J. Appl. Prob. **42,** 632–642 (2005)
Printed in Israel

SENSITIVITY OF HIDDEN MARKOV MODELS

ALEXANDER YU. MITROPHANOV,*
ALEXANDRE LOMSADZE * AND
MARK BORODOVSKY,** *Georgia Institute of Technology*

Abstract

We derive a tight perturbation bound for hidden Markov models. Using this bound, we show that, in many cases, the distribution of a hidden Markov model is considerably more sensitive to perturbations in the emission probabilities than to perturbations in the transition probability matrix and the initial distribution of the underlying Markov chain. Our approach can also be used to assess the sensitivity of other stochastic models, such as mixture processes and semi-Markov processes.

Keywords: Hidden Markov model; Markov chain; mixture; semi-Markov process; perturbation bound; sensitivity analysis

2000 Mathematics Subject Classification: Primary 93B35
Secondary 62M09; 60J10; 60E05

1. Introduction

Hidden Markov models (HMMs) and generalizations thereof are widely used in many fields of science and engineering (see, e.g. [15], [24], [28]–[29]). One of the most important problems in HMM theory is that of parameter estimation. The relatively long history of this problem started with the articles [3] and [25], which demonstrated the consistency and asymptotic normality of the maximum likelihood (ML) estimator for some classes of finite HMM. Later, these results were extended to other classes of HMM (see [7], [12], [16]). Another direction of research in the area of parameter estimation for HMMs is the development of computationally efficient methods of maximizing the likelihood function. One of the most popular methods is the Baum–Welch algorithm, which in fact is the expectation–maximization algorithm for HMMs (see [1], [4], [28]). Non-ML estimation procedures for HMMs have also been considered (e.g. the Bayesian maximum a-posteriori estimation described in [15]).

The accuracy of the estimates generated by the above-listed or other methods is sometimes known and, to some extent, can be controlled (e.g. by setting the appropriate tolerance for the iterative Baum–Welch procedure). To decide what accuracy is sufficient, it is desirable to know the effects of small perturbations in the parameter values on the distribution of the HMM under study. Here we develop an inequality-based approach to sensitivity analysis for HMMs. Numerous results have been obtained in the inequality-based perturbation theory for Markov chains, both in discrete and continuous time (see [8], [11], [14], [19]–[23], [33], and references therein). We show that these results play an important role in the perturbation theory for HMMs.

Received 19 January 2005; revision received 1 June 2005.
* Postal address: School of Biology, Georgia Institute of Technology, 310 Ferst Drive, Atlanta, GA 30332-0230, USA.
** Postal address: School of Biomedical Engineering, Georgia Institute of Technology, 313 Ferst Drive, Atlanta, GA 30332-0535, USA. Email address: mark@amber.biology.gatech.edu

The current study was motivated by the problem of parameter estimation for hidden Markov models in bioinformatics. While such models have proved very useful in the area of gene finding, the complexity of genomes and the finiteness of available training data make parametrization of such models a nontrivial task (see [5], [6], [17]). This problem becomes especially hard when (almost) no prior information is available about the genes in a genome, and unsupervised training methods for HMMs should be employed (see [5] and [6]). One of the important questions is, what parameters have more influence on the HMM's behavior, and thus should be treated with additional care?

In the context of speech recognition, it has been noticed that the behavior of finite HMMs is usually much more sensitive to changes in the emission probabilities than to changes in the transition probability matrix and the initial distribution of the underlying Markov chain [13]. Not long ago, the same observation was reported by researchers in bioinformatics [24]. To the best of the authors' knowledge, no theoretical explanation for this phenomenon has been obtained; the perturbation bound that we derive allows to prove and quantify it (see Sections 3 and 4).

The nature of our results is quite general, and our approach can be used outside the HMM setting. In particular, it is well suited for investigating the sensitivity of state space models, mixture processes, and semi-Markov processes (see Section 5).

2. Preliminaries

In this section, we define a hidden Markov model and introduce the necessary notation. For a background to HMMs, see the monograph [18] and the survey paper [10]. Define $\mathcal{S} = \{1, \dots, N\}$, $N \geq 2$, and let $\{X_n\}$, $n \in \mathbb{Z}_+ := \{0, 1, 2, \dots\}$, be a discrete-time, homogeneous Markov chain with state space $\mathcal{S}$. Let $\{Y_n\}$, $n \in \mathbb{Z}_+$, be a sequence of random variables which take values in some measurable space $(\mathcal{A}, \mathfrak{A})$ and satisfy the following conditions:

(a) the variables Y_n are conditionally independent given the sequence $\{X_n\}$;

(b) for each $l \in \mathbb{Z}_+$, Y_l depends on $\{X_n\}$ only through X_l.

The pair $\{(X_n, Y_n)\}$ is called a hidden Markov model. This term comes from the fact that, in applied settings, the variables X_n are usually unobserved, while the Y_n are observable. The sequence $\{X_n\}$ is sometimes called the *regime*.

Let $\boldsymbol{P}$ be the transition probability matrix of $\{X_n\}$. Let $\{f^{(i)}\}$, $i \in \mathcal{S}$, be a family of probability measures on $\mathfrak{A}$ such that $f^{(i)}$ is the distribution of Y_n under the condition $X_n = i$. It is clear that $\boldsymbol{P}$, $\{f^{(i)}\}$, and the initial distribution of $\{X_n\}$ completely define the HMM $\{(X_n, Y_n)\}$. Let p_n and d_n be the distributions of X_n and Y_n, respectively. We regard $\{(X_n, Y_n)\}$ as the unperturbed HMM and consider some perturbed HMM $\{(\tilde{X}_n, \tilde{Y}_n)\}$ with corresponding characteristics denoted by $\tilde{\boldsymbol{P}}$, $\{\tilde{f}^{(i)}\}$, $\tilde{d}_n$, and $\tilde{p}_n$ (where $\tilde{X}_n$ takes values in $\mathcal{S}$ and $\tilde{Y}_n$ takes values in $\mathcal{A}$). Our major goal is to obtain a perturbation bound for d_n in terms of the size of the perturbation in p_n and $\{f^{(i)}\}$.

