Neural Network Reading Group Presenter: Geelon So
Winter 2020, February 28 Scribe: Geelon So

Using SVD to learn HMMs

The paper *A spectral algorithm for learning hidden Markov models* (Hsu, Kakade, Zhang 2007) provides a simple and efficient algorithm to learn **hidden Markov models** (HMMs), which defines probabilities over sequences of hidden states (h_t) and observations (x_t). The approach relies on learning **observable operators** B_x for each observable x and a linear operator C such that:

$$\Pr[x_1, \dots, x_t] = C(B_{x_t} \cdots B_{x_1}).$$

The goal here is to:

- estimate the joint distribution $\Pr[x_1, \dots, x_t]$
- estimate the conditional distribution $\Pr[x_t | x_1, \dots, x_{t-1}]$,

for all sequence lengths t.

A naïve solution to estimating $P(x_1, \dots, x_t) = \Pr[x_1, \dots, x_t]$ up to ε-close total variation distance would be to draw something like n^t/ε^2 samples, which is exponential in the length of the sequence. This of course makes no use of the HMM assumptions. This paper shows that it is possible to estimate P using around $t^2 mn/\varepsilon^2$ samples, where n is the number of possible observables and m is the number of hidden states.

1 Hidden Markov models

A **hidden Markov model** defines a distribution over sequences of hidden states $(h_t)_{t=1}^\infty$ and observations $(x_t)_{t=1}^\infty$ such that:

1. the **hidden states** h_t come from the set $[m]$ and the **observations** x_t come from the set $[n]$
2. the **transition probability matrix** $T_{ij} \in \mathbb{R}^{m \times m}$ and **observation probability matrix** $O_{ij} \in \mathbb{R}^{n \times m}$ are defined as:
$$T_{ij} = \Pr[h_{t+1} = i | h_t = j] \quad \text{and} \quad O_{ij} = \Pr[x_t = i | h_t = j]$$
3. the **initial state distribution** $\boldsymbol{\pi} \in \mathbb{R}^m$ is defined:
$$\pi_i = \Pr[h_1 = i].$$
4. the following **conditional independence** properties are satisfied:
   a. conditioned on the previous hidden state h_{t-1}, the **current hidden state** h_t is sampled independently from all other events in the history
   b. conditioned on the current hidden state h_t, the **current observation** x_t is sampled independently from all other events in the history.

And so, we have the following graphical model:

$$\begin{tikzcd}
\cdots \arrow[r] & h_{t-1} \arrow[r] \arrow[d] & h_t \arrow[r] \arrow[d] & h_{t+1} \arrow[r] \arrow[d] & \cdots \\
& x_{t-1} & x_t & x_{t+1} &
\end{tikzcd}$$

One standard way to represent HMMs is through a graph:

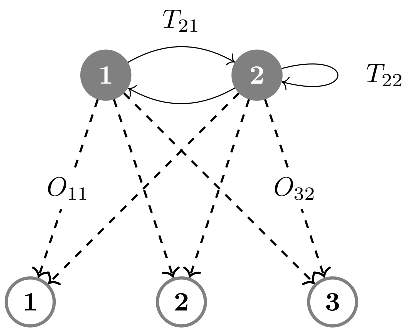

Figure 1: The gray nodes are hidden states and the white nodes are observations. The solid arrows correspond to transition probabilities; for example, T_{21} is the probability of transitioning to state 2 from state 1, $\Pr[h_{t+1} = 2 | h_t = 1]$. The dashed arrows are observation probabilities, where O_{11} is $\Pr[x_t = 1 | h_t = 1]$.

1.1 Observable operator model

If $\boldsymbol{\pi}_t$ is the state distribution at time t, then the state distribution at time $t+1$ is just $\boldsymbol{\pi}_{t+1} = T\boldsymbol{\pi}_t$. We can imagine the $(\boldsymbol{\pi}_t)$ tracing out some trajectory in the probability simplex $\Delta^{m-1} \subset \mathbb{R}^m$. But more generally, T just describes how (probability) mass on the hidden states travel; say we have $\mathbf{w} \in \mathbb{R}^m$ as follows:

$$w_i = \Pr[h_t = i \ \wedge \ \text{event } E \text{ holds}] \quad \Rightarrow \quad (T\mathbf{w})_i = \Pr[h_{t+1} = i \ \wedge \ \text{event } E \text{ holds}].^1$$

Note that $\mathbf{w}$ here does not even need to be normalized.

