Sample of HMM parameters estimation using Gibbs sampling method

Ryo Ozaki
Ritsumeikan University
Graduate School of Information Science and Engineering
Emergent Systems Laboratory
ryo.ozaki@em.ci.ritsumei.ac.jp

December 11, 2018

1 Graphical model

This section shows the graphical model of hidden Markov model (HMM) which were used this paper.

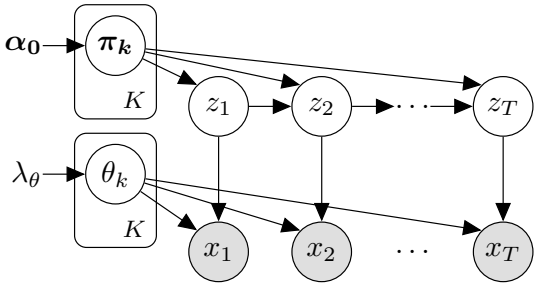

Figure 1: Graphical model of HMM.

The probabilistic generative process of HMM is shown below.

$$\begin{aligned}
\boldsymbol{\pi_0} &\sim \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha_0}) && (1)\\
\boldsymbol{\pi_k} &\sim \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha_0}) && (2)\\
\theta_k &\sim p(\theta|\lambda_\theta) && (3)\\
z_1 &\sim \mathrm{Cat}(z|\boldsymbol{\pi_0}) && (4)\\
z_t &\sim \mathrm{Cat}(z|\boldsymbol{\pi}_{\boldsymbol{z_{t-1}}}) && (5)\\
x_t &\sim p(x|\theta_{z_t}) && (6)\\
\\
k &= 1, 2, \ldots, K && (7)\\
t &= 1, \ldots, T && (8)
\end{aligned}$$

Here, Dir represents the Dirichlet distribution, Cat represents the categorical distribution. And, you can use the any emission distribution $p(x|\theta_t)$ and prior

distribution $p(\theta|\lambda_\theta)$. For example, set $p(x|\theta_t)$ to multivariate normal distribution and set $p(\theta|\lambda_\theta)$ to normal-inverse-Wishart distribution.

2 Posterior distribution

This section shows the posterior distributions.

2.1 Posterior distribution of z_t

When sample the latent variables sequence, basically we use the blocked Gibbs sampling. In this section, we shows the sampling algorithm using blocked Gibbs sampling.

In the blocked Gibbs sampling of HMM, the latent variable sequence $z_{1:T}$ are sampled by the conditional posterior distribution $p(z_{1:T}|x_{1:T}, \boldsymbol{\pi_{0:K}}, \theta_{1:K})$. The distribution is little bit redundancy, so, we don't write the $\boldsymbol{\pi_{0:K}}$ and $\theta_{1:K}$ often in condition part. Here, $p(z_{1:T}|x_{1:T})$ can be described as follows.

$$
\begin{aligned}
p(z_{1:T}|x_{1:T}) &= p(z_1|x_{1:T})p(z_{2:T}|z_1, x_{1:T}) && (9)\\
&= p(z_1|x_{1:T})p(z_2|z_1, x_{1:T})p(z_3|z_1, z_2, x_{1:T}) && (10)\\
&\vdots && (11)\\
&= \prod_{t=1}^{T} p(z_t|z_{1:t-1}, x_{1:T}) && (12)
\end{aligned}
$$

And then, the term $p(z_t = i|z_{1:t-1}, x_{1:T})$ will be calculate as follows.

$$
\begin{aligned}
p(z_t = i|z_{1:t-1}, x_{1:T}) &\overset{z_t}{\propto} p(x_{1:T}|z_{1:t-1}, z_t = i)p(z_t = i|z_{1:t-1}) && (13)\\
&= p(x_{1:t-1}|z_{1:t-1})p(x_{t:T}|z_t = i)p(z_t = i|z_{t-1}) && (14)\\
&\overset{z_t}{\propto} p(x_{t:T}|z_t = i)p(z_t = i|z_{t-1}) && (15)\\
&= p(x_t|z_t = i)p(x_{t+1:T}|z_t = i)p(z_t = i|z_{t-1}) && (16)\\
&= p(x_t|z_t = i)p(z_t = i|z_{t-1})\beta_t(i) && (17)
\end{aligned}
$$