To measure the closeness between probability distributions, we use the variation distance. For a measurable space $(\mathcal{X}, \mathfrak{X})$, let $\mathcal{M}(\mathfrak{X})$ be the class of finite signed measures on $\mathfrak{X}$. In our case, $(\mathcal{X}, \mathfrak{X})$ is either $(\mathcal{A}, \mathfrak{A})$ or $(\mathcal{S}, \mathfrak{S})$, where $\mathfrak{S}$ is the class of all subsets of $\mathcal{S}$. For $q \in \mathcal{M}(\mathfrak{X})$, we define the total variation norm by

$$\|q\| = |q|(\mathcal{X}),$$

that is, $\|q\|$ is the total variation of q on $\mathcal{X}$. Note that $|q| = q_+ + q_-$, where q_+ and q_- are the positive and negative parts of q, respectively. For a probability measure p, $\|p\| = 1$; if $\tilde{p}$ is a probability measure, then $\|\tilde{p} - p\| \leq 2$.

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

We shall also use $\|\cdot\|$ to denote the ℓ_1-norm (absolute entry sum) for vectors and the corresponding induced norm (maximum absolute row sum, since we regard vectors as row vectors) for matrices; this will cause no confusion. Thus,

$$\|\tilde{p}_n - p_n\| = \|\tilde{\boldsymbol{p}}_n - \boldsymbol{p}_n\|, \tag{1}$$

where $\tilde{\boldsymbol{p}}_n$ and $\boldsymbol{p}_n$ are the distribution vectors of $\tilde{X}_n$ and X_n, respectively. Using the vector ℓ_1-norm, we define the ergodicity coefficient, $\tau(\boldsymbol{R})$, of a real matrix $\boldsymbol{R} = (r^{(ij)})$ (see [31]) as follows:

$$\tau(\boldsymbol{R}) := \sup_{\substack{\|\boldsymbol{v}\|=1 \\ \boldsymbol{v}\boldsymbol{e}^\top=0}} \|\boldsymbol{v}\boldsymbol{R}\| = \tfrac{1}{2}\max_{i,j}\sum_k |r^{(ik)} - r^{(jk)}|,$$

where the supremum is taken over real vectors $\boldsymbol{v}$, $\boldsymbol{e}$ is the row vector of all ones, and '$^\top$' denotes transpose. Notice that if the rows of $\boldsymbol{R}$ are probability vectors, then $\tau(\boldsymbol{R}) \le 1$.

To prove our perturbation bound for HMMs (see Theorem 1), we need the following definition.

Definition 1. A mapping $T : \mathcal{S} \times \mathfrak{A} \to \mathbb{R}$, such that $\mathcal{B} \mapsto T(i, \mathcal{B})$ is a measure belonging to $\mathcal{M}(\mathfrak{A})$ for any fixed $i \in \mathcal{S}$, is called a transition signed measure from $(\mathcal{S}, \mathfrak{S})$ to $(\mathcal{A}, \mathfrak{A})$.

(The general definition of a transition signed measure also requires that $i \mapsto T(i, \mathcal{B})$ be $\mathfrak{S}$-measurable for any fixed $\mathcal{B} \in \mathfrak{A}$ (see [9, Part 2]); this is satisfied automatically.) We denote by $\mathcal{T}_{\mathcal{S}\mathcal{A}}$ the family of all transition signed measures from $(\mathcal{S}, \mathfrak{S})$ to $(\mathcal{A}, \mathfrak{A})$. For any $T(i, \mathcal{B}) \in \mathcal{T}_{\mathcal{S}\mathcal{A}}$, $i \in \mathcal{S}$, $\mathcal{B} \in \mathfrak{A}$, we introduce the operator $T : \mathcal{M}(\mathfrak{S}) \to \mathcal{M}(\mathfrak{A})$ acting as follows:

$$(qT)(\mathcal{B}) := \int_{\mathcal{S}} q(\mathrm{d}x)T(x, \mathcal{B}) = \sum_{i\in\mathcal{S}} q(\{i\})T(i, \mathcal{B}),$$

where $q \in \mathcal{M}(\mathfrak{S})$ and $\mathcal{B} \in \mathfrak{A}$. For such operators, we define the norm by

$$\|T\| = \sup_{\|q\|=1} \|qT\|.$$

We are now ready to state and prove our results.

3. The perturbation bound for HMMs: derivation

The following theorem provides a sensitivity bound for HMMs.

Theorem 1. *For $z_n := \tilde{d}_n - d_n$, the following inequality holds:*

$$\|z_n\| \le d(\tilde{f}, f) + \kappa_f \|\tilde{\boldsymbol{p}}_n - \boldsymbol{p}_n\|, \qquad n \in \mathbb{Z}_+. \tag{2}$$

Here

$$d(\tilde{f}, f) = \max_{i\in\mathcal{S}} \|\tilde{f}^{(i)} - f^{(i)}\| \quad \textit{and} \quad \kappa_f = \tfrac{1}{2}\max_{i,j\in\mathcal{S}} \|f^{(i)} - f^{(j)}\|.$$

Proof. Define a transition signed measure $F(i, \mathcal{B})$ by

$$F(i, \mathcal{B}) = f^{(i)}(\mathcal{B}), \qquad i \in \mathcal{S},\ \mathcal{B} \in \mathfrak{A}.$$

