EECS 225A Statistical Signal Processing Lecture 9 - 02/18/2020

State Estimation for Hidden Markov Processes

Lecturer: Jiantao Jiao *Scribe: Jiantao Jiao*

1 Markov Triplet

We begin by introducing the Markov triplet, which is useful in the design and analysis of algorithms for state estimation in Hidden Markov Processes.

Let X, Y, Z denote discrete random variables. We say that X, Y, and Z form a Markov triplet (which we denote as $X - Y - Z$) given the following relationships:

$$X - Y - Z \iff p(x, z|y) = p(x|y)p(z|y) \tag{1}$$
$$\iff p(z|x, y) = p(z|y) \tag{2}$$
$$\iff p(x|y, z) = p(x|y) \tag{3}$$

If X, Y, Z form a Markov triplet, their joint distribution enjoys the important property that it can be factored into the product of two functions ϕ_1 and ϕ_2.

Lemma 1. $X - Y - Z \iff \exists\ \phi_1, \phi_2\ s.t.\ p(x, y, z) = \phi_1(x, y)\phi_2(y, z)$.

Proof (Proof of necessity) Assume that $X - Y - Z$. Then by definition of the Markov chain, we have

$$p(x, y, z) = p(x|y, z)p(y, z) = p(x|y)p(y, z).$$

Now taking $\phi_1(x, y) = p(x, y)$ and $\phi_2(y, z) = p(y, z)$ proves this direction.

(Proof of sufficiency) Now, the functions ϕ_1, ϕ_2 exist as mention in the statement. Then we have:

$$\begin{aligned} p(x|y, z) &= \frac{p(x, y, z)}{p(y, z)} \\ &= \frac{p(x, y, z)}{\sum_{\tilde{x}} p(\tilde{x}, y, z)} \\ &= \frac{\phi_1(x, y)\phi_2(y, z)}{\phi_2(y, z)\sum_{\tilde{x}} \phi_1(\tilde{x}, y)} \\ &= \frac{\phi_1(x, y)}{\sum_{\tilde{x}} \phi_1(\tilde{x}, y)}. \end{aligned}$$

The final equality shows that $p(x|y, z)$ is not a function of z. It implies that

$$\begin{aligned} p(x|y) &= \frac{\sum_z p(x, y, z)}{\sum_{x,z} p(x, y, z)} \\ &= \frac{\sum_z p(y, z)p(x|y, z)}{\sum_{x,z} p(y, z)p(x|y, z)} \\ &= \frac{p(y)p(x|y, z)}{\sum_x p(y)p(x|y, z)} \\ &= \frac{p(y)p(x|y, z)}{p(y)} \\ &= p(x|y, z), \end{aligned}$$

showing $X - Y - Z$. Here in the third equality we used the property that $p(x|y, z)$ is not a function of z. □

2 Hidden Markov Processes

Recall the notation x_m^n (with $m \leq n$), which stands for the sequence $(x_m, x_{m+1}, \ldots, x_n)$. For example, $x_1^n = (x_1, x_2, \ldots, x_n)$ is the first n components and $x_t^n = (x_t, x_{t+1}, \ldots, x_n)$ is the last $n-t+1$ components of the n-tuple $(x_1, x_2, \ldots, x_n)$. We often shorten $x^n = x_1^n$. For completeness, if $n \leq 0$, then we think of x^n as the empty set, as in $p(x_1 \mid x^0) = p(x_1)$. We also abbreviate $p_{X,Y}(x, y)$ as $p(x, y)$, $p_{Y|X}(y|x)$ as $p(y|x)$ when its meaning is clear from the context.

