Proceedings of the
47th IEEE Conference on Decision and Control
Cancun, Mexico, Dec. 9-11, 2008

Probability of Error Bounds for Failure Diagnosis and Classification in Hidden Markov Models

Eleftheria Athanasopoulou and Christoforos N. Hadjicostis

Abstract— **In this paper we consider a formulation of the failure diagnosis problem in stochastic systems as a maximum likelihood classification problem: a diagnoser observes the system under diagnosis online and determines which candidate model (e.g., a fault-free model or a faulty model) is more likely given the observations. We are interested in measuring *a priori* the diagnosis/classification capability of the diagnoser by computing offline the probability that the diagnoser makes an incorrect decision (irrespective of the actual observation sequence) as a function of the observation step. We focus on hidden Markov models and compute an upper bound on this probability as a function of the length of the sequence observed. We also find necessary and sufficient conditions for this bound to decay to zero exponentially with the number of observations.**

I. Introduction

In this paper our goal is to evaluate the diagnosis (classification) capability of maximum likelihood diagnosis (classification) schemes. Given two candidate hidden Markov models (HMMs) along with their *priors*, we would like to compute offline the *a priori* probability that the diagnoser will make an incorrect decision as a function of the number of observation steps, irrespective of the actual observation sequence. To avoid high computational complexity, we focus on finding an upper bound on this probability of error and obtain necessary and sufficient conditions under which this bound goes to zero exponentially with the number of observations.

Much work has been done in failure diagnosis of discrete event systems, including probabilistic diagnosis or diagnosis of stochastic finite automata [1]–[5]. The work in [3] introduces two notions of stochastic diagnosability (namely A- and AA-diagnosability), both of which refer to asymptotic diagnosis properties as the observation time tends to infinity, and provides conditions that guarantee them. Although our failure model is not the same as the one in [3], the use of the probability of error as a measure of the diagnosis capability is related to the notion of stochastic diagnosability. The biggest difference is that our bound on the probability of diagnosis error is not only concerned with the asymptotic behavior of the system but also gives us information regarding the probability of error as a function of the observation interval.

In our previous work [4], [5] we formulated the failure diagnosis problem in discrete event systems as a maximum likelihood classification problem. (The challenge in [4], [5] was to find ways to deal with erroneous observations rather than the application of the forward algorithm; for simplicity, erroneous observations are *not* considered in this paper.) For HMMs this can be done using a recursive algorithm similar to the forward algorithm, which solves the evaluation problem in hidden Markov models (HMMs) and is used frequently in speech recognition, pattern recognition applications, and bioinformatics (see [6]–[9] and references therein). There are also relations between our work and information theoretic approaches that have been developed to capture the distance between HMMs [7].

In summary, the contribution of this paper in the framework of maximum likelihood diagnosis is two-fold: (i) we propose a measure of diagnosis capability by quantifying the *a priori* probability that the diagnoser makes an incorrect decision, (ii) we calculate a bound on the probability that the diagnoser makes an incorrect decision, and necessary and sufficient conditions for the bound to go to zero exponentially with the number of observation steps.

II. Preliminaries

A. FSM, Markov Chain, and Hidden Markov Model Notation

A finite state machine (FSM) can be described by (Q, X, δ, q_0), where $Q = \{0, 1, 2, ..., |Q|-1\}$ is the set of states; X is the finite set of inputs; δ is the state transition function; and q_0 is the initial state. The FSMs we consider here are event-driven and we use n to denote the time epoch between the occurrence of the n^{th} and $(n+1)^{st}$ input. The state $Q[n+1]$ of the FSM at time epoch $n+1$ is specified by its state $Q[n]$ at time epoch n and its input $X[n+1]$ via the state transition function δ as $Q[n+1] = \delta(Q[n], X[n+1])$.

We denote a time homogeneous Markov chain by $(Q, X, \Delta, \pi[0])$, where $Q = \{0, 1, 2, ..., |Q| - 1\}$ is the set of states; X is the set of inputs; $\pi[0]$ is the initial state probability distribution vector; and Δ captures the state transition probabilities, i.e., $\Delta(q, x_i, q') = P(Q[n+1] = q' \mid Q[n] = q, X[n+1] = x_i)$, for $q, q' \in Q$, $x_i \in X$. If we denote the state transition probabilities by $a_{jk} = P\{(Q[n+1] = j) \mid (Q[n] = k)\}$, the state transition matrix of the Markov chain associated with the given system is $\mathcal{A} = (a_{jk})_{j,k=0,1,...,|Q|-1}$. (To keep the notation clean, the rows and columns of all matrices are indexed starting from

This material is based upon work supported in part by the National Science Foundation under NSF Career Award No 0092696 and NSF ITR Award No 0426831. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of NSF.

E. Athanasopoulou is with the Coordinated Science Laboratory, and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. C. N. Hadjicostis is with the Department of Electrical and Computer Engineering, University of Cyprus, and also with the Coordinated Science Laboratory, and the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign. Corresponding author: C. N. Hadjicostis, 110 Green Park, 75 Kallipoleos Avenue, P.O. Box 20537, 1678 Nicosia, Cyprus. E-mails: {athanaso, chadjic@illinois.edu}.