The total probability formula gives

$$d_n = p_n F, \qquad \tilde{d}_n = \tilde{p}_n \tilde{F}, \tag{3}$$

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

for all $n \in \mathbb{Z}_+$, where the operator $\tilde{F}$ is defined for $\{(\tilde{X}_n, \tilde{Y}_n)\}$ in a similar way to F. Thus, we have

$$\|z_n\| = \|\tilde{p}_n(\tilde{F} - F) - (p_n - \tilde{p}_n)F\| \le \|\tilde{p}_n\|\|\tilde{F} - F\| + \|(p - \tilde{p})F\|. \tag{4}$$

Since $\|\tilde{p}_n\| \equiv 1$, the first term on the right-hand side of (4) is just $\|\tilde{F} - F\|$. For the second term, we have

$$\|(p - \tilde{p})F\| = \|\tilde{p} - p\|\left\|\frac{\tilde{p} - p}{\|\tilde{p} - p\|}F\right\| \le \|\tilde{p} - p\| \sup_{q \in \mathcal{M}_0} \|qF\|, \tag{5}$$

where $\mathcal{M}_0 = \{q \in \mathcal{M}(\mathfrak{S})\colon \|q\| = 1,\ q(\mathcal{S}) = 0\}$. It follows from Corollary 2.2 of [9, Part 2, Chapter 2] that

$$\|\tilde{F} - F\| = \max_{i \in \mathcal{S}} |\tilde{f}^{(i)} - f^{(i)}|(\mathcal{A}) = d(\tilde{f}, f),$$
$$\sup_{q \in \mathcal{M}_0} \|qF\| = \kappa_f.$$

This, together with (1), (4), and (5), proves the theorem.

Corollary 1. *Suppose that $\{X_n\}$ is ergodic, which means that there exist a probability vector $\boldsymbol{\pi}$ and positive numbers $C < \infty$ and $\rho < 1$ such that, for all $\boldsymbol{p}_0$,*

$$\|\boldsymbol{p}_n - \boldsymbol{\pi}\| \le C\rho^n, \qquad n \in \mathbb{Z}_+.$$

In this case, there exists a measure $\delta \in \mathcal{M}(\mathfrak{A})$ such that, for all $\boldsymbol{p}_0$,

$$\|d_n - \delta\| \le C\kappa_f\rho^n, \qquad n \in \mathbb{Z}_+.$$

The quantity κ_f in (2) never exceeds 1, and it can be arbitrarily small depending on how similar the distributions $f^{(i)}$ are to each other. The next two theorems give a necessary and a sufficient condition for the inequality $\kappa_f < 1$ to hold.

Theorem 2. *The inequality $\kappa_f < 1$ holds only if, for every pair $(i, j) \in \mathcal{S} \times \mathcal{S}$, $i \ne j$, there exists a set $\mathcal{C} \in \mathfrak{A}$ such that*

$$|(f^{(i)} - f^{(j)})(\mathcal{C})| < f^{(i)}(\mathcal{C}) + f^{(j)}(\mathcal{C}). \tag{6}$$

Proof. Suppose that, for some pair $(i_0, j_0) \in \mathcal{S} \times \mathcal{S}$, $i_0 \ne j_0$, there is no set $\mathcal{C} \in \mathfrak{A}$ such that (6) holds. Let $\bigcup_l \mathcal{A}_l$ be a finite partition of $\mathcal{A}$ into measurable subsets. We then have

$$\sum_l |(f^{(i_0)} - f^{(j_0)})(\mathcal{A}_l)| = \sum_l f^{(i_0)}(\mathcal{A}_l) + \sum_l f^{(j_0)}(\mathcal{A}_l) = 2.$$

Taking the supremum over all partitions $\bigcup_l \mathcal{A}_l$, we find that $\|f^{(i_0)} - f^{(j_0)}\| = 2$. This, together with the definition of κ_f, gives $\kappa_f = 1$.

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

Theorem 3. *The inequality $\kappa_f < 1$ holds if, for every pair $(i, j) \in \mathcal{S} \times \mathcal{S}$, $i \neq j$, there exists a set $\mathcal{C} \in \mathfrak{A}$ such that at least one of the following conditions is satisfied:*

(a) $(f^{(i)} - f^{(j)})_-(\mathcal{C}) = 0$ *and* $f^{(j)}(\mathcal{C}) > 0$;

(b) $(f^{(j)} - f^{(i)})_-(\mathcal{C}) = 0$ *and* $f^{(i)}(\mathcal{C}) > 0$.

Proof. Consider an arbitrary pair $(i, j) \in \mathcal{S} \times \mathcal{S}$, $i \neq j$. Without loss of generality, assume that, for this pair, condition (b) is satisfied. We then have

$$\|f^{(i)} - f^{(j)}\| = |f^{(j)} - f^{(i)}|(\mathcal{C}) + |f^{(j)} - f^{(i)}|(\mathcal{X} \setminus \mathcal{C}). \tag{7}$$

Since $|q| = q_+ + q_-$ and $q = q_+ - q_-$ for every measure $q \in \mathcal{M}(\mathfrak{A})$,

$$|f^{(j)} - f^{(i)}|(\mathcal{C}) = (f^{(j)} - f^{(i)})_+(\mathcal{C}) = (f^{(j)} - f^{(i)})(\mathcal{C}) < (f^{(j)} + f^{(i)})(\mathcal{C}).$$

This, together with (7), gives

$$\|f^{(i)} - f^{(j)}\| < (f^{(j)} + f^{(i)})(\mathcal{C}) + (f^{(j)} + f^{(i)})(\mathcal{X} \setminus \mathcal{C}) = (f^{(j)} + f^{(i)})(\mathcal{X}) = 2.$$

Since our pair (i, j) is arbitrary, $\kappa_f < 1$.