Here, $p(x_{t+1:T}|z_t = i)$ is called "backward message," and it defined as $\beta_t(i)$. And here, latent variables $z_{1:t-1}$ have already been sampled when sampling z_t. Therefore, we can think that latent variables $z_{1:t-1}$ are constant irrespective of z_t. However, you should be care full that the latent variables $z_{t+1:T}$ are still probabilistic variables when sampling z_t.

$$
\beta_t(i) \overset{\text{def}}{=} p(x_{t+1:T}|z_t = i) \tag{18}
$$

And moreover, $\beta_t(i)$ can be calculated recursively.

$$\begin{aligned}
\beta_t(i) &= \sum_j p(x_{t+1:T}, z_{t+1} = j | z_t = i) && (19)\\
&= \sum_j p(x_{t+1:T} | z_{t+1} = j, z_t = i) p(z_{t+1} = j | z_t = i) && (20)\\
&= \sum_j p(x_{t+1:T} | z_{t+1} = j) p(z_{t+1} = j | z_t = i) && (21)\\
&= \sum_j p(x_{t+1} | z_{t+1} = j) p(x_{t+2:T} | z_{t+1} = j) p(z_{t+1} = j | z_t = i) && (22)\\
&= \sum_j p(x_{t+1} | z_{t+1} = j) p(z_{t+1} = j | z_t = i) \beta_{t+1}(j) && (23)
\end{aligned}$$

In addition, the $\beta_{T-1}(i)$ is equals to as follows, and the initial value $\beta_T(j)$ is as follows.

$$\begin{aligned}
\beta_{T-1}(i) &= \sum_j p(x_T | z_T = j) p(z_T = j | z_{T-1} = i) && (24)\\
&= \sum_j p(x_T | z_T = j) p(z_T = j | z_{T-1} = i) \beta_T(j) && (25)\\
\beta_T(j) &= 1 && (26)
\end{aligned}$$

Summarize, the posterior distribution of latent variable sequence is as follows.

$$\begin{aligned}
p(z_{1:T} | x_{1:T}) &= \prod_{t=1}^{T} p(z_t | z_{1:t-1}, x_{1:T}) && (27)\\
p(z_t = i | z_{1:t-1}, x_{1:T}) &\overset{z_t}{\propto} p(x_t | z_t = i) p(z_t = i | z_{t-1}) \beta_t(i) && (28)\\
\beta_t(i) &= \sum_j p(x_{t+1} | z_{t+1} = j) p(z_{t+1} = j | z_t = i) \beta_{t+1}(j) && (29)\\
\beta_T(i) &= 1 && (30)
\end{aligned}$$

Therefore, the posterior distribution of z_t is as follows.

$$\begin{aligned}
p(z_t = i | x_{1:T}, z_{1:t-1}, \boldsymbol{\pi_{0:K}}, \theta_{1:K}) &\overset{z_t}{\propto}\\
& p(x_t | z_t = i, \theta_{1:K}) p(z_t = i | z_{t-1}, \boldsymbol{\pi_{0:K}}) \beta_t(i) && (31)
\end{aligned}$$

In addition, the array which the value of posterior distribution arranged from

$k=1$ to $k=K$ is as follows.

$$\begin{aligned}
\begin{bmatrix} p(z_t=1|x_{1:T},z_{1:t-1}) \\ p(z_t=2|x_{1:T},z_{1:t-1}) \\ \vdots \\ p(z_t=i|x_{1:T},z_{1:t-1}) \end{bmatrix} &= \eta \cdot \begin{bmatrix} p(x_t|z_t=1)p(z_t=1|z_{t-1})\beta_t(1) \\ \vdots \\ p(x_t|z_t=i)p(z_t=i|z_{t-1})\beta_t(i) \\ \vdots \\ p(x_t|z_t=K)p(z_t=i|z_{t-1})\beta_t(K) \end{bmatrix} \\
&\overset{z_t}{\propto} \begin{bmatrix} p(x_t|z_t=1)p(z_t=1|z_{t-1})\beta_t(1) \\ \vdots \\ p(x_t|z_t=i)p(z_t=i|z_{t-1})\beta_t(i) \\ \vdots \\ p(x_t|z_t=K)p(z_t=i|z_{t-1})\beta_t(K) \end{bmatrix}
\end{aligned} \tag{32}$$

Here, η is defined as the constant term, and it will be able to calculate the value using the restoration.