978-1-4244-3124-3/08/$25.00 ©2008 IEEE

0 and not 1.) The state transition matrix $\mathcal{A}$ captures how state probabilities evolve in time via the evolution equation $\pi[n+1] = \mathcal{A}\pi[n]$. Here, $\pi[n]$ is a $|Q|$-dimensional vector, whose j^{th} entry denotes the probability that the Markov chain is in state j at step n.

An HMM is described by a five-tuple $(Q, Y, \Delta, \Lambda, \boldsymbol{\rho}[0])$, where $Q = \{0, 1, 2, ..., |Q|-1\}$ is the set of states; Y is the set of outputs; Δ captures the state transition probabilities; Λ captures the output probabilities associated with transitions; and $\boldsymbol{\rho}[0]$ is the initial state probability distribution vector. More specifically, for $q, q' \in Q$, $x_i \in X$, and $\sigma \in Y$, the state transition probabilities are given by $\Delta(q, x_i, q') = P(Q[n+1] = q' \mid Q[n] = q, X[n+1] = x_i)$ and the output probabilities associated with transitions are given by $\Lambda(q, \sigma, q') = P(Q[n+1] = q', \lambda(Q[n], X[n+1]) = \sigma \mid Q[n] = q, X[n+1] = x_i)$, where λ denotes the output function that assigns output σ to the transition from state $Q[n]$ under input $X[n+1]$. We define the $|Q| \times |Q|$ matrix $\mathcal{A}_\sigma$, associated with output $\sigma \in Y$ of the HMM, as follows: an entry at the $(j,k)^{th}$ position of $\mathcal{A}_\sigma$ captures the probability of a transition from state k to state j that produces output σ. Note that $\sum_{\sigma \in Y} \mathcal{A}_\sigma = \mathcal{A}$, i.e., a matrix whose $(j,k)^{th}$ entry denotes the probability of taking a transition from state k to state j. The joint probability of the state at step n and the observation sequence $y[1], \ldots, y[n]$ is captured by the vector $\boldsymbol{\rho}[n]$ where the entry $\boldsymbol{\rho}[n](j)$ denotes the probability that the HMM is in state j at step n and the sequence $y_1^n = y[1], \ldots, y[n]$ has been observed. More formally, $\boldsymbol{\rho}[n](j) = P(Q[n] = j, Y_1^n = y_1^n)$ (note that $\boldsymbol{\rho}$ is not necessarily a probability vector).

B. Likelihood Calculation

Given the observation sequence $Y_1^L = y_1^L =< y[1], y[2], ..., y[L] >$, the *priors*, and the initial state probability distributions for two candidate models, the diagnoser implements the maximum *a posteriori* probability (MAP) rule, by comparing $P(S_1 \mid y_1^L) \underset{<}{\overset{>}{}} P(S_2 \mid y_1^L) \Rightarrow \frac{P(y_1^L \mid S_1)}{P(y_1^L \mid S_2)} \underset{<}{\overset{>}{}} \frac{P_2}{P_1}$, and deciding in favor of S_1 (S_2) if the left (right) quantity is larger. For candidate HMM S_i, $i = 1, 2$, we can update $\boldsymbol{\rho}_i$ recursively as $\boldsymbol{\rho}_i[n+1] = \mathcal{A}_{i,y[n+1]}\, \boldsymbol{\rho}_i[n], \quad n = 0, 1, ..., L-1$. If L is the last step, the probability that the observation sequence was produced by FSM S_i is equal to the sum of the entries of $\boldsymbol{\rho}_i[L]$, i.e., $P(y_1^L \mid S_i) = \sum_{j=0}^{|Q|-1} \boldsymbol{\rho}_i[L](j)$. This recursive algorithm is the standard forward algorithm that is used to solve the evaluation problem in HMMs.

III. Probability of Error

We start by conditioning on a given observation sequence and we compute online the conditional probability that the diagnoser makes the incorrect decision as follows:

$$\begin{aligned} P(\text{error at } L \mid y_1^L) &= P(\text{decide } S_2 \text{ at } L,\ S_1 \mid y_1^L) + \\ &\quad P(\text{decide } S_1 \text{ at } L,\ S_2 \mid y_1^L) \\ &= P(\text{decide } S_2 \text{ at } L \mid S_1,\ y_1^L) \cdot P(S_1 \mid y_1^L) + \\ &\quad P(\text{decide } S_1 \text{ at } L \mid S_2,\ y_1^L) \cdot P(S_2 \mid y_1^L) \\ &= \min\{P(S_2 \mid y_1^L),\ P(S_1 \mid y_1^L)\}. \end{aligned}$$

Since both *posteriors* are already computed (for use in the MAP rule comparison), the probability of error given the observation sequence y_1^L as a function of L can be easily computed online along with the maximum likelihood decision. At each step, the diagnoser chooses the model with the larger *posterior* and makes an error with probability equal to the *posterior* of the other model (of course, the *posteriors* are normalized so that they sum up to one).

Our goal is to find a measure of the diagnosis capability of our diagnosis scheme *a priori*, i.e., before any observation is made. The probability of error at step L is given by $P(\text{error at } L) = \sum_{y_1^L} \left(P(y_1^L) \cdot \min\{P(S_2 \mid y_1^L),\ P(S_1 \mid y_1^L)\} \right)$. To perform such computation, we need to find each possible observation sequence y_1^L, along with its probability of occurring, and use it to compute the *posterior* of each model conditioned on this observation sequence. To avoid the possibly prohibitively high computational complexity (especially for large L) we will focus on obtaining an easily computable upper bound and then show that, under certain conditions, this bound on the probability of error decays exponentially to zero with the number of observation steps L.