Example 1. (*Finite observation space.*) Let $\mathcal{A} = \{1, \ldots, M\}$, $M \geq 2$, and let $\mathfrak{A}$ be the class of all subsets of $\mathcal{A}$. We have $\|z_n\| = \|z_n\|$, $z_n := \tilde{\boldsymbol{d}}_n - \boldsymbol{d}_n$, where $\tilde{\boldsymbol{d}}_n$ and $\boldsymbol{d}_n$ are the distribution vectors of $\tilde{Y}_n$ and Y_n, respectively. The measures $f^{(i)}$ are defined by the emission probabilities $f^{(ij)} := \mathrm{P}[Y_n = j \mid X_n = i]$, $i \in \mathcal{S}$, $j \in \mathcal{A}$. The emission probabilities $\tilde{f}^{(ij)}$ for $\{(\tilde{X}_n, \tilde{Y}_n)\}$ are defined in a similar manner. Writing $\boldsymbol{F} = (f^{(ij)})$ and $\tilde{\boldsymbol{F}} = (\tilde{f}^{(ij)})$, we have

$$d(\tilde{f}, f) = \|\tilde{\boldsymbol{F}} - \boldsymbol{F}\|, \qquad \kappa_f = \tau(\boldsymbol{F}).$$

The inequality $\kappa_f < 1$ holds if and only if, for every pair $(i, j) \in \mathcal{S} \times \mathcal{S}$, $i \neq j$, there exists a state $k \in \mathcal{S}$ such that $f^{(ik)} > 0$ and $f^{(jk)} > 0$. This can be proved directly, but can also be obtained as a corollary to Theorems 2 and 3.

Example 2. (*Continuous observation space.*) Let $\mathcal{A} = \mathbb{R}$, and let $\mathfrak{A}$ be the Borel σ-algebra on $\mathbb{R}$. Suppose that the distributions $f^{(i)}$ and $\tilde{f}^{(i)}$ have continuous densities $\phi^{(i)}(t)$ and $\tilde{\phi}^{(i)}(t)$, respectively, for $t \in \mathbb{R}$. In this case,

$$d(\tilde{f}, f) = \max_{i \in \mathcal{S}} \int_{\mathbb{R}} |\tilde{\phi}^{(i)}(t) - \phi^{(i)}(t)| \, \mathrm{d}t, \qquad \kappa_f = \tfrac{1}{2} \max_{i,j \in \mathcal{S}} \int_{\mathbb{R}} |\phi^{(i)}(t) - \phi^{(j)}(t)| \, \mathrm{d}t.$$

We have $\kappa_f < 1$ if, for every pair $(i, j) \in \mathcal{S} \times \mathcal{S}$, $i \neq j$, there exists an interval $[a, b]$, $a < b$, such that $\phi^{(i)}(t) > 0$ and $\phi^{(j)}(t) > 0$ for all $t \in [a, b]$. This can be proved directly, but can also be obtained as a corollary to Theorem 3.

Theorem 1 shows that the distributions d_n may be strongly influenced by the perturbations in $f^{(i)}$, while the influence of perturbations in $\boldsymbol{p}_n$ is weak if the $f^{(i)}$ do not differ much. However, this theorem says nothing about the tightness of the inequality (2). Also, (2) does not explicitly show the effect of perturbations in $\boldsymbol{P}$ or $\boldsymbol{p}_0$ on d_n for $n > 0$. These questions will be addressed in Section 4.

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

4. The perturbation bound for HMMs: implications

The perturbation bound provided by Theorem 1 suggests a strong dependence of d_n on the distributions $f^{(i)}$. However, (2) is just an upper bound, so we may think that a tighter bound can be obtained which will not have this feature. Below we prove that (2) is tight, that is, there are examples when this inequality turns into an equality.

Theorem 4. *If the HMM $\{(X_n, Y_n)\}$ is as in Example 1 and the matrix $\boldsymbol{F}$ has at least two positive columns, then there exists a perturbed HMM, $\{(\tilde{X}_n, \tilde{Y}_n)\}$, such that*

(a) $d(\tilde{f}, f) \neq 0$ *and* $\tilde{\boldsymbol{p}}_n = \boldsymbol{p}_n$, *for all* $n \in \mathbb{Z}_+$,

(b) *the inequality (2) becomes an equality for all* $n \in \mathbb{Z}_+$.

Proof. Let k and l be the numbers of two positive columns of $\boldsymbol{P}$, and let $p_{\min}$ be the smallest entry in the kth and lth columns. Choose a number $r \in (0, p_{\min})$. Define $\boldsymbol{r}$ to be a vector of dimension M whose kth entry is r and whose lth entry is $-r$, all other entries being zero. Let $\boldsymbol{D}$ be a matrix of dimension $N \times M$ whose rows are all equal to $\boldsymbol{r}$. It is clear that $\boldsymbol{F} + \boldsymbol{D}$ is a row-stochastic matrix. Define the perturbed HMM $\{(\tilde{X}_n, \tilde{Y}_n)\}$ by $\tilde{\boldsymbol{P}} = \boldsymbol{P}$, $\tilde{\boldsymbol{F}} = \boldsymbol{F} + \boldsymbol{D}$, and $\tilde{\boldsymbol{p}}_0 = \boldsymbol{p}_0$. Thus, we have $\tilde{\boldsymbol{p}}_n = \boldsymbol{p}_n$ for all $n \in \mathbb{Z}_+$. Therefore,

$$\|z_n\| = \|\tilde{\boldsymbol{p}}_n(\tilde{\boldsymbol{F}} - \boldsymbol{F})\| = \|\tilde{\boldsymbol{p}}_n \boldsymbol{D}\| = \|\boldsymbol{r}\|, \qquad n \in \mathbb{Z}_+.$$

Since the right-hand side of (2) equals $\|\tilde{\boldsymbol{F}} - \boldsymbol{F}\| = \|\boldsymbol{r}\|$, the theorem follows.

