Hidden Markov Model

Yunkyu Song

6/8, 2020

Abstract

In recent years, a hidden Markov model (HMM) has been successfully used in varieties of academic fields including statistics, computer science, computational biology, and so on. In this paper, we state important quantities in estimations involving HMM and MLE and give a survey of one way to compute the maximum likelihood estimator of unknown parameter in a situation where the HMM of interest is not fully specified with some assumptions, which will be stated in the paper.

Contents

1 Introduction

Simply speaking, a Hidden Markov Model (HMM) is a Markov chain that has an observable component $\{Y_i\}$ and a hidden component $\{X_i\}$; that is, HMM is a bivariate stochastic process $\{X_i, Y_i\}_{\geq 0}$ such that $\{X_i\}$ is a Markov chain, $\{Y_i\}$ is a family of random variables that are independent conditionally on $\{X_i\}$, and each Y_n has a conditional distribution that only depends on X_n. A rigorous development of HMM was first done by Baum, Eagon, Petrie, Soules and Weiss in late 1960s and early 1970s, although the term was first invented by L. P. Neuwirth - according to [5]. Since then, there has been various applications of HMM in a wide range of areas such as speech recognition, financial mathematics, cryptoanalysis, gene prediction, and so on.

In this paper, we aim to explore the basic of HMM and MLE, and also the basic of one of the popular way to estimate the unknown parameter of HMM, with a somewhat focus on stating difficulties and limits on computing important quantities.

2 Hidden Markov Model

Due to the difficulty of defining conditional independence in general settings, most literature discussing HMM uses more sophisticated than what we had before. In this section, we give a formal definition of HMM, state three elementary estimation problems in HMM, and focus on one problem we would need to compute likelihood in later sections.

2.1 Formal definition of HMM

2.1.1 Kernel

Given measurable spaces $(E, \mathcal{E})$, which we will call a **state space**, and $(F, \mathcal{F})$, a map $\phi : E \times \mathcal{F} \to \overline{\mathbb{R}}_{\geq 0}$ is called a **kernel** from $(E, \mathcal{E})$ to $(F, \mathcal{F})$, if it satisfies the following:

1. For a fixed $x \in E$, $P_x \equiv \phi(x, \cdot) : \mathcal{F} \to \mathbb{R}_{\geq 0}$ is a measure, so that $(F, \mathcal{F}, P_x)$ is a measure space.
2. For a fixed $A \in \mathcal{F}$, $\phi_A(\cdot) \equiv \phi(\cdot, A)$ is measurable; that is, $\phi_A : (E, \mathcal{E}) \to (\overline{\mathbb{R}}_{\geq 0}, \mathcal{B}_{\geq 0})$ is a $\mathcal{E}$-$\mathcal{B}_{\geq 0}$ measurable function.

If P_x is a probability measure for each $x \in E$, ϕ is called a **transition kernel**. Additionally, if $(E, \mathcal{E}) = (F, \mathcal{F})$, ϕ is called a **Markov transition kernel** on $(E, \mathcal{E})$. Given ϕ is a transition kernel and μ is a measure on $(F, \mathcal{F})$ that dominates P_x for each $x \in E$, with appropriate assumptions about σ-finiteness of the measures (every measure space is assumed to be σ-finite in this paper), for each $x \in E$, there exists a measurable map

$p : (E \times F, \mathcal{E} \otimes \mathcal{F}) \to (\overline{\mathbb{R}}_{\geq 0}, \mathcal{B}_{\geq 0})$ such that

$$\phi(x, A) = \int_A p(x, y)\mu(dy) \equiv \int_A p(x, y)d\mu$$

for any $A \in \mathcal{A}$ by the Radon-Nikodym theorem, and p is called a **transition density** with respect to μ.

We state several properties of transition kernels below.

1. If ϕ_1 is a transition kernel from $(E, \mathcal{E})$ to $(F, \mathcal{F})$ and ϕ_2 is a transition kernel from $(F, \mathcal{F})$ to $(G, \mathcal{G})$, then the product of ϕ_1 and ϕ_2 defined as

   $$\phi_1 \otimes \phi_2(x, A) \equiv \int \phi_1(x, dy)\phi_2(y, A) \equiv \int \phi_2(y, A)d\phi_1(x, \cdot) \text{ for } x \in E, A \in \mathcal{Y}$$

   is a transition kernel from $(E, \mathcal{E})$ to $(G, \mathcal{G})$. The product is associative; that is, if ϕ_3 is a transition kernel from $(G, \mathcal{G})$ to $(H, \mathcal{H})$, $(\phi_1 \otimes \phi_2) \otimes \phi_3 = \phi_1 \otimes (\phi_2 \otimes \phi_3)$. $\otimes$ may be omitted often times.