Definition 2. *A* Markov process $\{X_n\}_{n\geq 1}$ *is a stochastic process* $X_1, X_2, X_3 \ldots$ *such that for all* $n \geq 2$, X_n *is independent of* X^{n-2} *given* X_{n-1}*, i.e.,* $X_n - X_{n-1} - X^{n-2}$ *is a Markov triplet.*

In other words, a process is Markov if the density of a variable conditioned on previous variables depends only on the immediately preceding variable, as in

$$p(x_n \mid x^{n-1}) = p(x_n \mid x_{n-1}). \tag{4}$$

The property (4) is known as the Markov property. For example, the joint probability density $p(x^n)$ of the first n terms of a Markov process is given by

$$p(x^n) = \prod_{t=1}^{n} p(x_t \mid x^{t-1}) = \prod_{t=1}^{n} p(x_t \mid x_{t-1}), \tag{5}$$

where in the last step we used the Markov property.

Definition 3. *Let* $\{X_n\}_{n\geq 1}$ *be a Markov process.* $\{Y_n\}_{n\geq 1}$ *is a* Hidden Markov Process *if the conditional probability density* $p(y^n \mid x^n)$ *factors as*

$$p(y^n \mid x^n) = \prod_{i=1}^{n} p(y_i \mid x_i). \tag{6}$$

The Hidden Markov Process $\{Y_n\}_{n\geq 1}$ is a noisy observation of an underlying Markov state process $\{X_n\}_{n\geq 1}$ through a "memoryless" noisy channel. We assume that the state transition probabilities $p(x_t \mid x_{t-1})$ in (5) and the channel noise probabilities $p(y_i \mid x_i)$ in (6) are all well defined. The noisy channel is memoryless, because the observation of random variable Y_i depends only on the state variable X_i through the "channel" conditional density $p(y_i \mid x_i)$.

Combining (5) and (6) gives the joint density

$$p(x^n, y^n) = p(x^n)p(y^n \mid x^n) \tag{7}$$

$$= \left(\prod_{t=1}^{n} p(x_t \mid x_{t-1})\right)\left(\prod_{i=1}^{n} p(y_i \mid x_i)\right) \tag{8}$$

3 Undirected graphical models

Let us derive a graphical representation of the joint distribution expressed in (8) as follows:

(a) Create nodes on the graph representing each random variable: $X_1, ..., X_n, Y_1, ..., Y_n$.

(b) For each pairing of nodes, create an edge if there is some factor in (8) which contains both variables.

Example: Figure 1 shows the undirected graphical model for the hidden Markov process in our setting.

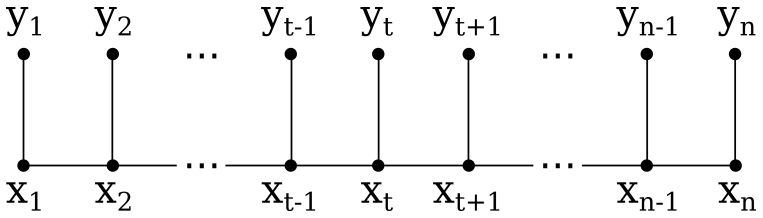

Figure 1: The graph for our hidden Markov process

For any partitioning of the graph into three node sets S_1, S_2, and S_3 such that any path from S_1 to S_3 passes some node in S_2, these sets form a Markov triplet $S_1 - S_2 - S_3$ (for a rigorous proof, see [1, Theorem 7]). Clearly, for any $B_1 \subset S_1, B_3 \subset S_3$, $B_1 - S_2 - B_3$ is also a Markov triplet. Using the graph in Figure 1, we can extract the following additional conditional independence relations.

$$(X^{t-1}, Y^t) - X_t - (X_{t+1}^n, Y_{t+1}^n) \quad \text{(a)}$$
$$(X^{t-1}, Y^{t-1}) - X_t - (X_{t+1}^n, Y_t^n) \quad \text{(b)}$$
$$X^{t-1} - (X_t, Y^{t-1}) - (X_{t+1}^n, Y_t^n) \quad \text{(c)}$$
$$X^{t-1} - (X_t, Y^n) - X_{t+1}^n \quad \text{(d)}$$

It is quite easy to establish principles (a)-(d) via graphs. For example, principles (a), (b), and (d) can be proved via Figure 2, 3, and 4 shown below.