Corollary 2. *Suppose that, for all finite HMMs $\{(X_n, Y_n)\}$ with $\boldsymbol{F}$ as in Theorem 4, we have a bound of the form*

$$\|z_n\| \leq C_1 \|\tilde{\boldsymbol{F}} - \boldsymbol{F}\| + C_2 \|\tilde{\boldsymbol{p}}_n - \boldsymbol{p}_n\|, \qquad n \in \mathbb{Z}_+,$$

where C_1 and C_2 depend only on $\boldsymbol{P}$ and $\boldsymbol{F}$. Then we must have $C_1(\boldsymbol{P}, \boldsymbol{F}) \geq 1$ for all such HMMs.

Theorem 5. *If the HMM $\{(X_n, Y_n)\}$ is as in Example 2 and $\phi^{(i)}(t) > 0$ for all $t \in \mathbb{R}$ and $i \in \mathcal{S}$, then there exists a perturbed HMM $\{(\tilde{X}_n, \tilde{Y}_n)\}$ which possesses the properties (a) and (b) as stated in Theorem 4.*

Proof. Consider the interval $[-a, a]$, $a > 0$. Let $\varepsilon > 0$ be such that $\phi^{(i)}(t) > \varepsilon$ for all $i \in \mathcal{S}$ and $t \in [-a, a]$. Define the function $s(t)$ by

$$s(t) = \begin{cases} \varepsilon \sin\left(\dfrac{t\pi}{a}\right), & t \in [-a, a], \\ 0, & t \in \mathbb{R} \setminus [-a, a], \end{cases}$$

and define the perturbed chain $\{(\tilde{X}_n, \tilde{Y}_n)\}$ by $\tilde{\boldsymbol{P}} = \boldsymbol{P}$, $\tilde{\boldsymbol{p}}_0 = \boldsymbol{p}_0$, and $\tilde{\phi}^{(i)}(t) = \phi^{(i)}(t) + s(t)$, for all $i \in \mathcal{S}$ and $t \in \mathbb{R}$ (it is clear that $\tilde{\phi}^{(i)}(t)$ are valid probability density functions). Thus, all $\tilde{f}^{(i)} - f^{(i)}$ are identical, and we have

$$\|z_n\| = \|\tilde{p}_n(\tilde{F} - F)\| = \|\tilde{f}^{(1)} - f^{(1)}\|.$$

Since the right-hand side of (2) equals $\|\tilde{F} - F\| = \|\tilde{f}^{(1)} - f^{(1)}\|$, the theorem follows.

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

Remark 1. An analogue of Corollary 2 holds for an HMM defined as in Theorem 5.

Remark 2. For the HMMs $\{(X_n, Y_n)\}$ considered in Theorems 4 and 5, $\kappa_f < 1$.

The tightness of (2) implies that this bound cannot be improved by a constant factor. Theorems 4 and 5, Corollary 2, and Remark 1 show that it is also impossible to improve the first term on the right-hand side of (2) by a constant factor; for important special cases of HMM, we cannot in fact improve it by *any* factor. This makes us believe that a strong dependence of d_n on the functions $f^{(i)}$ is a real property of HMMs, and that (2) quantifies it adequately. However, improvements in the second term of the bound (2) may be possible, and the actual influence of perturbations in $\boldsymbol{p}_n$ may be even weaker than is suggested by (2).

We now pass to the discussion of the influence of the perturbations in $\boldsymbol{P}$ and $\boldsymbol{p}_0$ on the distributions of Y_n. To study this influence, we can use Theorem 1 and apply perturbation bounds for Markov chains to bound $\|\tilde{\boldsymbol{p}}_n - \boldsymbol{p}_n\|$ (for different bounds, see [8], [14], [23]). We should mention that perturbation bounds for the stationary distribution require that $\{X_n\}$ have a unique stationary distribution, and perturbation bounds which are uniform over $n \in \mathbb{N} := \{1, 2, \ldots\}$ require that $\{X_n\}$ be ergodic. The uniform perturbation bounds have the form

$$\sup_{n\in\mathbb{N}} \|\boldsymbol{x}_n\| \leq \kappa_1 \|\boldsymbol{x}_0\| + \kappa_2 \|\boldsymbol{E}\|,$$

where $\boldsymbol{x}_n = \tilde{\boldsymbol{p}}_n - \boldsymbol{p}_n$, $\boldsymbol{E} = \tilde{\boldsymbol{P}} - \boldsymbol{P}$, and κ_1 and κ_2 are numbers depending on the parameters of the unperturbed Markov chain. From this expression and the inequality (2), it is clear that if $\kappa_f \kappa_1 < 1$ and $\kappa_f \kappa_2 < 1$, then the distributions of Y_n are less sensitive to perturbations in $\boldsymbol{P}$ and $\boldsymbol{p}_0$ than to perturbations in $\{f^{(i)}\}$. Below we consider situations in which simple approximate bounds can be obtained.