In particular, for each observation $x \in [m]$, consider the **observation operator** $A_x : \mathbb{R}^m \to \mathbb{R}^m$ defined:

$$A_x = T\text{diag}(O_{x,1}, \ldots, O_{x,m}).$$

Therefore, given a vector $\mathbf{w}$ of probability masses on each hidden state, $A_x\mathbf{w}$ gives the probability masses of observing x then ending up on each of the hidden states in the next time step. If $\mathbf{w}$ were a (normalized) probability distribution over hidden states, then $\mathbf{1}^\top A_x \mathbf{w}$, the L^1-length of $A_x\mathbf{w}$, is the probability of observing x from that configuration of probability mass over hidden states. More generally,

Lemma 1. *For any sequence $(x_1, \ldots, x_t) \in [n]^t$,*

$$\Pr[x_1, \ldots, x_t] = \mathbf{1}^\top A_{x_t} \cdots A_{x_1} \boldsymbol{\pi}.$$

It follows that if we wish to estimate the joint distribution, we might try to estimate these observation operators A_x for $x \in [m]$. The problem of course is that the hidden states are hidden, so we cannot view the action of A_x on the hidden state distribution directly. Our interaction is mediated through the observable operator $O : \mathbb{R}^m \to \mathbb{R}^n$. Assuming that O is full rank, then the hidden state space $\mathbb{R}^m$ is isomorphic to its image $O(\mathbb{R}^m)$ in $\mathbb{R}^n$. This suggests the following high-level sketch:

1. **recover the subspace** range(O) with a set of m linearly independent vectors $U \subset \text{range}(O) \subset \mathbb{R}^n$
2. **construct a representation space** isomorphic to the hidden state space via $U^\top O : \mathbb{R}^m \to \mathbb{R}^m$
3. **estimate the observation operators** on the representation space $B_x = (U^\top O) A_x (U^\top O)^{-1}$

[1] Technically, assume that $E \in \mathcal{F}_t$, where $(\mathcal{F}_\tau)_{\tau=1}^\infty$ is the natural filtration of the stochastic process $(h_\tau, x_\tau)_{\tau=1}^\infty$.

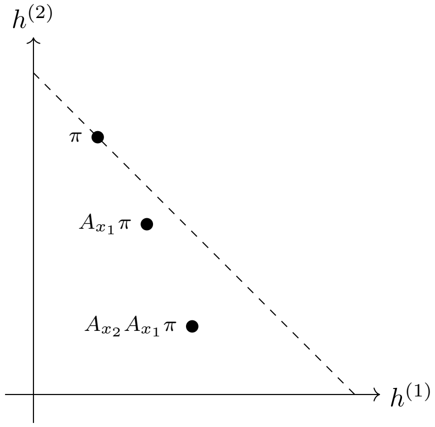

Figure 2: The dashed line is the probability simplex; $\boldsymbol{\pi}$ is the initial state distribution. $A_{x_1}\boldsymbol{\pi}$ describes the probability masses on the hidden states after seeing the observation x_1. Similarly, $A_{x_2}A_{x_1}\boldsymbol{\pi}$ describes the same after seeing the observations (x_1, x_2). The lengths of these vectors is 1, the probability of seeing x_1, and the probability of seeing (x_1, x_2), respectively.