2. If Q is a Markov transition kernel on $(E, \mathcal{E})$, then its iterates is defined recursively as:

   $$Q^1 = Q, Q^n = Q \otimes Q^{n-1}$$

   and it satisfies Chapman-Kolmogorov equation defined as: $Q^{n+m} = Q^n Q^m$.

   If there exists a measure μ on $(E, \mathcal{E})$ that dominates $Q(x, \cdot)$ for every $x \in E$, then the formula for the transition density of $Q^n(x, \cdot)$ with respect to μ is given by

   $$\int_{X^{n-1}} q(x, x_1) \cdots q(x_{n-1}, \cdot)\mu(dx_1) \cdots \mu(dx_{n-1})$$

   where $q(x, \cdot)$ is a transition density of $Q(x, \cdot)$.

3. Transition kernels operate on real measurable functions. Let ϕ_1 be defined as in the first property. If f is a real measurable function on G, then with an assumption that the integral is well defined, let $\phi_1 f$ to be defined as

   $$\phi_1 f(x) \equiv \int Q(x, dy)f(y) \text{ for } x \in E$$

   is a real measurable function. An alternative notation of $\phi_1 f(x)$ is $\phi_1(e, f)$.

An adapted stochastic process $\{X_i\}_{\geq 0}$ to a filtration $\{\mathcal{E}_i\}_{\geq 0}$ on a state probability space $(E, \mathcal{E}, P)$ with a family of Markov transition kernels on the state space $\{Q_i\}_{\geq 0}$ is called an **inhomogeneous Markov chain** under P with respect to $\{\mathcal{E}_i\}$ and with $\{Q_i\}$, if, for every $k \geq 0$ and $A \in \mathcal{E}$, it satisfies

$$P(X_k \in A|\mathcal{E}_k) = Q_k(X_k, A)$$

If $|\{Q_i\}| = 1$, $\{X_i\}_{\geq 0}$ is called a **homogeneous Markov chain**. The probability measure P_0 defined as $P_0(A) = Q_0(X_0 \in A)$ is called an **initial measure**.

2.1.2 Hidden Markov Model

We finally give a formal definition of Hidden Markov Model. Let $(E, \mathcal{E})$, which we will call an and $(F, \mathcal{F})$ be measurable spaces, Q be a Markov transition kernel on $(E, \mathcal{E})$, which is called a **signal state space**, and

G be a transition kernel from $(E, \mathcal{E})$ to $(F, \mathcal{F})$, which is called an **observation state space**. Let a joint Markov transition kernel T be defined as

$$T((x,y),C) = \iint_C Q(x,dx')G(x',dy') \equiv \iint_C dQ(x,\cdot)dG(x',\cdot)$$

for $(x,y) \in E \times F$ and $C \in \mathcal{E} \otimes \mathcal{F}$. Also, let ν be a probability measure on $(E, \mathcal{E})$. A bivariate Markov chain $\{X_i, Y_i\}_{\geq 0}$ with a Markov transition kernel T and an initial measure $\nu \otimes G$ is called a **hidden Markov model**. If $E \times F$ is countable, it is discrete. If there exists a probability measure μ on $(F, \mathcal{F})$ that dominates $G(x,\cdot)$ for any $x \in E$, then $\{X_i, Y_i\}_{\geq 0}$ is **partially dominated**. Note that if $\{X_i, Y_i\}_{\geq 0}$ is partially dominated, then $G(x,\cdot)$ has a transition density g with respect to μ, so $T((x,y),C)$ can be written as

$$\iint_C Q(x,dx')g(x',y')\mu(dy') \equiv \iint_C g(x',y')dQ(x,\cdot)d\mu$$

If $\{X_i, Y_i\}_{\geq 0}$ is partially dominated and if there exists a probability measure λ on $(E, \mathcal{E})$ that dominates ν and $Q(x,\cdot)$ for any $x \in E$ additionally, then $\{X_i, Y_i\}_{\geq 0}$ is **fully dominated**. Note that if $\{X_i, Y_i\}_{\geq 0}$ is fully dominated, then T is dominated by a product measure $\lambda \otimes \mu$ and admits a density t with respect to the product measure so that T can be written as:

$$\iint_C t((x,y),(x',y'))d\mu d\lambda$$

Throughout the paper, we generally limit our focus on dominated models.