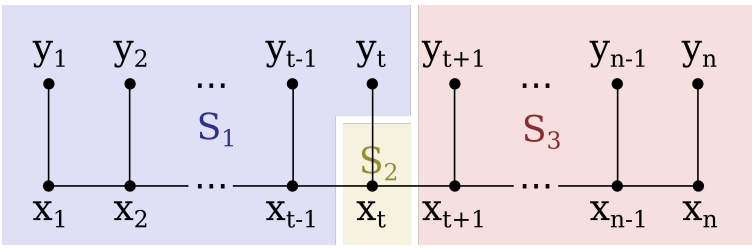

Figure 2: The graph partition for (a)

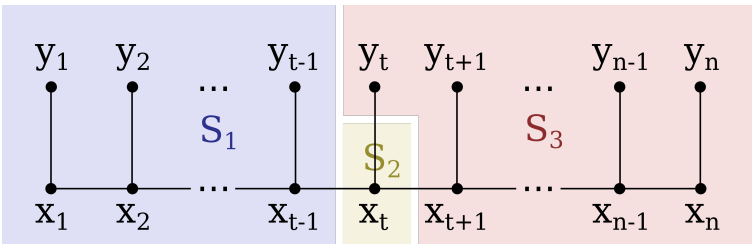

Figure 3: The graph partition for (b)

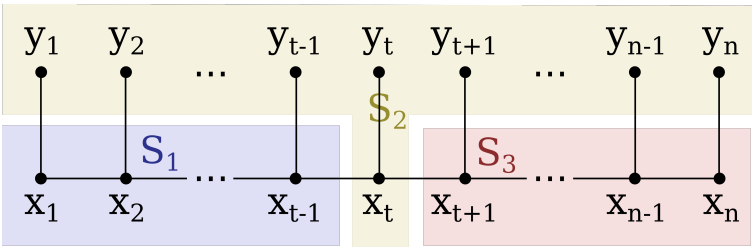

Figure 4: The graphical proof of (d)

4 Inference

Given that the noisy process $\{Y_n\}_{n\geq 1}$ is observed, we would like to estimate $\{X_n\}_{n\geq 1}$. One way to do this is to calculate the posterior distribution $p(x_t \mid y^t)$. The naive approach would rewrite

$$\begin{aligned} p(x_t \mid y^t) &= \frac{p(x_t, y^t)}{p(y^t)} \\ &= \frac{\sum_{x^{t-1}} p(x^t, y^t)}{\sum_{x^t} p(x^t, y^t)} \end{aligned}$$

by marginalizing the joint density $p(x^n, y^n)$ and dividing by the marginal density $p(y^t)$. This summing approach is usually infeasible, however, since each sum above goes over an exponential number of terms. The factorized structure in (8) suggests that we can do much better.

4.1 Causal Inference with Forward Recursion

Indeed the clever factorization which is the source of the HMP graph in Figure 1 is the basis for a recursive inference procedure termed *forward recursion*. Suppose at some index t we know $p(x_t \mid y^{t-1})$ and the noisy channel conditional distribution $p(y_t \mid x_t)$. We can then write

$$\begin{aligned} p(x_t \mid y^t) &= \frac{p(x_t, y_t, y^{t-1})}{\sum_{x_t} p(x_t, y_t, y^{t-1})} \\ &= \frac{p(y_t, y^{t-1} \mid x_t) p(x_t)}{\sum_{x_t} p(y_t, y^{t-1} \mid x_t) p(x_t)} \\ &= \frac{p(y_t \mid x_t) p(y^{t-1} \mid x_t) p(x_t)}{\sum_{x_t} p(y_t \mid x_t) p(y^{t-1} \mid x_t) p(x_t)} \\ &= \frac{p(y_t \mid x_t) p(x_t \mid y^{t-1}) p(y^{t-1})}{\sum_{x_t} p(y_t \mid x_t) p(x_t \mid y^{t-1}) p(y^{t-1})} \\ &= \frac{p(x_t \mid y^{t-1}) p(y_t \mid x_t)}{\sum_{x_t} p(x_t \mid y^{t-1}) p(y_t \mid x_t)}, \end{aligned}$$

where in the third step we made use of the independence assumption $Y^{t-1} - X_t - Y_t$ (property (b)). We make the following two definitions:

- $\beta_t(x_t) = p(x_t \mid y^{t-1})$ posterior on the state x_t given past observations y^{t-1},
- $\alpha_t(x_t) = p(x_t \mid y^t)$ posterior on the state x_t given all observations y^t up to index t.