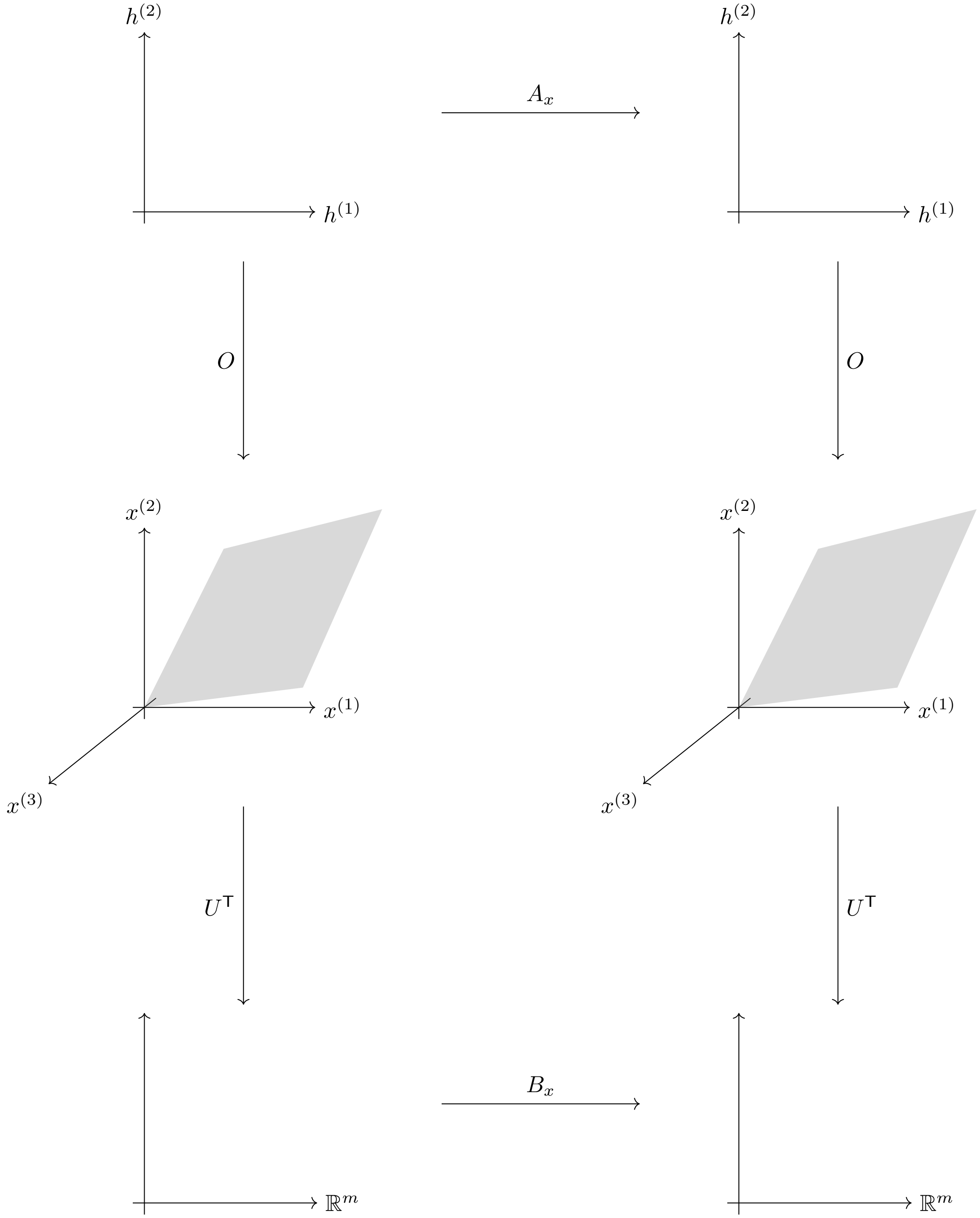

Figure 3: The observable operators A_x act on the hidden state space. If O is full rank, then it maps the hidden state space injectively into the observable space. Performing subspace recovery, say through SVD, allows us to find an isomorphic representation of that linear subspace U and the corresponding B_x on the representation space.

2 Spectral learning of HMMs

The learning model and assumptions:

1. **Learning model**: we obtain samples generated from the HMM from the same initial state distribution.
2. **HMM rank condition**: $\pi > 0$ element-wise, and O and T are full-rank.
3. **Invertibility condition**: $U^\mathsf{T} O$ is invertible, where $U \in \mathbb{R}^{n \times m}$. In other words, U defines m linearly independent vectors in range(O).