Proposition 1. *Suppose $\{X_i, Y_i\}_{\geq 0}$ is a Markov Chain over $E \times F$ with ν and T defined as above. Then, for any integer j, $\{f_i : 1 \leq i \leq j\} \subset M_b(F)$, and $\{k_1 < \cdots < k_j\}$, $E_\nu[\prod_{i=1}^{j} f_i(Y_{k_i})] = \prod_{i=1}^{j} \int_F f_i(y)G(X_{k_i}, dy)$.*

Corollary 1. *$Y_{k_1}, \cdots, Y_{k_j}$ are conditionally independent given $(X_{k_1}, \cdots, X_{k_j})$.*

Corollary 2. *For $m \notin \{k_1, \cdots, k_j\}$, Y_m and $(X_{k_1}, \cdots, X_{k_j})$ are conditionally independent given X_m.*

Note that the Corollary 1 and the Corollary 2 give the conditional independence properties of HMM that we had in our first definition of HMM. Although the definition of HMM given here does not specify, the term "hidden" would make sense only if $\{X_i\}$ is indeed hidden. In that sense, $\{X_k\}$ here can be seen as an imaginary intermediate process that is only useful in defining the distribution of the observed process, $\{Y_i\}$.

2.2 Smoothing recursion

One of the questions in estimation that arise naturally given a set up of HMM is: given a model and observations $Y_0, \cdots, Y_n$, what can be said about hidden component(s) X_k(s)? Roughly speaking, if $0 \leq k \leq n$, this is called a smoothing problem; if $k \geq n$, this is called a prediction problem; if the observations are not fixed and $k = n$, this is called a filtering problem. In this paper, we will mainly consider smoothing, and a way to compute the likelihood function, which will be defined here.

2.2.1 Smoothing

For any $f \in M_b((E \times F)^{n+1})$, $E_\nu[f(X_0, Y_0, \cdots, X_n, Y_n)]$ can be written as

$$\int \cdots \int f(x_0, y_0, \cdots, x_n, y_n) \cdot \nu(dx_0) g(x_0, y_0) \cdot \prod_{i=1}^{n} (Q(x_{i-1}, dx_i) g(x_i, y_i)) \mu_n(dy_0, \cdots, dy_n)$$
$$\equiv \int \cdots \int f(x_0, y_0, \cdots, x_n, y_n) \cdot \nu(dx_0) g(x_0, y_0) \cdot \prod_{i=1}^{n} (Q(x_{i-1}, dx_i) g(x_i, y_i)) d\mu^{\otimes(n+1)}$$

.

With μ_n as above, we define $L_{\nu,n}$, the **likelihood function** of the observations, to be the probability density function of $Y_0, \cdots, Y_n$ with respect to μ_n and $\ell_{\nu,n}$, the **log likelihood function**, to be $\log(L_{\nu,n})$. The likelihood function is given explicitly for $(y_0, \cdots, y_n) \in F^{n+1}$ as

$$L_{\nu,n}(y_0, \cdots, y_n) = \int \cdots \int \nu(dx_0) g(x_0, y_0) \prod_{i=1}^{n} (Q(x_{i-1}, dx_i) g(x_i, y_i))$$

Marginalizing with respect to hidden components, $E[f(Y_{0:n})]$ can be written as

$$\int \cdots \int f(y_{0:n}) L_{\nu,n}(y_{0:n}) \mu_n(dy_0, \cdots, dy_n)$$

Let $\phi_{\nu,k:l|n}$ be a version of the conditional distribution of $X_{k:l}$ given $Y_{0:n}$; that is,

1. $\phi_{\nu,k:l|n}$ is a transition kernel from F^{n+1} to E^{n+1}.
2. For any $f \in M_b(E^{l-k+1})$, $E_\nu[f(X_{k:l})] =_{a.s} \int \cdots \int f(x_{k:l}) \phi_{\nu,k:l|n}(Y_{0:n}, dx_{k:l})$.

If we are interested in $\phi_{\nu,0:n|n}$, we are considering a **joint smoothing** problem. If we are only interested in a conditional distribution of one hidden element: $\phi_{\nu,k|n}$ for $k \in \{0\} \cup [n]$, then we are considering a a (**marginal**) **smoothing** problem.

The proposition below gives a constructive approach to define the quantities above in terms of the elements of the hidden Markov model.

Proposition 2. *Suppose a HMM is partially dominated and $L_{\nu,n}(y_{0:n}) > 0$ for $y_{0:n} \in F^{n+1}$. Then, for any $f \in M_b(E^{n+1})$, the following equation holds.*

$$\phi_{\nu,0:n|n}(y_{0:n}, f) = \frac{\int \cdots \int f(x_{0:n}) \nu(dx_0) g(x_0, y_0) \prod_{i=1}^{n} (Q(x_{i-1}, dx_i) g(x_i, y_i))}{L_{\nu,n}(y_{0:n})}$$

Although the proposition only concerns the joint smoothing kernel, it also implicitly handles a case for the marginal smoothing kernel: $\phi_{\nu,k|n}(y_{0:n}, f) \equiv \int \cdots \int f(x_k) \phi_{\nu,0:n|n}(y_{0:n}, dx_{0:k})$ for $k \in \{0\} \cup [n]$.