A diagnoser that uses the MAP rule necessarily chooses model S_1 (S_2) if the observation sequence cannot be produced by S_2 (S_1), with no risk of making an incorrect decision. However, if the observation sequence can be produced by both models, the diagnoser chooses the model with the highest *posterior*, thereby risking to make an incorrect decision. The bound we obtain considers the worst case scenario where, when both models are consistent with the observation sequence y_1^L (i.e., when $P(S_i \mid y_1^L) > 0$ for $i = 1$ and 2), the diagnoser always makes the incorrect decision, and is given by

$$\begin{aligned} &P(\text{error at } L) \\ &= \sum_{y_1^L} \min\{P(S_1 \mid y_1^L), P(S_2 \mid y_1^L)\} \cdot P(y_1^L) \\ &= 1 - \sum_{y_1^L} \max\{P(S_1 \mid y_1^L), P(S_2 \mid y_1^L)\} \cdot P(y_1^L) \\ &= 1 - \sum_{\substack{y_1^L : P(S_i \mid y_1^L) = 0 \\ \text{for } i=1 \text{ or } 2}} P(y_1^L) - \\ &\quad \sum_{\substack{y_1^L : P(S_i \mid y_1^L) > 0 \\ \text{for } i=1 \text{ and } 2}} \max\{P(S_1 \mid y_1^L), P(S_2 \mid y_1^L)\} \cdot P(y_1^L) \\ &\leq 1 - \sum_{\substack{y_1^L : P(S_i \mid y_1^L) = 0 \\ \text{for } i=1 \text{ or } 2}} P(y_1^L) - \frac{1}{2} \sum_{\substack{y_1^L : P(S_i \mid y_1^L) > 0 \\ \text{for } i=1 \text{ and } 2}} P(y_1^L) \\ &= 1 - \sum_{\substack{y_1^L : P(S_i \mid y_1^L) = 0 \\ \text{for } i=1 \text{ or } 2}} P(y_1^L) - \frac{1}{2} \Big(1 - \sum_{\substack{y_1^L : P(S_i \mid y_1^L) = 0 \\ \text{for } i=1 \text{ or } 2}} P(y_1^L) \Big) \\ &= \frac{1}{2} \Big(1 - \sum_{\substack{y_1^L : P(S_i \mid y_1^L) = 0 \\ \text{for } i=1 \text{ or } 2}} P(y_1^L) \Big) \\ &= \frac{1}{2} \Big(1 - P_1 \sum_{\substack{y_1^L : S_2 \\ \text{incons.}}} P(y_1^L \mid S_1) - P_2 \sum_{\substack{y_1^L : S_1 \\ \text{incons.}}} P(y_1^L \mid S_2) \Big). \end{aligned}$$

In the previous formulas we used the fact that, when both

S_1 and S_2 are consistent with the observations, then the maximum of their *posteriors* is greater than half.

Interestingly enough, given that the actual system is S_1, to compute the bound on the probability of error as a function of the observation step, all we need to compute is the probability that the *posterior* of S_2 is equal to zero conditioned on the actual system being S_1 (i.e., the probability that the underlying system is S_1 and it has generated a sequence that is inconsistent with S_2).

IV. Calculation of Bound on Probability of Error

Initially, our objective is to capture the set of observation sequences that are consistent with S_1 but not with S_2, i.e., to capture the set of output sequences that can be produced by S_1 but not by S_2. Once we have identified this set of output sequences, we need to find its probability of occurring. First, we construct the Markov chain $S_{12|1}$ (respectively MC $S_{12|2}$) to help us compute the bound on the probability that S_2 (respectively S_1) becomes inconsistent with the observations given that the actual model is S_1 (respectively S_2). In particular, we explain how to construct MC $S_{12|1}$ starting from HMMs S_1 and S_2 in the following five steps (a similar procedure can be followed to construct MC $S_{12|2}$).

Step 1. Construct FSMs S_{1ND} and S_{2ND} from HMMs S_1 and S_2 respectively.
The set of input sequences that S_{iND} accepts is the set of output sequences that S_i is capable of producing (where $i = 1, 2$). Recall that HMM S_i, is denoted by $(Q_i, Y, \Delta_i, \Lambda_i, \boldsymbol{\rho}_i[0])$ (without loss of generality[1] we assume that $Y_1 = Y_2 = Y$). Ignoring the transition probabilities of HMM S_i, we build the possibly nondeterministic FSM S_{iND} which has the same set of states as S_i and its set of inputs is equal to the set of outputs of S_i. The state transition functionality of S_{iND} is determined by the output functionality of S_i which is captured by Λ_i (although the probabilities are not important at this point). More formally, FSM S_{iND} is denoted by $S_{iND} = (Q_{iND}, X_{iND}, \delta_{iND}, q_{iND0})$, where $Q_{iND} = Q_i$; $X_{iND} = Y$; $q_{iND0} = \{j \text{ s.t. } \boldsymbol{\rho}_i[0](j) > 0\}$ (i.e., q_{iND0} includes all states of S_i with nonzero initial probability); and $\delta_{iND}(q_{iND}, \sigma) = q'_{iND}$ if $\Lambda_i(q_{iND}, \sigma, q'_{iND}) \neq 0$.