Our goal will be to compute the quantity:

$$\Pr[x_1, \ldots, x_t] = \mathbf{1}^\mathsf{T} A_{x_t} \cdots A_{x_1} \boldsymbol{\pi},$$

but as suggested before, we need to access these quantities through the representation $U^\mathsf{T} O : \mathbb{R}^m \to \mathbb{R}^m$, where the domain $\mathbb{R}^m$ is the space in which the hidden state distributions live, and the range $\mathbb{R}^m$ is the representation space. In particular, we can define representations of $\boldsymbol{\pi}$, A_x and $\mathbf{1}$ through:

$$\Pr[x_1, \ldots, x_t] = \underbrace{\left(\mathbf{1}^\mathsf{T} (U^\mathsf{T} O)^{-1}\right)}_{\mathbf{b}_\infty^\mathsf{T}} \underbrace{\left((U^\mathsf{T} O) A_{x_t} (U^\mathsf{T} O)^{-1}\right)}_{B_{x_t}} \cdots \underbrace{\left((U^\mathsf{T} O) A_{x_1} (U^\mathsf{T} O)^{-1}\right)}_{B_{x_1}} \underbrace{\left((U^\mathsf{T} O) \boldsymbol{\pi}\right)}_{\mathbf{b}_1}.$$

In order to produce such a representation, we can look to estimating the following:

1. $P_1 \in \mathbb{R}^n$ the initial observation distribution,
$$[P_1]_i = \Pr[x_1 = i]$$
2. $P_{2,1} \in \mathbb{R}^{n \times n}$ the distribution over (x_2, x_1),
$$[P_{2,1}]_{ij} = \Pr[x_2 = i, x_1 = j]$$
3. $P_{3,x,1} \in \mathbb{R}^{n \times n}$ the conditional distribution over (x_3, x, x_1) for fixed x,
$$[P_{3,x,1}]_{ij} = \Pr[x_3 = i, x_2 = x, x_1 = j].$$

In fact, it is not too hard to algebraically verify:

$$\mathbf{b}_1 = U^\mathsf{T} P_1 \qquad \mathbf{b}_\infty^\mathsf{T} = (P_{2,1} U)^+ P_1 \qquad B_x = \left(U^\mathsf{T} P_{3,x,1}\right) \left(U^\mathsf{T} P_{2,1}\right)^+ \ \forall x \in [n],$$

where X^+ denotes the Moore-Penrose pseudo-inverse of a matrix X.

And to obtain such a U satisfying the above invertibility condition, we can perform SVD on $P_{2,1}$:

Lemma 2 (One choice of U)**.** *Assume $\boldsymbol{\pi} > 0$ and O, T have full rank. Then,* rank$(P_{2,1}) = m$. *Moreover, if U is the matrix of left singular vectors of $P_{2,1}$ corresponding to non-zero singular values, then* range$(U) =$ range(O)*, so $U^\mathsf{T} O$ is invertible.*

Proof. It is straightforward to verify that:

$$P_{2,1} = OT\mathrm{diag}(\boldsymbol{\pi}) O^\mathsf{T},$$

where diag$(\boldsymbol{\pi})$ is the diagonal matrix with π_i as entries. If O and T are full rank and $\boldsymbol{\pi} > 0$, then $P_{2,1}$ is also rank m and has the same range as O. So, if U are the left singular vectors of $P_{2,1}$ with non-zero singular values, then U also has rank m and the same range as O. □

2.1 LearnHMM algorithm

Define the algorithm LEARNHMM(m, N), for m the number of states and N the sample size, to return HMM model parameterized by $\{\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_\infty, \widehat{B}_x \, \forall x \in [n]\}$.