2.2.2 Forward-Backward decomposition

With a fixed n, we are interested in a **fixed interval smoothing**, a task to compute a marginal smoothing kernel $\phi_{\nu,k|n}$ for any $k \in \{0\} \cup [n]$. We define the **forward kernel** $\alpha_{\nu,k}$ to be the non-negative finite kernel

from $(F^{k+1}, \mathcal{F}^{k+1})$ to $(E, \mathcal{E})$ that satisfies

$$\alpha_{\nu,k}(y_{0:k}, f) = \int \cdots \int f(x_k)\nu(dx_0)g(x_0, y_0)\prod_{i=1}^{k} Q(x_{i-1}, dx_i)g(x_i, y_i)$$

and the **backward function** $\beta_{k|n}$ to be the non-negative measurable function on $F^{n-1} \times E$ defined as

$$\beta_{k|n}(y_{k+1:n}, x) = \int \cdots \int Q(x, dx_{k+1})g(x_{k+1}, y_{k+1}) \prod_{i=k+2}^{n} Q(x_{i-1}, dx_i)g(x_i, y_i)$$

Then, we may apply the proposition 2 (with an assumption that the likelihood is not null) to obtain the **forward – backward decomposition**:

$$\phi_{\nu,k|n}(y_{0:n}, f) = \frac{\int \cdots \int f(x)\nu(dx_0)g(x_0, y_0)\prod_{i=1}^{n}(Q(x_{i-1}, dx_i)g(x_i, y_i))}{L_{\nu,n}(y_{0:n})} = \frac{\int f(x)\alpha_{\nu,k}(y_{0:k}, dx)\beta_{k|n}(y_{k+1:n}, x)}{L_{\nu,n}(y_{0:n})}$$

This will be our main tool for fixed interval smoothing. The use of **forward and backward variables** ($\alpha_{\nu,k|n}$ and $\beta_{k|n}$) was first done by Baum and his collaborators, when they developed the theory of HMM and considered the case where E is finite, in which case $\alpha_{\nu,k|n}(y_{0:k}, x)$ and $\beta_{k|n}(y_{k+1:n}, x)$ can be interpreted (after scaling) as $P(Y_0 = y_0, \cdots, Y_k = y_k, X_k = x)$ and $P(Y_{k+1} = y_{k+1}, \cdots, Y_n = y_n | X_k = x)$ respectively. In general, without assuming the finiteness of the state space, $\alpha_{\nu,k|n}(y_{0:k}, x)$ and $\beta_{k|n}(y_{k+1:n}, x)$ maybe interpreted as the joint density of $Y_{0:k}$ with respect to μ_{k+1} and distribution of X_k, and conditional joint density of $Y_{k+1:n}$ given $X_k = x$ respectively. From now on, we use the notation that often omits y's clarified by the notations section in appendix.

2.2.3 Forward-Backward recursions

An important feature of the forward-backward decomposition is that both forward-backward variables can be written in recursive way, as shown in the proposition below.

Proposition 3. *For $k \in \{0\} \cup [n]$ and $f \in M_b(E)$, the forward measure and the backward function can be written recursively as following.*

1. $$\alpha_{\nu,0}(f) = \int f(x)g_0(x)\nu(dx) \text{ and } \alpha_{\nu,k}(f) = \int f(x') \int \alpha_{\nu,k-1}(dx)Q(x, dx')g_k(x')$$

2. $$\beta_{n|n}(x) = 1 \text{ and } \beta_{k|n}(x) = \int Q(x, dx')g_k(x')\beta_{k+1|n}(x')$$

With the notation for the product of transition kernels defined in the first section, we may write some expressions above in terms of the product of transition kernels defined in the first section. In particular, we have $\alpha_{\nu,k}(f) = \alpha_{\nu,k-1}Q(fg_k)$, $\beta_{k|n} = Q(g_{k+1}\beta_{k+1|n})$, and $\phi_{\nu,k|n}(f) = \frac{\alpha_{\nu,k}(f\beta_{k|n})}{L_{\nu,n}}$. With those in mind, we give a way to compute compute the likelihood functions (observations up to index $k \in \{0\} \cup [n]$) in the proposition below.