Step 2. Construct FSMs S_{1D} and S_{2D} from FSMs S_{1ND} and S_{2ND} respectively.
We can think of FSM S_{iD} as an observer for S_i because each state of S_{iD} contains the set of states that S_i may be in given the observation sequence. The number of states of S_{iD}, i.e., the deterministic version of S_{iND} could be as many as $2^{|Q_{iND}|}$. Although this may raise complexity issues, it is very common in practical scenarios for S_{iD} to have roughly the same number of states as S_{iND} [10]. Following the procedure of subset construction [10] we use S_{iND} to build the deterministic, equivalent machine $S_{iD} = (Q_{iD}, X_{iD}, \delta_{iD}, q_{iD0})$, where Q_{iD} contains subsets of states in the set Q_i (recall that $Q_{iND} = Q_i$); the set of inputs are the same as the set of inputs of S_{iND}, i.e., $X_{iD} = Y$ (recall that $X_i = Y$); $q_{iD0} = q_{i0}$; and δ_{iD} is determined from S_{iND} by the procedure of subset construction, i.e., for $Q_S \subset Q_i$ and $\sigma \in Y$, $\delta_{iD}(Q_S, \sigma) = \bigcup_k$ s.t. $_{\delta_{iND}(j,\sigma)=k \text{ for } j \in Q_S} k$.

Step 3. Construct FSM S_{2DNC} from FSM S_{2ND}.
Next, we append the inconsistent state NC to S_{2D} to obtain FSM S_{2DNC}. As mentioned earlier, FSM S_{2D} accepts all sequences that can be produced by S_2. FSM S_{2DNC} accepts not only the sequences that can be produced by S_2, but also all other sequences (that cannot be produced by S_2). In fact, all sequences that cannot be produced by S_2 will lead S_{2DNC} to its inconsistent state NC. More specifically, $S_{2DNC} = (Q_{2DNC}, X_{2DNC}, \delta_{2DNC}, q_{2DNC0})$, where $Q_{2DNC} = Q_{2D} \cup \{NC\}$; $X_{2DNC} = Y$; $q_{2DNC0} = q_{2D0}$ and δ_{2DNC} is given by $\delta_{2DNC}(q_{2D}, \sigma) =$

$$\begin{cases} \delta_{2D}(q_{2DNC}, \sigma), & \text{if } q_{2DNC} \neq NC, \delta_{2D}(q_{2DNC}, \sigma) \neq \emptyset, \\ NC, & \text{otherwise.} \end{cases}$$

Step 4. Construct FSM S_{1D2DNC} from FSMs S_{1D} and S_{2DNC}.
To capture the set of observations that can be produced by S_1 but not by S_2, we need to build the product FSM S_{1D2DNC}. FSM S_{1D2DNC} accepts all sequences that can be produced by S_1; from all of these sequences, the ones that cannot be produced by S_2 lead S_{1D2DNC} to a state of the form $\{q_{1D}, NC\}$. More specifically, $S_{1D2DNC} = S_{1D} \times S_{2DNC}$, i.e., $S_{1D2DNC} = (Q_{1D2DNC}, X_{1D2DNC}, \delta_{1D2DNC}, q_{0,1D2DNC})$, where $Q_{1D2DNC} = Q_{1D} \times Q_{2DNC}$; $X_{1D2DNC} = Y$ (recall that $X_{1D} = X_{2DNC} = Y$), $q_{0,1D2DNC} = q_{0,1D} \times q_{0,2DNC}$; and δ_{1D2DNC} is given by $\delta_{1D2DNC}(\{q_{1D}, q_{2DNC}\}, \sigma) = \{\delta_{1D}(q_{1D}, \sigma), \delta_{2DNC}(q_{2DNC}, \sigma)\}$, $\sigma \in Y$. Note that $\delta_{1D2DNC}(\{q_{1D}, q_{2DNC}\}, \sigma)$ is undefined if $\delta_{1D}(q_{1D}, \sigma)$ is undefined.

Step 5. Construct MC $S_{12|1}$ from FSM S_{1D2DNC} or from S_1, S_{1D}, and S_{2D}.
To compute the probabilities of the sequences captured by S_{1D2DNC} we construct the Markov chain $S_{12|1} = (Q_{12|1}\ X_{12|1}, \Delta_{12|1}, \boldsymbol{\rho}_{12|1}[0])$, where $Q_{12|1} = Q_1 \times Q_{1D2DNC}$; $X_{12|1} = Y$; $\boldsymbol{\rho}_{12|1}[0](\{q_1, q_{0,1D2DNC}\}) = \boldsymbol{\rho}_1[0](q_1)$, for every $q_1 \in Q_1$ and zero otherwise;[2] and $\Delta_{12|1}$ is given by $\Delta_{12|1}$ $(\{q_1, q_{1D}, q_{2DNC}\}, \sigma, \{q'_1, \delta_{1D}(q_{1D}, \sigma), \delta_{2DNC}(q_{2DNC}, \sigma)\}) = \Lambda_1(q_1, \sigma, q'_1)$ for all $\sigma \in Y$ s. t. $\delta_{1D}(q_{1D}, \sigma) \neq \emptyset$. We group all states of the form $\{q_1, q_{1D}, NC\}$ in one new state and call it NC; we also add a self-loop at state NC with probability one.