If $\{X_n\}$ is ergodic then there exists an integer m such that $\tau(\boldsymbol{P}^m) < 1$. The inequality (3.17) of [23] gives

$$\sup_{n\in\mathbb{N}} \|\boldsymbol{x}_n\| \leq \sup_{n\in\mathbb{N}} (\tau(\boldsymbol{P}^m))^{\lfloor n/m \rfloor} \|\boldsymbol{x}_0\| + \frac{m\|\boldsymbol{E}\|}{1 - \tau(\boldsymbol{P}^m)}, \tag{8}$$

where $\lfloor x \rfloor$ is the largest integer less than or equal to x. An important special case is when $\tau(\boldsymbol{P}) < 1$. Such matrices $\boldsymbol{P}$ are sometimes called *scrambling matrices* [30]. Note that positive stochastic matrices are scrambling; the HMMs whose sensitivity was investigated in [13] and [24] had positive transition matrices. If $\boldsymbol{P}$ is scrambling then, by setting $m = 1$ in (8) and combining it with (2), we arrive at

$$\begin{aligned}
\sup_{n\in\mathbb{N}} \|z_n\| &\leq d(\tilde{f}, f) + \kappa_f \tau(\boldsymbol{P}) \|\boldsymbol{x}_0\| + \frac{\kappa_f \|\boldsymbol{E}\|}{1 - \tau(\boldsymbol{P})} \\
&= d(\tilde{f}, f) + \kappa_f \tau(\boldsymbol{P}) \|\boldsymbol{x}_0\| + \kappa_f \|\boldsymbol{E}\| (1 + \tau(\boldsymbol{P}) + \tau^2(\boldsymbol{P}) + \cdots) \\
&= d(\tilde{f}, f) + \kappa_f \|\boldsymbol{E}\| + O(\kappa_f \tau(\boldsymbol{P}))
\end{aligned}$$

(since $\sup_{n\in\mathbb{N}} (\tau(\boldsymbol{P}))^n = \tau(\boldsymbol{P})$). If both κ_f and $\tau(\boldsymbol{P})$ are small enough, then the term $O(\kappa_f \tau(\boldsymbol{P}))$ can be neglected, and we obtain an approximate bound which is equivalent to (2) with $\|\tilde{\boldsymbol{p}}_n - \boldsymbol{p}_n\|$ substituted with $\boldsymbol{E}$. Thus, at this level of approximation, the initial distribution of $\{X_n\}$ has no effect on d_n – its influence is a 'second-order effect'.

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

If $m > 1$ then $\sup_{n\in\mathbb{N}}(\tau(\boldsymbol{P}^m))^{\lfloor n/m\rfloor} = 1$, and (2) and (8) give

$$\sup_{n\in\mathbb{N}}\|z_n\| \le d(\tilde{f}, f) + \kappa_f\|\boldsymbol{x}_0\| + \frac{m\kappa_f\|\boldsymbol{E}\|}{1-\tau(\boldsymbol{P}^m)}$$
$$= d(\tilde{f}, f) + \kappa_f(\|\boldsymbol{x}_0\| + m\|\boldsymbol{E}\|) + O(\kappa_f\tau(\boldsymbol{P}^m)).$$

In this case, the influence of perturbations in $\boldsymbol{P}$ and the influence of perturbations in $\boldsymbol{p}_0$ are of the same order for small m. However, even for large m, if the distributions $f^{(i)}$ are similar enough to each other, then d_n is much less sensitive to perturbations in $\boldsymbol{p}_0$ and $\boldsymbol{P}$ than to perturbations in $f^{(i)}$. For large m, as in the case $m = 1$, the perturbations in $\boldsymbol{p}_0$ are likely to have the smallest effect on d_n.

5. Generalizations and applications

We have presented a new approach to sensitivity analysis for HMMs. We proved that, in many important cases, the distributions of an HMM show a weaker dependence on the transition probabilities than on the emission probabilities. This property has an important statistical implication: the transition probabilities may typically be more difficult to estimate.

It should be noted that our results admit broad generalizations. The reason for this possibility lies in the nature of the problem, as well as in our operator-theoretic treatment (see Theorem 1), which works in a variety of situations. As a matter of fact, Theorem 1 provides a sensitivity bound for mixtures of distributions and, thus, can be extended to any situation where the distribution of interest can be represented as a mixture.

One direction for generalization is related to the features of the underlying Markov chain $\{X_n\}$. Clearly, we may consider cases of inhomogeneous chains and chains on an infinite state space. (HMMs whose underlying Markov chain has a general state space are frequently called state space models (see, e.g. [12]).) Theorem 1 can be generalized to all of these cases, and can be used in conjunction with the corresponding perturbation bounds (see Section 1). We can also combine Corollary 1 with convergence bounds for Markov chains (see [11], [14], [19], [23], and [32]) to bound the speed of convergence of an HMM to stationarity.

Yet another direction for extension is as follows. Instead of considering the Markov chain $\{X_n\}$, we may consider regime processes which are not necessarily Markov. All we need is to have perturbation bounds for the distributions of such processes. We finish this section by giving four examples.

Example 3. (*A mixture process.*) In our definition of an HMM, the random variables X_n are dependent. If we define $\{X_n\}$ to be a sequence of independent and identically distributed random variables taking values in $\mathcal{S}$ with probabilities p_i, $i \in \mathcal{S}$, then we obtain a mixture process, in which the distribution of all Y_n is the mixture $\sum_{i\in\mathcal{S}} p_i f^{(i)}$ (see [10] and references therein). This case is especially simple because we do not need to derive perturbation bounds for the regime process; we just substitute the size of the perturbation in p_i directly into the analogue of (2). Obviously, Theorems 2 and 3 are applicable, and analogues of Theorems 4 and 5 can be obtained, showing the same qualitative stability features as for HMMs.

Example 4. (*HMMs and continuous-time Markov chains.*) HMMs whose regimes are derived from a continuous-time Markov chain arise in ion channel modeling (see, e.g. [27]). A channel is modeled by a finite, homogeneous, continuous-time Markov chain with generator $\boldsymbol{Q}$. The matrix $\boldsymbol{Q}$ is usually assumed to be reversible. At discrete time moments $n\Delta$, $n \in \mathbb{N}$, $\Delta > 0$, we sample the chain, but the recordings are corrupted by noise. The resulting stochastic process

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

is an HMM with continuous observation space, whose underlying Markov chain has transition probability matrix $\exp(\Delta \boldsymbol{Q})$. The sensitivity of this chain with respect to perturbations in $\boldsymbol{Q}$ can be analyzed by using Theorem 1 in conjunction with perturbation bounds for the distribution of a (reversible) continuous-time Markov chain.