Proposition 4. *For $k \in \{0\} \cup [n]$,*

$$\alpha_{\nu,k}(\beta_{k|n}) = L_{\nu,n} \text{ and } \alpha_{\nu,k}(1) = L_{\nu,k}$$

For explicit implementations of the algorithms that uses this recursion for finite space cases, see the algorithm 3.1 and the algorithm 3.2 from [4].

3 Maximum Likelihood Estimator in Hidden Markov Models

Usually, in practice, the model of interest is not fully specified, and some of its parameters can only be guessed using observed data. For complex models, the use of direct estimators such as those using moment or least squares methods may not be sufficient. One of the estimators for unknown parameters that can be used in such case is **Maximum Likelihood Estimator** (MLE), which we will discuss in the section. To describe methods that we will state concisely, we assume that the likelihood function of interest may be written as the marginal of a higher dimensional function, and adopt the terminology from [2], which refers the higher dimensional function as the complete data likelihood; in this setting, the incomplete data refers to the actual observed data and the complete data is a hidden higher dimensional random variable. Furthermore, we will assume that the HMMs discussed here have likelihood functions computable in practice using smoothing recursions discussed previously and are fully dominated. In this section, we define the estimator (MLE) that we consider for unknown parameters, give a method of computing it (EM) when the straightforward computation is not feasible, discuss those in terms of HMM, and then discuss when the estimator might be appropriate to use.

3.1 Maximum Likelihood Estimator

3.1.1 Problem Statement

Given a measure space $(X, \mathcal{X}, \lambda)$ and an open subset $\Theta \subset \mathbb{R}^d$, which is called a **parameter set**, consider a family of functions $\{f(\cdot;\theta)\}$ that are non-negative and integrable with respect to λ and indexed by $\theta \in \Theta$. The task is to find θ that maximize

$$L(\theta) \equiv \int f(x;\theta) d\lambda$$

Often times, it is convenient to instead maximize $\ell(\theta) \equiv \log(L(\theta))$. Note that $p(x;\theta) = \frac{f(x;\theta)}{L(\theta)}$ is can be thought as a probability density with respect to λ. In the context of HMM, we will consider a case where f is the joint probability density function of two X and Y, hidden components and observed components. In the context of the discussion in [2], X is the missing data, f is the complete data likelihood, and L is the likelihood available for estimating θ. In typical example presented in the first courses in statistics or machine learning, the task is simple as $f(\cdot;\theta)$ is simple. However, in general, as $L(\theta)$ involves high dimensional integration, it is not easy to directly evaluate the function. One of the most popular methods to solve this task is the **Expectation Maximization** (EM) algorithm, which will be described below.

3.2 Expectation Maximization

3.2.1 Assumptions

We assume the followings.

1. For any $\theta \in \Theta$, $L(\theta)$ is positive.
2. For any $(\theta, \theta') \in \Theta^2$, $\int |\nabla_\theta \log(p(x;\theta))|(p(x;\theta')d\lambda < \infty$
3. $\{f(\cdot;\theta)\}$ is an exponential family; that is, $f(\cdot;\theta)$ can be written as following.

$$f(\cdot;\theta) = e^{\psi(\theta)^t S(x) - c(\theta)} h(x)$$

   for some vector valued functions of the same dimension S and ψ, real valued function c, and a real valued non-negative function h.

4. $L(\theta)$ is continuously differentiable with respect to θ on Θ.
5. For each fixed θ', $\mathcal{H}(\theta;\theta')$ defined below is continuously differentiable with respect to θ on Θ; in practice, this assumption is somewhat restrictive.

The first assumption guarantees that $p(x;\theta)$ is well-defined for each $\theta \in \Theta$, and the importance of other assumptions may be explained later.

3.2.2 Intermediate quantity of EM

With a convention that $0 \cdot \infty = 0$ to deal with a trivial case for a set on which both $f(x;\theta)$ and $p(x;\theta')$ vanish, the **intermediate quantity** of EM is a family of real valued functions on Θ indexed by $\theta' \in \Theta$ defined as

$$\mathcal{Q}(\theta;\theta') \equiv \int \log(f(x;\theta))p(x;\theta')d\lambda$$

The second assumption guarantees that $p(x;\theta)$ and $p(x;\theta')$ are absolutely continuous with respect to each other, so along with the convention, the quantity is well-defined. The intermediate quantity may be interpreted as $E[\log(f(X;\theta))]$, where the probability density of X is $p(x;\theta')$.