Alternatively we can build MC $S_{12|1}$ from S_1, S_{1D}, and S_{2D} as follows: $S_{12|1} = (Q_{12|1}, X_{12|1}, \Delta_{12|1}, \boldsymbol{\rho}_{12|1}[0])$, where $Q_{12|1} = Q_1 \times Q_{1D} \times Q_{2D}$; $X_{12|1} = Y$; $\boldsymbol{\rho}_{12|1}[0](\{q_1, q_{1D}, q_{2D}\}) = \boldsymbol{\rho}_1[0](q_1)$, for every $q_1 \in Q_1$, $q_{1D} \in Q_{1D}$, and $q_{2D} \in Q_{2D}$; and $\Delta_{12|1}$ is given by $\Delta_{12|1}(\{q_1, q_{1D}, q_{2DNC}\}, \sigma, \{q'_1, \delta_{1D}(q_{1D}, \sigma), \delta_{2DNC}(q_{2DNC}, \sigma)\}) = \Lambda_1(q_1, \sigma, q'_1)$, for all $\sigma \in$

[1] We can always redefine $Y = Y_1 \cup Y_2$ to be the output of both machines if Y_1 and Y_2 are different.

[2] Abusing notation, we use $\boldsymbol{\rho}_{12|1}[0](\{q_1, q_{1D2DNC}\})$ to denote the entry of $\boldsymbol{\rho}_{12|1}[0]$ that corresponds to state $\{q_1, q_{1D2DNC}\}$; of course, $\boldsymbol{\rho}_1[0](q_1)$ denotes the entry of $\boldsymbol{\rho}_1[0]$ that corresponds to state q_1.

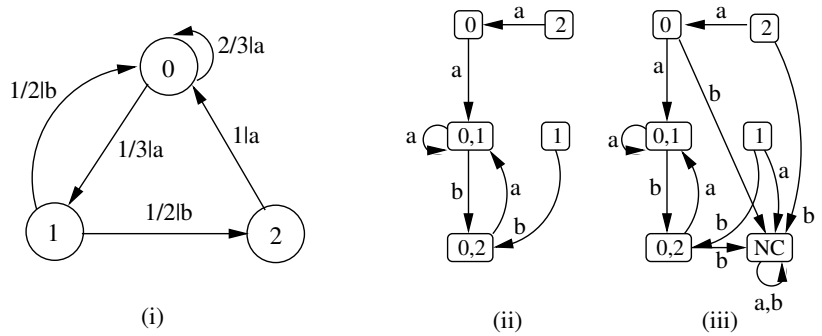

Fig. 1. State transition diagrams of (i) HMM S_1, (ii) FSM S_{1D}, and (iii) FSM S_{1DNC} of Example 1.

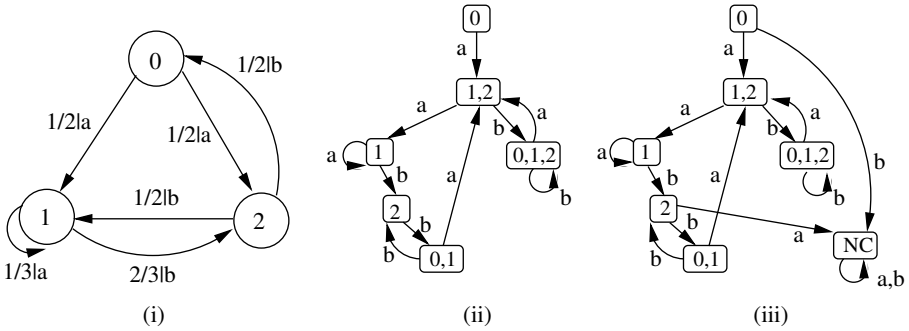

Fig. 2. State transition diagrams of (i) HMM S_2, (ii) FSM S_{2D}, and (iii) FSM S_{2DNC} of Example 1.

Y s. t. $\delta_{1D}(q_{1D},\sigma) \neq \emptyset$ and $\delta_{2D}(q_{2D},\sigma) \neq \emptyset$ or $\Delta_{12|1}\ (\{q_1,q_{1D},q_{2DNC}\},\sigma,\{q_1',\delta_{1D}(q_{1D},\sigma),NC\}) = \Lambda_1(q_1,\sigma,q_1')$, for all $\sigma \in Y$ s. t. $\delta_{1D}(q_{1D},\sigma) \neq \emptyset$ and $\delta_{2D}(q_{2D},\sigma) = \emptyset$. As mentioned before, we group all states of the form $\{q_1,q_{1D},NC\}$ in one new state and call it NC; then we add a self-loop at state NC with probability one.