Example 5. (*Multidimensional distributions of HMMs.*) The theory developed so far concerns the one-dimensional distributions of HMMs. Let us take a look at multidimensional distributions, which are frequently of interest in applications of HMMs. Suppose that we would like to obtain sensitivity bounds for the joint distribution of the random variables $Y_n, Y_{n+1}, \ldots, Y_{n+m}$ for a discrete HMM (see Example 1). By the definition of an HMM, we can write

$$\begin{aligned}&\mathrm{P}[Y_n = i_0,\ Y_{n+1} = i_1, \ldots, Y_{n+m} = i_m] \\ &\quad = \sum_{j_0, j_1, \ldots, j_m} \mathrm{P}[X_n = j_0,\ X_{n+1} = j_1, \ldots, X_{n+m} = j_m] \prod_{k=0}^{m} f^{(j_k i_k)}.\end{aligned}$$

We see that this formula defines a mixture distribution. Therefore, we can use an argument similar to the proof of Theorem 1 to obtain a perturbation bound for the vectors formed by the probabilities $\mathrm{P}[Y_n = i_0, Y_{n+1} = i_1, \ldots, Y_{n+m} = i_m]$ (such a vector will have M^{m+1} components, each corresponding to a possible choice of $i_0, i_1, \ldots, i_m$). This bound, together with the analogues of the theorems in Section 4, shows that the multidimensional distributions are more sensitive to changes in the emission probabilities than to changes in the finite-dimensional distributions of the Markov chain $\{X_n\}$. Obtaining perturbation bounds for the latter distributions in terms of the perturbations in the transition probabilities seems to be a difficult task.

Example 6. (*A semi-Markov process.*) Let $\{g^{(i)}\}$, $i \in \mathcal{S}$, be a family of probability distributions on $\mathbb{R}_+$, and let $\boldsymbol{S} = (s^{(ij)})$ be a stochastic matrix of dimension $N \times N$. Suppose that the distributions $g^{(i)}$ have continuous densities $\gamma^{(i)}(t)$, $t \geq 0$, and set $Q_{ij}(t) = g^{(i)}(t) s^{(ij)}$. Let V be a semi-Markov process on $\mathcal{S}$ determined by the kernel $Q_{ij}(t)$ and the initial distribution vector $\boldsymbol{s}_0$ (for the definition and basic properties of semi-Markov processes, see the classic work [26]). The Markov chain governing the jumps of V has transition matrix $\boldsymbol{S}$ and distribution vector $\boldsymbol{s}_n = (s_n^{(i)})$, $i \in \mathcal{S}$, $n \in \mathbb{Z}_+$. We also define (in a similar way) a perturbed semi-Markov process, $\tilde{V}$, with corresponding characteristics $\tilde{g}^{(i)}$, $\tilde{\gamma}^{(i)}(t)$, and $\tilde{\boldsymbol{s}}_n$. Note that V, as well as $\tilde{V}$, is a special type of semi-Markov process, in which sojourn times depend only on the current state. Such processes arise in applications, e.g. in ion channel modeling [2].

The duration of the nth sojourn of the process V has distribution denoted by h_n, which is given by

$$h_n(\mathcal{B}) = \sum_{i,j \in \mathcal{S}} s_{n-1}^{(i)} g^{(i)}(\mathcal{B}) s^{(ij)} = \sum_{i \in \mathcal{S}} s_{n-1}^{(i)} g^{(i)}(\mathcal{B}), \qquad n \in \mathbb{N}. \tag{9}$$

Here, $\mathcal{B}$ belongs to the Borel σ-algebra on $\mathbb{R}_+$. A similar formula holds for the corresponding distributions $\tilde{h}_n$ defined for $\tilde{V}$. The equality (9) is analogous to the expressions (3) for the distributions of HMMs. Therefore, we can use the argument in the proof of Theorem 1 to obtain the following perturbation bound for the distributions of the sojourn times of V:

$$d(\tilde{h}_n, h_n) \leq d(\tilde{g}, g) + \kappa_g \|\tilde{\boldsymbol{s}}_{n-1} - \boldsymbol{s}_{n-1}\|, \qquad n \in \mathbb{N},$$

https://doi.org/10.1239/jap/1127322017 Published online by Cambridge University Press

where

$$d(\tilde{h}_n, h_n) = \int_{\mathbb{R}_+} |\tilde{\chi}_n(t) - \chi_n(t)|\,\mathrm{d}t,$$

$$d(\tilde{g}, g) = \max_{i \in \mathcal{S}} \int_{\mathbb{R}_+} |\tilde{\gamma}^{(i)}(t) - \gamma^{(i)}(t)|\,\mathrm{d}t,$$

$$\kappa_g = \tfrac{1}{2} \max_{i,j \in \mathcal{S}} \int_{\mathbb{R}_+} |\gamma^{(i)}(t) - \gamma^{(j)}(t)|\,\mathrm{d}t;$$

$\tilde{\chi}_n(t)$ and $\chi_n(t)$ are probability density functions corresponding to $\tilde{h}_n$ and h_n, respectively. As in the case of HMMs, this bound should be used in conjunction with perturbation bounds for Markov chains. Note that if $\gamma^{(i)}(t) > 0$ for all $i \in \mathcal{S}$ and $t > 0$, then $\kappa_g < 1$.