In addition, define the **entropy** of the probability density $p(\cdot;\theta')$ as

$$\mathcal{H}(\theta,\theta') \equiv -\int \log(p(\theta;\theta'))p(x;\theta')d\lambda$$

Then,

$$\mathcal{H}(\theta,\theta') - \mathcal{H}(\theta',\theta') = -\int \log(\frac{p(\theta;\theta')}{p(\theta';\theta')})p(x;\theta')d\lambda$$

is called the **relative entropy** between the $p(\cdot;\theta)$ and $p(\cdot;\theta')$. With a direct application of Jensen's inequality, one can see that this quantity is always is strictly positive, except for the case of equality when $p(\cdot;\theta) =_{a.e} p(\cdot;\theta')$ with respect to λ.

Note that the intermediate quantity may be written in terms of the entropy and the log likelihood.

$$\mathcal{Q}(\theta;\theta') = \ell(\theta) - \mathcal{H}(\theta,\theta')$$

Below is called the fundamental inequality of EM.

Proposition 5. *Under the first three assumptions, the following inequality holds.*

$$\ell(\theta) - \ell(\theta') \geq \mathcal{Q}(\theta;\theta') - \mathcal{Q}(\theta';\theta')$$

Under all assumptions, for any $\theta' \in \Theta$, $\mathcal{Q}(\theta;\theta')$ is continuously differentiable with respect to θ on Θ and

$$\nabla_\theta \ell(\theta') = \nabla_\theta \mathcal{Q}(\theta;\theta') \text{ evaluated at } \theta = \theta'$$

3.2.3 Expectation Maximization algorithm

The EM algorithm, in essence, is mainly described in two steps (E and M steps) as below.

1. Choose θ_0 and set $d \neq 0$ and $i = 0$
2. Repeat below two steps while $d \neq 0$
   (a) E step: determine $\mathcal{Q}(\theta;\theta_i)$
   (b) M step: let θ_{i+1} to be $\arg\max_{\theta\in\Theta} \mathcal{Q}(\theta;\theta_i)$
   (c) Update variables: set $d = \theta_{i+1} - \theta_i$ and $i = i + 1$
3. return θ_i

The algorithm is choosing θ_{i+1} in a way that $\mathcal{Q}(\theta_{i+1};\theta_i) - \mathcal{Q}(\theta_i;\theta_i) \geq 0$. Then, by the fundamental inequality of EM, the algorithm is choosing θ_{i+1} in a way that $\ell(\theta_{i+1})$ is greater than $\ell(\theta_i)$. Thus, the algorithm is a monotone optimization algorithm. Also, suppose that the algorithm converges and returns some θ_*. Then, $\mathcal{Q}(\cdot;\theta_*)$ takes a maximum at θ_*, so $\nabla_\theta \mathcal{Q}(\cdot;\theta_*) = \nabla_\theta L$ takes 0 when $\theta = \theta_*$.

In terms of computation in practice, it is important to have the following conditions for each steps: for E step, we need to compute $\mathcal{Q}(\theta;\theta_i)$ given θ_i in a reasonable computational cost; for M step, we need to find a closed form maximization for $\mathcal{Q}(\theta;\theta_i)$ taking θ as a variable. Here, the assumption regarding exponential family is relevant. With the assumption, $\mathcal{Q}(\theta;\theta')$ can be written as

$$\mathcal{Q}(\theta;\theta') = \varphi(\theta)^t \int S(x)p(x;\theta')d\lambda - c(\theta) + \int p(x;\theta')\log(h(x))d\lambda$$

Since the last term does not vary depending on θ, it is irrelevant and does not need to be computed. Then, the desirable conditions reduce to: for E step, we need to compute expectation of $S(X)$ where S has a density of $p(s;\theta')$ with respect to λ; for M step, we want to have a closed form for a maximization of $\varphi(\theta)^t s - c(\theta)$ as a function of θ for any s in the convex hull of $S(X)$; X here denotes a space, not a random variable. When the condition for the M step is not satisfied, there are ways to get away with it sometimes; see the section 10.5 of [1]. Thus, the main limitation of EM occurs when the condition for E step is satisfied, which is mainly the case when X (the space) is not finite. To see a way to deal with in, see the chapter 11 of [1]. Here, we highlighted some selected important features of the EM algorithm. In order to see more detailed general information about EM, see [2].

3.2.4 Convergence of EM

Above, we established that if the algorithm terminates, it must be that $\nabla_\theta \mathcal{Q}(\cdot;\theta_*) = \nabla_\theta L$ takes 0 when $\theta = \theta_*$. However, there is no guarantee that the θ_i will converge, and even if it does, there is no guarantee that θ_* is not a saddle point of ℓ. It turns out that this convergence issue is hard to verify in practice, and there are more stringent conditions required to guarantee the convergence. In order to see the results regarding convergence, see [6].

