *Biophysical journal*, 58(3):731, 1990.

Dasgupta, Sanjoy. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 634–644, 1999.

Feldman, Jon, O'Donnell, Ryan, and Servedio, Rocco A. Learning mixtures of product distributions over discrete domains. *SIAM Journal on Computing*, 37(5):1536–1564, 2008.

Finesso, Lorenzo. Consistent estimation of the order for Markov and hidden Markov chains. Technical report, DTIC Document, 1990.

Fredkin, Donald R and Rice, John A. On aggregated markov processes. *Journal of Applied Probability*, pp. 208–214, 1986.

Fredkin, Donald R and Rice, John A. Maximum likelihood estimation and identification directly from single-channel recordings. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 249(1325):125–132, 1992.

Hsu, Daniel, Kakade, Sham M., and Zhang, Tong. A spectral algorithm for learning hidden Markov models. In *COLT*, 2009.

Huang, Qingqing, Ge, Rong, Kakade, Sham, and Dahleh, Munther. Minimal realization problems for hidden markov models. *arXiv preprint arXiv:1411.3698*, 2014.

Ito, Hisashi, Amari, S-I, and Kobayashi, Kingo. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *Information Theory, IEEE Transactions on*, 38(2):324–333, 1992.

Jaeger, Herbert. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000.

Jefferies, M. E. and Yeap, W. *Robotics and cognitive approaches to spatial mapping*, volume 38. Springer, 2008.

Kontorovich, A. and Weiss, R. Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *Journal of Applied Probability*, 51:1–14, 2014.

Kontorovich, Aryeh, Nadler, Boaz, and Weiss, Roi. On learning parametric-output HMMs. In *Proceedings of The 30th International Conference on Machine Learning*, pp. 702–710, 2013.

Kritchman, Shira and Nadler, Boaz. Non-parametric detection of the number of signals: Hypothesis testing and random matrix theory. *IEEE Transactions on Signal Processing*, 57(10):3930–3941, 2009.

Leggetter, CJ and Woodland, P. C. Speaker adaptation of continuous density HMMs using multivariate linear regression. In *ICSLP*, volume 94, pp. 451–454, 1994.

Leroux, Brian G. Maximum-likelihood estimation for hidden Markov models. *Stochastic processes and their applications*, 40(1):127–143, 1992.

McCallum, R Andrew. Instance-based utile distinctions for reinforcement learning with hidden state. In *ICML*, pp. 387–395, 1995.

Newey, Whitney K. Uniform convergence in probability and stochastic equicontinuity. *Econometrica*, 59:1161–1167, 1991.

Petrie, T. Probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, pp. 97–115, 1969.

Rosales, Rafael, Stark, J Alex, Fitzgerald, William J, and Hladky, Stephen B. Bayesian restoration of ion channel records using hidden Markov models. *Biophysical journal*, 80(3):1088–1103, 2001.

Shani, G., Brafman, R. I., and Shimony, S. E. Model-based online learning of POMDPs. In *ECML*, pp. 353–364. Springer, 2005.

Siddiqi, Sajid M., Boots, Byron, and Gordon, Geoffrey J. Reduced-rank Hidden Markov Models. In *AISTAT*, 2010.

Stanke, M. and Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(suppl 2):ii215–ii225, 2003.

Stewart, G.W. and Sun, Ji-guang. *Matrix Perturbation Theory*. Academic Press, 1990.

Titterington, D Michael, Smith, Adrian FM, Makov, Udi E, et al. *Statistical analysis of finite mixture distributions*, volume 7. Wiley New York, 1985.

Vanluyten, Bart, Willems, Jan C, and De Moor, Bart. Equivalence of state representations for hidden Markov models. *Systems & Control Letters*, 57(5):410–419, 2008.

White, Langford B, Mahony, Robert, and Brushe, Gary D. Lumpable hidden Markov models-model reduction and reduced complexity filtering. *Automatic Control, IEEE Transactions on*, 45(12):2297–2306, 2000.

Witkoskie, James B and Cao, Jianshu. Single molecule kinetics. i. theoretical analysis of indicators. *The Journal of chemical physics*, 121(13):6361–6372, 2004.

Zatuchna, Z. V. and Bagnall, A. Learning mazes with aliasing states: An LCS algorithm with associative perception. *Adaptive Behavior*, 17(1):28–57, 2009.