6. Conclusion

We have shown that, in general, the behavior of HMMs tends to be more sensitive to the choice of the emission probabilities than to the choice of the transition probabilities. This conclusion supports the approach taken by the developers of gene-finding algorithms of the GENEMARK^{TM} family (see [5], [6], [17]). The heuristically derived values are used as initial estimates for the emission probabilities in the self-training version of the HMM-based algorithm (see [6]). These estimates reflect the actual compositional properties of the genomic DNA sequence, and seem to be crucial to the convergence of the iterative training procedure to biologically relevant values. The initial choice of the transition probabilities is far more arbitrary; our experience shows that altering these probabilities produces small relative changes in the final values of the HMM parameters as well as in final gene predictions. The results of this paper provide grounds to suggest that the approach to HMM parameter estimation described in [5], [6], and [17] may serve as a guideline for the developers of self-training versions of other HMM-based bioinformatic algorithms.

Acknowledgements

This work has been supported by the NIH grant HG00783 to Mark Borodovsky. The authors are grateful to the anonymous referee for valuable remarks on an earlier version of the paper.

References

[1] Archer, G. E. B. and Titterington, D. M. (2002). Parameter estimation for hidden Markov chains. *J. Statist. Planning Infer.* **108,** 365–390.

[2] Ball, F., Milne, R. K. and Yeo, G. F. (1991). Aggregated semi-Markov processes incorporating time interval omission. *Adv. Appl. Prob.* **23,** 772–797.

[3] Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite Markov chains. *Ann. Math. Statist.* **37,** 1554–1563.

[4] Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41,** 164–171.

[5] Besemer, J. and Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Res.* **27,** 3911–3920.

[6] Besemer, J., Lomsadze, A. and Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.* **29,** 2607–2618.

[7] Bickel, P. G., Ritov, Y. and Rydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.* **26,** 1614–1635.

[8] Cho, G. E. and Meyer, C. D. (2001). Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra Appl.* **335,** 137–150.

[9] COHEN, J. E., KEMPERMAN, J. H. B. AND ZBĂGANU, GH. (1998). *Comparisons of Stochastic Matrices.* Birkhäuser, Boston, MA.
[10] EPHRAIM, Y. AND MERHAV, N. (2002). Hidden Markov processes. *IEEE Trans. Inf. Theory* **48,** 1518–1569.
[11] GRANOVSKY, B. L. AND ZEIFMAN, A. I. (2000). Nonstationary Markovian queues. *J. Math. Sci. (New York)* **99,** 1415–1438.
[12] JENSEN, J. L. AND PETERSEN, N. V. (1999). Asymptotic normality of the maximum likelihood estimator in state space models. *Ann. Statist.* **27,** 514–535.
[13] JUANG, B.-H. AND RABINER, L. R. (1985). A probabilistic distance measure for hidden Markov models. *AT&T Tech. J.* **64,** 391–408.
[14] KARTASHOV, N. V. (1996). *Strong Stable Markov Chains.* VSP, Utrecht.
[15] KOSKI, T. (2001). *Hidden Markov Models for Bioinformatics.* Kluwer, Dordrecht.
[16] LEROUX, B. G. (1992). Maximum likelihood estimation for hidden Markov models. *Stoch. Process. Appl.* **40,** 127–143.
[17] LUKASHIN, A. L. AND BORODOVSKY, M. (1998). GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26,** 1107–1115.
[18] MCDONALD, I. AND ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-Valued Time Series.* Chapman and Hall, London.
[19] MITROPHANOV, A. YU. (2003). Stability and exponential convergence of continuous-time Markov chains. *J. Appl. Prob.* **40,** 970–979.
[20] MITROPHANOV, A. YU. (2004). The spectral gap and perturbation bounds for reversible continuous-time Markov chains. *J. Appl. Prob.* **41,** 1219–1222.
[21] MITROPHANOV, A. YU. (2005). Ergodicity coefficient and perturbation bounds for continuous-time Markov chains. *Math. Ineq. Appl.* **8,** 159–168.
[22] MITROPHANOV, A. YU. (2005). Estimates of sensitivity to perturbations for finite homogeneous continuous-time Markov chains. *Teor. Veroyat. Primen.* **50,** 371–379 (in Russian). English translation to appear in *Theory Prob. Appl.*
[23] MITROPHANOV, A. YU. (2005). Sensitivity and convergence of uniformly ergodic Markov chains. To appear in *J. Appl. Prob.*
[24] PESHKIN, L. AND GELFAND, M. S. (1999). Segmentation of yeast DNA using hidden Markov models. *Bioinformatics* **15,** 980–986.
[25] PETRIE, T. (1969). Probabilistic functions of finite state Markov chains. *Ann. Math. Statist.* **40,** 97–115.
[26] PYKE, R. (1961). Markov renewal processes: definitions and preliminary properties. *Ann. Math. Statist.* **32,** 1231–1242.
[27] QIN, F., AUERBACH, A. AND SACHS, F. (2000). A direct optimization approach for hidden Markov modelling for single channel kinetics. *Biophys. J.* **79,** 1915–1927.
[28] RABINER, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77,** 257–284.
[29] ROSALES, R., STARK, J. A., FITZGERALD, W. J. AND HLADKY, S. B. (2001). Bayesian restoration of ion channel recordings using hidden Markov models. *Biophys. J.* **80,** 1088–1103.
[30] SENETA, E. (1981). *Nonnegative Matrices and Markov Chains.* Springer, New York.
[31] SENETA, E. (1984). Explicit forms for ergodicity coefficients and spectrum localization. *Linear Algebra Appl.* **60,** 187–197.
[32] ZEIFMAN, A. I. (1995). Upper and lower bounds on the rate of convergence for nonhomogeneous birth and death processes. *Stoch. Process. Appl.* **59,** 157–173.
[33] ZEIFMAN, A. I. AND ISAACSON, D. L. (1994). On strong ergodicity for nonhomogeneous continuous-time Markov chains. *Stoch. Process. Appl.* **50,** 263–273.