3.3 Expectation Maximization in Hidden Markov Models

We assume that observations $Y_{0:n}$ are available for a model whose parameters are not fully known. Also, we consider that ν is known or is fixed completely depending on θ, as we have no means of estimating it. Since we assume that the model is fully dominated, X_0 has a density ν and X_k for $k \in [n]$ has a density $\phi_{k|n}$ with respect to λ given observations $Y_{0:n}$. Also, the joint probability density function of $X_{0:n}$ and $Y_{0:n}$ with respect to $\lambda^{\otimes(n+1)} \otimes \mu^{\otimes(n+1)}$ that depends on unknown θ is given by

$$f_n(x_{0:n}, y_{0:n}; \theta) = \nu(x_0;\theta) g(x_0, y_0; \theta) \prod_{i=1}^{n} q(x_{i-1}, x_i; \theta) f(x_i, y_i; \theta)$$

Then, by plugging f_n in the definition of it and expressing the logarithm of product in the sum of logarithms, the intermediate quantity of EM can be written as

$$\mathcal{Q}(\theta;\theta') = E_{\theta'}[\log(\nu(X_0;\theta))|Y_{0:n}] + \sum_{i=0}^{n} E_{\theta'}[g_k(X_k;\theta))|Y_{0:n}] + \sum_{i=0}^{n-1} E_{\theta'}[q(X_k;\theta))|Y_{0:n}]$$

Then, we can evalutate the intermediate EM quantity by computing expectations, given θ', under $\phi_{k|n}(\cdot,\theta')$ and bivariate smoothing distributions $\phi_{k:k+1}(\cdot;\theta')$, which can be computed using a variant of forward-backward approach discussed previously. For a concrete implementation of the EM algorithm for HMM, see the algorithm 6.1 of [4].

3.4 Consistency

Although we stated that MLE is an estimator for unknown parameters, we never stated why it is a "good" parameter. The "good" property of MLE in HMM will be discussed here.

We first define the concept of consistency. For an unknown parameter θ^*, an estimator $\widehat{\theta}_n$ that depends on the sample size n is called **consistent** if

$$\widehat{\theta}_n \to_p \theta^* \text{ as } n \to \infty$$

For example, LLN gives that the sample mean of iid random variables is a consistent estimator for the mean. Under 6 assumptions specified and justified by the consistency theorem from the page 481 in [3], MLE is a consistent estimator in HMM.

4 Appendix

4.1 Notations

1. $\overline{\mathbb{R}}$ denotes an extended real line and $\mathcal{B}_{\geq 0}$ denotes the Borel σ-algebra on an extended non-negative real line.

2. $\equiv$ symbol may be used to define new things.

3. $\mu^{\otimes n}$ denotes the product measure defined as $\mu^{\otimes 1} = \mu$ and $\mu^{\otimes n} = \mu^{\otimes(n-1)} \otimes \mu$

4. We denote P_ν and E_ν for probability and expectation associated with the model with an initial measure ν. When we write $a.s$, we mean almost surely with respect to P_ν.

5. Given a measurable space $(X, \mathcal{S})$, $M_b(X)$ denote the bounded $\mathcal{S}$-measurable functions. We may often skip specifying measurable spaces in many contexts if it is clear what we are talking about from the context or definitions given in previous parts.

6. We denote $x_n, x_{n+1}, \cdots, x_{m-1}, x_m$ as $x_{n:m}$.

7. Since we focus on given observations $y_{0:n}$, we may omit $y_{0:n}$ or y_k for functions that takes those as inputs. For example, instead of $L_{\nu,n}(y_{0:n})$, we may just write $L_{\nu,n}$; instead of $g(x, y_k)$, we may write $g_k(x)$; the same for $\phi_n(y_{0:n}, \cdot)$ and $\phi_n(\cdot)$, $\alpha_{\nu,k}(y_{0:k}, \cdot)$ and $\alpha_{\nu,k}(\cdot)$, and $\beta_{k|n}(y_{k+1:n}, \cdot)$ and $\beta_{k|n}(\cdot)$. As we slightly change our notation, we may make a slight adjustment in terminologies. For instance, $\alpha_{\nu,k}(\cdot)$ may be called as the forward measure at index k, which refers to $\alpha_{\nu,k}(y_{0:k}, \cdot)$.

8. Given a vector $V \in \mathbb{R}^n$, V^t denotes its transpose.

9. We use $;\theta$ to specify the unknown parameter as θ.

10. We denote $P_{\nu,\theta}$ and $E_{\nu,\theta}$ for probability and expectation associated with the model with an initial measure ν parameterized by θ for models with not fully know parameters.

4.2 Proofs

We state sketches of the proofs for the propositions in the paper.