Notice that any path in $S_{12|1}$ that ends up in state NC represents a sequence that can be produced by S_1 but not by S_2; the probability of such path is easily computed using the Markovian property. Recall that our objective is to calculate the probability that HMM S_2 is inconsistent with the observations given that the observations are produced by S_1 (i.e., $\sum_{\substack{y_1^L:S_2\\ \text{incons.}}} P(y_1^L \mid S_1)$). Therefore, we are interested in the probability of $S_{12|1}$ being in the inconsistent state NC as a function of the observation step given by $P(S_{12|1}$ in state NC at $L) = \pi_{12|1}[L](NC)$, where $\pi_{12|1}[L](NC)$ denotes the entry of $\boldsymbol{\pi}_{12|1}[L]$ that captures the probability that $S_{12|1}$ is in the inconsistent state NC at L. Note that $\pi_{12|1}[L] = \mathcal{A}_{12|1}^L \boldsymbol{\pi}_{12|1}[0]$, where $\mathcal{A}_{12|1}$ is the matrix that captures the transition probabilities for MC $S_{12|1}$ and $\boldsymbol{\pi}_{12|1}[0] = \boldsymbol{\rho}_{12|1}[0]$.

Proposition 1: The probability of error as a function of the observation step is given by

$$P(\text{error at } L) \leq$$
$$\tfrac{1}{2}\Big(1 - P_1 \cdot \sum_{\substack{y_1^L:S_2\\ \text{incons.}}} P(y_1^L \mid S_1) - P_2 \cdot \sum_{\substack{y_1^L:S_1\\ \text{incons.}}} P(y_1^L \mid S_2)\Big)$$
$$= \tfrac{1}{2} - \tfrac{1}{2}P_1 \cdot \pi_{12|1}[L](NC) - \tfrac{1}{2}P_2 \cdot \boldsymbol{\pi}_{12|2}[L](NC),$$

where $\pi_{12|1}[L](NC)$ captures the probability of $S_{12|1}$ being in state NC at step L, $\boldsymbol{\pi}_{12|2}[L](NC)$ captures the probability of $S_{12|2}$ being in state NC at step L, and P_1 and P_2 denote the *priors* of S_1 and S_2. □

Example 1: We consider two candidate HMMs S_1 and S_2 with $Q_1 = Q_2 = \{0,1,2\}$, $Y_1 = Y_2 = \{a,b\}$, initial state $\{0\}$, and transition functionality, as shown in Figures 1.(i) and 2.(i), where each transition is labeled by $p_i \mid \sigma$, i.e., the probability of the transition and the output it produces. Following the procedure of subset construction we construct the deterministic FSMs S_{1D} and S_{2D} as shown in Figures 1.(ii) and 2.(ii), respectively (notice that we include states $\{1\}$ and $\{2\}$ in the state transition diagram of S_{1D} for completeness although they are not reachable from the initial state $\{0\}$). Adding the inconsistent state for each machine we get FSMs S_{1DNC} and S_{2DNC} as shown in Figures 1.(iii) and 2.(iii), respectively. Then, we construct MCs $S_{12|1}$ and $S_{12|2}$ with state transition diagrams as shown in Figures 3 and 4, respectively. For example, the sequence $< a\ a\ b\ a >$ can be produced by S_1 but not by S_2 (hence, given that this is the observation sequence, the probability that the diagnoser makes an incorrect decision is zero). In fact, all sequences in $S_{12|1}$ that end up in state NC can be produced by S_1 but not by S_2. □

V. Properties of Bound on Probability of Error

The inconsistent state NC in MC $S_{12|1}$ is an absorbing state by construction. Therefore, the probability that $S_{12|1}$ is in state NC does not decrease as a function of the observation step; the same property holds for $S_{12|2}$. From Proposition 1 it is clear that the bound on the probability of error is a nonincreasing function of the observation step.

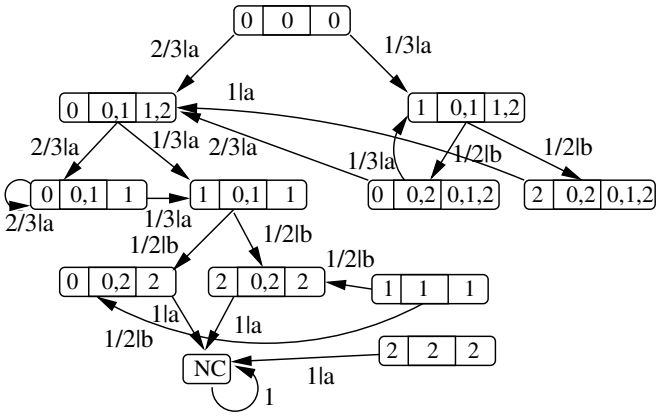

Fig. 3. State transition diagram of MC $S_{12|1}$ of Example 1.

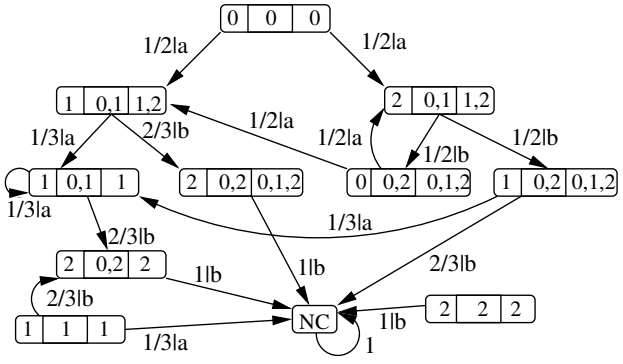

Fig. 4. State transition diagram of MC $S_{12|2}$ of Example 1.