1. Independently sample N observation triples (x_1, x_2, x_3) from HMM to form empirical estimates $\widehat{P}_1$, $\widehat{P}_{2,1}$ and $\widehat{P}_{3,x,1}$ for all $x \in [n]$.
2. Compute SVD of $\widehat{P}_{2,1}$ and let $\widehat{U}$ be the matrix of left singular vectors corresponding to the m largest singular vectors.
3. Compute the model parameters:
   (a) $\widehat{\mathbf{b}}_1 = \widehat{U}^\mathsf{T} \widehat{P}_1$
   (b) $\widehat{\mathbf{b}}_\infty = \left(\widehat{P}_{2,1}^\mathsf{T} \widehat{U}\right)^+ P_1$
   (c) $\widehat{B}_x = \widehat{U}^\mathsf{T} \widehat{P}_{3,x,1} \left(\widehat{U}^\mathsf{T} \widehat{P}_{2,1}\right)^+$ for all $x \in [n]$

2.2 Sample complexity analysis

Theorem 3 (Sample complexity, loose)**.** *Let $0 < \varepsilon, \delta < 1$. Then for all $t \geq 1$, if T and O are full rank and:*

$$N = \Omega_{\sigma_m(O), \sigma_m(P_{2,1})} \left(\frac{t^2}{\varepsilon^2} \cdot mn \cdot \log \frac{1}{\delta} \right),$$

then with probability at least $1 - \delta$, LEARNHMM(m, N) *satisfies:*

$$\|P(X_1, \ldots, X_t) - \widehat{P}(X_1, \ldots, X_t)\|_1 \leq \varepsilon.$$

Here, the $\sigma_m(O)$ and $\sigma_m(P_{2,1})$ are mth largest singular values, and $\Omega_{\sigma_m(O), \sigma_m(P_{2,1})}$ takes them to be constant order. Because we are representing the hidden state space through the map $U^\mathsf{T} O$, we need to make sure that this map is **well-conditioned**—that not only is it invertible, but that it doesn't collapse any probability vectors too much.

As a high-level approach to proving Theorem 3, we take the following steps:

1. Show that if the number of samples N is sufficiently large, then the errors:

   $$\|\widehat{P}_1 - P\|_2, \quad \|\widehat{P}_{2,1} - P_{2,1}\|_2, \quad \|\widehat{P}_{3,x,1} - P_{3,x,1}\|_2 \tag{1}$$

   are small with high probability. This concentration bound makes use of McDiarmid's inequality.
2. Notice that we use $\widehat{P}_{2,1}$ to obtain $\widehat{U}$. We would like range$(\widehat{U}) \approx$ range(O), and in particular, we want $\widehat{U}^\mathsf{T} O$ to be as well-conditioned as $U^\mathsf{T} O$. Using results from matrix perturbation theory, if U and $\widehat{U}$ are close in L^2, then range$(U) \approx$ range$(\widehat{U})$ (in the sense of angles) and the singular values also close.
3. This previous point implies that $\widehat{U}$ is a good choice of map $\widehat{U} : \mathbb{R}^n \to \mathbb{R}^m$. So, while we motivated the choice of U above through Lemma 2 to be the first m left singular vectors of $P_{2,1}$, we could also choose $\widehat{U}$. According to this choice, $\widehat{\mathbf{b}}_1$, $\widehat{\mathbf{b}}_\infty$ and $\widehat{B}_x$ are estimates of:

   $$\widetilde{\mathbf{b}}_1 = \widehat{U}^\mathsf{T} P_1 \qquad \widetilde{\mathbf{b}}_\infty^\mathsf{T} = \left(P_{2,1} \widehat{U}\right)^+ P_1 \qquad \widetilde{B}_x = \left(\widehat{U}^\mathsf{T} P_{3,x,1}\right) \left(\widehat{U}^\mathsf{T} P_{2,1}\right)^+ \, \forall x \in [n],$$

   and the following errors can be bounded in terms of the above estimation errors in Equation 1 and $\sigma_m(O)$, $\sigma_m(P_{2,1})$,

   $$\delta_\infty := \left\| (\widehat{U}^\mathsf{T} O)^\mathsf{T} \widehat{\mathbf{b}}_\infty - \mathbf{1} \right\|_\infty, \quad \Delta_x := \left\| (\widehat{U}^\mathsf{T} O)^{-1} \widehat{B}_x (\widehat{U}^\mathsf{T} O) - A_x \right\|_1, \quad \delta_1 := \left\| (\widehat{U}^\mathsf{T} O)^{-1} \widehat{\mathbf{b}}_1 - \boldsymbol{\pi} \right\|_1.$$