4.2.1 Proposition 1

For any $h \in M_b(E^p)$,

$$\begin{aligned} E_\nu[\prod_{i=1}^{j} f_i(Y_{k_i}) h(X_{k_1}, \cdots X_{k_j})] &= \int \cdots \int \nu(dx_0) G(x_0, dy_0) (\prod_{i=1}^{k_j} Q(x_{i-1}, dx_i) G(x_i, dy_i)) \\ &\qquad (\prod_{i=1}^{j} f_i(y_{k_i})) h(x_{k_1}, \cdots, x_{k_j}) \\ &= \int \cdots \int \nu(dx_0) G(x_0, dy_0) \prod_{i=1}^{k_j} Q(x_{i-1}, dx_i)) h(x_{k_1}, \cdots, x_{k_j}) \\ &\int \cdots \int (\prod_{i \notin \{k_1, \cdots, k_j\}} G(x_i, dy_i)) (\prod_{i \in \{k_1, \cdots, k_j\}} \int f_i(y_i) G(X_i, dy_i) \end{aligned}$$

Then, since $\int G(x_i, dy_i) = 1$,

$$E_\nu[\prod_{i=1}^{j} f_i(Y_{k_i})h(X_{k_1}, \cdots X_{k_j})] = E_\nu[h(X_{k_1}, \cdots X_{k_j}) \prod_{i=1}^{j} \int f_i(y_i)G(X_i, dy_i)]$$

This is taken from the section 2.2 of [1].

4.2.2 Corollary 1 and 2

1 is immediate from the proposition 1. For 2, see the page 45 of [1].

4.2.3 Proposition 2

Consider the definitions of the quantities involved in the equation.

4.2.4 Proposition 3

First,

$$\alpha_{\nu,k}(f) = \int_{x_k \in E} f(x_k) \int_{x_{k-1} \in E} [\int \cdots \int_{x_0 \in E, \cdots x_{k-1} \in E} \nu(dx_0)g_0(x_0) \prod_{i=1}^{k-1} Q(x_{i-1}, dx_i)g_i(x_i)]Q(x_{k-1}, dx_k)g_k(x_k)$$

Note that the term within the bracket [] is $\alpha_{\nu,k-1}dx_{k-1}$. The similar reasoning holds for the backward function. This is taken from the section 3.2 of [1].

4.2.5 Proposition 4

Recalling the definitions of likelihood function and $\phi_{\nu,k|n}$ and the proposition 2, we have that

$$\phi_{\nu,k|n}(1) = 1$$

On the other hand, by considering the forward-backward decomposition, we have

$$\phi_{\nu,k|n}(1) = \frac{\int 1 \cdot \alpha_{\nu,k}(dx)\beta_{k|n}(x)}{L_{\nu,n}} = \frac{\alpha_{\nu,k}(\beta_{k|n})}{L_{\nu,n}}$$

Thus, it follows that

$$\alpha_{\nu,k}(\beta_{k|n}) = L_{\nu,n}$$

Considering that $\beta_{n|n} = 1$,

$$\alpha_{\nu,n}(1) = L_{\nu,n}$$

Since the forward measure having an index k only depends on the observation upto k, so any index can be a final index, hence, in general,

$$\alpha_{\nu,k}(1) = L_{\nu,k}$$

This is taken from the section 3.2 of [1].

4.2.6 Proposition 5

The justification for the inequality comes from the fact that the difference between the left hand side and the right hand side is the relative entropy between $p(\cdot;\theta)$ and $p(\cdot;\theta')$, which was shown to be non-negative. The justification for the second result comes from the decomposition of $\mathcal{Q}$ and the fact that $\mathcal{H}(\cdot;\theta')$ takes a maximum at θ'; to see the brief discussion that relates to entropy, see the page 350 of [1].

References

[1] Olivier Cappé, Eric Moulines, and Tobias Rydén. *Inference in hidden Markov models*. Springer Science & Business Media, 2006.

[2] Arthur P Dempster, Nan M Laird, and Donald B Rubin. "Maximum likelihood from incomplete data via the EM algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1 (1977), pp. 1–22.

[3] Randal Douc et al. "Consistency of the maximum likelihood estimator for general hidden Markov models". In: *the Annals of Statistics* 39.1 (2011), pp. 474–513.

[4] Ramon van Handel. "Hidden markov models". In: *Unpublished lecture notes* (2008).

[5] Martin Sewell. "Hidden markov models". In: *Department of Computer Science, UCL* (2008).

[6] CF Jeff Wu. "On the convergence properties of the EM algorithm". In: *The Annals of statistics* (1983), pp. 95–103.