Proposition 2: The bound on the probability of error given by Proposition 1 is a nonincreasing function of the number of observation steps. □

In fact, if MCs $S_{12|1}$ and $S_{12|2}$ have a unique absorbing state each, i.e., state NC, then the bound goes to zero as the number of observation steps increases. The expected number of steps to absorption, given that the initial state is the 0^{th} state of $S_{12|1}$, can be calculated using the fundamental matrix of the absorbing Markov chain $S_{12|1}$ [11]. If $\mathcal{A}_{T12|1}$ is the substochastic transition matrix of $S_{12|1}$ that captures the transitions among all transient states (all but NC) then the fundamental matrix is given by $\sum_{i=0}^{\infty} {\mathcal{A}_{T12|1}}^i = (\mathbf{I} - \mathcal{A}_{T12|1})^{-1}$ and its $(j,k)^{th}$ entry captures the expected number of transitions from state k to state j before absorption. The expected number of steps to absorption, given that the initial state is state $\{0\}$, is equal to the sum of the elements of the 0th column of the fundamental matrix. In fact, the rate of convergence to absorption depends on the largest eigenvalue of the substochastic matrix $\mathcal{A}_{T12|1}$ (because the rate of convergence of matrix $\mathcal{A}^m_{T12|1}$ is captured by the rate of convergence of $\lambda^m_{12|1}$, where $\lambda_{12|1}$ is the largest eigenvalue of $\mathcal{A}_{T12|1}$ and m denotes the number of steps [11]).

Let us now consider the scenario where neither $S_{12|1}$ nor $S_{12|2}$ includes the inconsistent state NC in their set of states. Then the bound on the probability of error will not go to zero; in fact, it will always be equal to half, thereby providing us with no useful information. This scenario corresponds to the case where all output sequences that can be produced by S_1 can also be produced by S_2 and vice versa. For this to be true, S_1 and S_2 need to be equivalent, i.e., generate the same regular language (i.e., the same set of output sequences). Of course, although the set of output sequences is the same for both models, the probabilities associated with an output sequence could be different for each model. The *posteriors* of the candidate models in this case would be strictly greater than zero for any observation sequence; hence the error in the MAP decision will always be nonzero. We can check whether S_1 and S_2 are equivalent using standard approaches with complexity $O((|Q_{1D}| + |Q_{2D}|)^2)$ [10]. We can also easily check equivalence by using S_{1D2DNC} and S_{1DNC2D} which we have already constructed: if the inconsistent state in either S_{1D2DNC} or S_{1DNC2D} (and consequently $S_{12|1}$ or $S_{12|2}$) can be reached starting from the initial state, then the two models are not equivalent.

If MC $S_{12|1}$ has no absorbing state and MC $S_{12|2}$ has only the state NC as an absorbing state, then the bound on the probability of error goes to the value $\frac{P_1}{2}$. This case corresponds to the language generated by S_1 being a subset of the language generated by S_2, i.e., the set of output sequences that can be produced by S_1 can also be produced by S_2. To check for this scenario, we can check whether the inconsistent state in S_{1D2DNC} is reachable from the initial state. We formalize the above discussion in the following proposition.

Proposition 3: For two HMMs S_1 and S_2, the upper bound on the probability of error for the diagnosis decision

- tends to zero exponentially with the number of observation steps, if (and only if) each of FSMs S_{1D2DNC} and S_{1DNC2D} has a unique absorbing state, namely the inconsistent state;
- tends to the value $P_1/2$ exponentially with the number of observation steps, if FSM S_{1D2DNC} has no inconsistent state and FSM S_{1DNC2D} has a unique absorbing state, i.e., the inconsistent state;
- tends to the value $P_2/2$ exponentially with the number of observation steps, if FSM S_{1D2DNC} has no inconsistent state and FSM S_{1DNC2D} has a unique absorbing state, i.e., the inconsistent state;
- is equal to $1/2$, if (and only if) FSMs S_{1D2DNC} and S_{1DNC2D} have no inconsistent states. □

Example 1 (continued): As shown in Figures 3 and 4, each of $S_{12|1}$ and $S_{12|2}$ have NC as the unique absorbing state. Assuming equal *priors* $P_1 = P_2 = 0.5$, the bound on the probability of error can be calculated; as expected, it goes to zero exponentially as the observation time increases (see Figure 5). After running simulations, half with the actual model being S_1 and the other half with the actual model being S_2, we obtain the empirical probability of error given S_1 (and given S_2) by recording the fraction of simulations for which the diagnoser incorrectly decided S_2 (and S_1, respectively). The empirical probability of error as a function of the observation step is shown in Figure 5. The expected

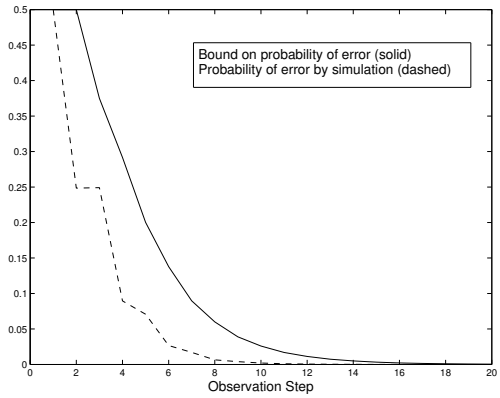

Fig. 5. Plot of the bound on the probability of error (solid) and the empirical probability of error obtained via simulations (dashed) in Example 1, as functions of the observation step.