4. We need to understand how error propagates when we multiply out:

$$\widehat{B}_{x_t} \cdots \widehat{B}_{x_1} \widehat{\mathbf{b}}_1.$$

   If $\Delta := \sum_{x \in [n]} \Delta_x$, then one can inductively show:

$$\sum_{(x_1,\ldots,x_t)} \left\| (\widehat{U}^\mathsf{T} O)^{-1} \left(\widehat{B}_{x_t} \cdots \widehat{B}_{x_1} \widehat{\mathbf{b}}_1 - \widetilde{B}_{x_t} \cdots \widetilde{B}_{x_1} \widetilde{\mathbf{b}}_1 \right) \right\|_1 \leq (1+\delta)^t \delta_1 + (1+\Delta)^t - 1,$$

   where the summation is over all possible length-t observations $(x_1, \ldots, x_t) \in [n]^t$. Thus, if $\widehat{\mathbf{b}}_1$ is close enough to $\widetilde{\mathbf{b}}_1$, so δ_1 is small, then when Δ is made small enough, the error is only additive in t. The right hand side is on the order of $\delta_1 + t\Delta$.

5. And finally, we need to bound the error when we estimate the L^1-length of $A_{x_t} \ldots A_{x_1} \boldsymbol{\pi}$ through computing $\widetilde{\mathbf{b}}_\infty^\mathsf{T} \widehat{B}_{x_t} \cdots \widehat{B}_{x_1} \widehat{\mathbf{b}}_1$. One can obtain:

$$\|P - \widehat{P}\|_1 = \sum_{(x_1,\ldots,x_t)} \left| \Pr[x_1, \ldots, x_t] - \widehat{\Pr}[x_1, \ldots, x_t] \right| \leq \delta_\infty + (1+\delta_\infty)\big((1+\delta)^t \delta_1 + (1+\Delta)^t - 1\big),$$

   where P and $\widehat{P}$ are the joint distribution over $(x_1, \ldots, x_t)$ and its estimate by LearnHMM.

Theorem 3 follows from setting N sufficiently large to make δ_1, δ_∞, and Δ small.

This sample complexity can in fact be improved; if n is very large, certain observations may almost never be seen. Let $\mathcal{S}(\varepsilon) = \{S \subset [n] : \Pr[x_2 \in S] \geq 1 - \varepsilon\}$. Then, define $n_0(\varepsilon_0)$ to be the size of the smallest set of observations that contains at least $1 - \varepsilon_0$ of the probability mass:

$$n_0(\varepsilon_0) := \min_{S \in \mathcal{S}(\varepsilon)} |S|.$$

Then, we have the following sample complexity bound:

Theorem 4 (Sample complexity)**.** *There exists $C > 0$ such that for all $0 < \varepsilon, \delta < 1$, for all $t \geq 1$, if T and O are full rank, if $\varepsilon_0 = \sigma_m(O)\sigma_m(P_{2,1}\varepsilon/(4t\sqrt{m})$ and:*

$$N = C \cdot \frac{t^2}{\varepsilon^2} \cdot \left(\frac{m}{\sigma_m(O)^2 \sigma_m(P_{2,1})^4} + \frac{m \cdot n_0(\varepsilon_0)}{\sigma_m(O)^2 \sigma_m(P_{2,1})^2} \right) \cdot \log \frac{1}{\delta},$$

then with probability at least $1 - \delta$, LearnHMM(m, N) *satisfies:*

$$\|P(X_1, \ldots, X_t) - \widehat{P}(X_1, \ldots, X_t)\|_1 \leq \varepsilon.$$

3 Extensions

- This paper also describes a technique to estimate the conditional distribution.
- This technique can be used to learn distributions that are ε-close to HMMs
- Later work shows how to kernelize this algorithm.