In this case, we have

$$\alpha_t(x_t) = \frac{\beta_t(x_t)p(y_t \mid x_t)}{\sum_{x_t} \beta_t(x_t)p(y_t \mid x_t)}. \tag{9}$$

Note that $\alpha_t(x_t)$ depends only on $\beta_t(x_t)$ and the channel probabilities $p(y_t \mid x_t)$. Equation (9) is also known as the *measurement update*, because it calculates the posterior probability of the state at the current index t by incorporating the measurement y_t. We can also compute the posterior $\beta_{t+1}(x_{t+1})$ as

$$\begin{aligned}
\beta_{t+1}(x_{t+1}) &= p(x_{t+1} \mid y^t) \\
&= \sum_{x_t} p(x_{t+1}, x_t \mid y^t) \\
&= \sum_{x_t} p(x_t \mid y^t)p(x_{t+1} \mid x_t, y^t) \\
&= \sum_{x_t} p(x_t \mid y^t)p(x_{t+1} \mid x_t),
\end{aligned}$$

where in the last step we used the relation $Y^t - X_t - X_{t+1}$ (property (a)). Hence,

$$\beta_{t+1}(x_{t+1}) = \sum_{x_t} \alpha_t(x_t)p(x_{t+1} \mid x_t). \tag{10}$$

Equation (10) is also known as the *time update*, because it updates the posterior probability by propagating the state forward one time index. Putting (9) and (10) together gives a recursive algorithm for estimating the posterior probabilities $\alpha_t(x_t)$ on the state x_t for each index t. Note that this algorithm is causal in that $\alpha_t(x_t)$ depends only on the posterior and transition probabilities at the previous indices $\tilde{t}$ with $1 \le \tilde{t} \le t$. The forward recursion algorithm is summarized below.

Algorithm: Forward Recursion

Initialize: $\beta_1(x_1) = p(x_1)$ (prior probability on state)

For $t \ge 1$:

1. $\alpha_t(x_t) = \frac{\beta_t(x_t)p(y_t \mid x_t)}{\sum_{x_t} \beta_t(x_t)p(y_t \mid x_t)}$ (measurement update)
2. $\beta_{t+1}(x_{t+1}) = \sum_{x_t} \alpha_t(x_t)p(x_{t+1} \mid x_t)$ (time update)

The forward recursion algorithm is quite general and flexible. The channel and transition probabilities need not be the same index to index.

In fact, the forward recursion also holds when the sequence $\{X_t\}$ is not Markov but a controlled Markov process. Suppose we allow some action A_t depending on Y^t, and the joint distribution of (X^n, Y^n, A^n) is

$$p(x^n, y^n, a^n) = \prod_{i=1}^{n} (p(x_i|x_{i-1}, a_{i-1})p(a_i|y^i)p(y_i|x_i)),$$

where we assume $x_0 = 0, a_0 = 0$. We observe the sequence Y^n and A^n. Then we have the new updates

$$\begin{aligned}
\alpha_t(x_t) &= \frac{\beta_t(x_t)p(y_t \mid x_t)}{\sum_{x_t} \beta_t(x_t)p(y_t \mid x_t)} \\
\beta_{t+1}(x_{t+1}) &= \sum_{x_t} \alpha_t(x_t)p(x_{t+1} \mid x_t, a_t(y^t)).
\end{aligned}$$

References

[1] L. Wasserman, "Undirected graphical models," [Online]. [Online]. Available: \url{http://www.stat.cmu.edu/~larry/=stat700/UG.pdf}