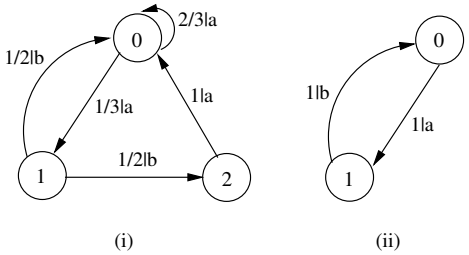

Fig. 6. State transition diagram of (i) HMM S_1 and (ii) HMM S_3 of Example 2.

time to absorption for $S_{12|1}$ is calculated to be 6.8 steps and the expected time to absorption for $S_{12|2}$ is 6.3 steps; hence, for equal *priors*, the mean number of steps for the probability of error to become zero is 6.55 steps. □

Example 2: We consider HMM S_1 of Example 1 and HMM S_3 as shown in Figure 6, and we assume equal *priors*. Notice that any output sequence that can be produced by S_3 can also be produced by S_1; thus, there is no inconsistent state in $S_{13|3}$ and the probability $\sum_{y_1^L:P(S_1|y_1^L)=0} P(y_1^L \mid S_3)$ is always equal to zero. On the other hand, $S_{13|3}$ has a unique absorbing inconsistent state. According to the proposition, we expect the bound on the probability of error to go to $P_1/2 = 0.25$. ¿From Figure 7 we see that, although the bound on the probability of error goes to 0.25, the simulations show that the empirical probability of error goes to zero as the number of steps increases; for this set of candidate models, the bound is not tight, even as the number of observation steps goes to infinity. □

VI. Conclusions

In this paper we consider a formulation of failure diagnosis as a maximum likelihood classification problem. Given candidate HMMs along with their *priors*, a diagnoser determines which candidate model has most likely produced the observation sequence of the system under diagnosis. We are interested in the *a priori* probability that the diagnoser makes an incorrect decision as a function of the observation step. Since the complexity for calculating the exact probability of error can be prohibitively high, we find an upper bound on this probability, as well as necessary and sufficient conditions for the bound to go exponentially to zero as the number of observation steps increases. These bounds can be used to bound the similarity/dissimilarity between HMMs and can therefore find applications in many areas where HMM classification is used, including pattern recognition applications and bioinformatics.

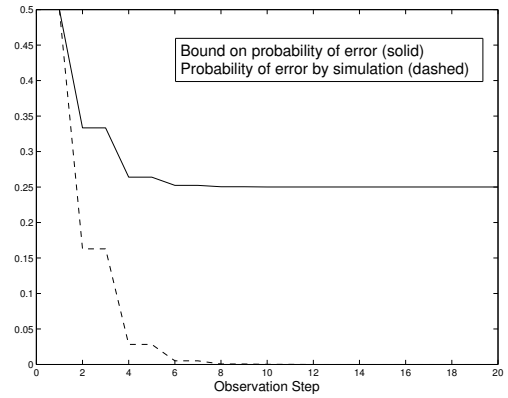

Fig. 7. Plot of the bound on the probability of error (solid) and the empirical probability of error obtained by simulation (dashed) in Example 2, as functions of the observation step.

References

[1] C. N. Hadjicostis, "Probabilistic fault detection in finite-state machines based on state occupancy measurements," *IEEE Trans. on Automatic Control,* vol. 50, no. 12, pp. 2078–2083, December 2005.

[2] M. Blanke, M. Kinnaert, J. Lunze, M. Staroswiecki, *Diagnosis and Fault-Tolerant Control.* Springer-Verlag, 2003.

[3] D. Thorsley and D. Teneketzis, "Diagnosability of stochastic discrete-event systems," *IEEE Trans. on Automatic Control,* vol. 50, no. 4, pp. 476–492, April 2005.

[4] E. Athanasopoulou, L. Li, and C. N. Hadjicostis, "Probabilistic failure diagnosis in finite state machines under unreliable observations," in *Proc. of WODES 2006, the* 8^{th} *International Workshop on Discrete Event Systems,* July 2006, pp. 301–306.

[5] E. Athanasopoulou, L. Li, and C. N. Hadjicostis, "Online posterior probability calculation for failure diagnosis in finite state machines based on unreliable sensor information," in *Proc. of DX-07, the* 18^{th} *International Workshop on Principles of Diagnosis,* May 2007, pp. 13–20.

[6] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE,* vol. 77, no. 2, pp. 7–13, February 1989.

[7] Y. Ephraim and N. Merhav, "Hidden Markov processes," *IEEE Trans. Information Theory,* vol. 48, no. 6, pp. 1518–1569, June 2002.

[8] A. M. Poritz, "Hidden Markov models: A guided tour," in *Proc. IEEE Conf. Acoustics, Speech, and Signal Processing,* vol. 1, April 1988, pp. 7–13.

[9] E. Vidal, F. Thollard, C. de la Higuera, F. Casacuberta and R. C. Carrasco, "Probabilistic finite-state machines—part I," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 7, pp. 1013–1025, July 2005.

[10] J. E. Hopcroft, R. Motwani, and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation.* Cambridge, MA: Addison Wesley, 2001.

[11] J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains.* 2^{nd} ed., New York: Springer-Verlag, 1976.