$$\eta \cdot \sum_{k=1}^{K} \{p(x_t|z_t=i)p(z_t=i|z_{t-1})\beta_t(i)\} = 1 \tag{33}$$

$$\eta = \frac{1}{\sum_{k=1}^{K} \{p(x_t|z_t=i)p(z_t=i|z_{t-1})\beta_t(i)\}} \tag{34}$$

2.2 Posterior distribution of $\boldsymbol{\pi_0}$

The posterior distribution of $\boldsymbol{\pi_0}$ is represented as follows.

$$\begin{aligned}
p(\boldsymbol{\pi_0}|\boldsymbol{\alpha_0}, \{z^s_{1:T_s}\}_{s=1,\dots,S}) &= p(\boldsymbol{\pi_0}|\boldsymbol{\alpha_0}, \{z^s_1\}_{s=1,\dots,S}) && (35)\\
&\overset{\boldsymbol{\pi_0}}{\propto} p(\{z^s_1\}_{s=1,\dots,S}|\boldsymbol{\pi_0})p(\boldsymbol{\pi_0}|\boldsymbol{\alpha_0}) && (36)\\
&= \prod_{s=1}^{S} \{p(z^s_1|\boldsymbol{\pi_0})\}p(\boldsymbol{\pi_0}|\boldsymbol{\alpha_0}) && (37)\\
&= \prod_{s=1}^{S} \{\mathrm{Cat}(z^s_1|\boldsymbol{\pi_0})\}\,\mathrm{Dir}(\boldsymbol{\pi_0}|\boldsymbol{\alpha_0}) && (38)\\
&= \mathrm{Mult}(\boldsymbol{m}|\boldsymbol{\pi_0})\,\mathrm{Dir}(\boldsymbol{\pi_0}|\boldsymbol{\alpha_0}) && (39)\\
&= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha_0^*}) && (40)
\end{aligned}$$

$$m_k = \sum_{s=1}^{S} \delta(z^s_1 = i) \tag{41}$$

$$\delta(\mathrm{CONDITION}) = \begin{cases} 0 & (\mathrm{CONDITION\ is\ false}) \\ 1 & (\mathrm{CONDITION\ is\ true}) \end{cases} \tag{42}$$

$$\boldsymbol{\alpha_0^*} = \boldsymbol{m} + \boldsymbol{\alpha_0} \tag{43}$$

Here, S represents the number of sequence of z, and $z^s_{1:T_s}$ represents the s-th sequence. Mult represents multinomial distribution.

2.3 Posterior distribution of $\boldsymbol{\pi_k}$

The posterior distribution of $\boldsymbol{\pi_k}$ is represented as follows.

$$
\begin{aligned}
Z_k &= \left\{ z_t^s \;\middle|\; z_{t-1}^s = k, s = 1, \ldots, S, t = 2, \ldots, T_s \right\} && (44)\\
p(\boldsymbol{\pi_k}|\boldsymbol{\alpha_0}, \{z^s_{1:T_s}\}_{s=1,\ldots,S}) &= p(\boldsymbol{\pi_k}|\boldsymbol{\alpha_0}, Z_k) && (45)\\
&\overset{\boldsymbol{\pi_k}}{\propto} p(Z_k|\boldsymbol{\pi_k}) p(\boldsymbol{\pi_k}|\boldsymbol{\alpha_0}) && (46)\\
&= \prod_{z_i \in Z_k} \{p(z_i|\boldsymbol{\pi_k})\} p(\boldsymbol{\pi_k}|\boldsymbol{\alpha_0}) && (47)\\
&= \prod_{z_i \in Z_k} \{\mathrm{Cat}(z_i|\boldsymbol{\pi_k})\} \,\mathrm{Dir}(\boldsymbol{\pi_k}|\boldsymbol{\alpha_0}) && (48)\\
&= \mathrm{Mult}(\boldsymbol{m}|\boldsymbol{\pi_k}) \,\mathrm{Dir}(\boldsymbol{\pi_k}|\boldsymbol{\alpha_0}) && (49)\\
&= \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha_k^*}) && (50)
\end{aligned}
$$

$$
\begin{aligned}
m_j &= \sum_{z_i \in Z_k} \delta(z_i = j) && (51)\\
\delta(\text{CONDITION}) &= \begin{cases} 0 & (\text{CONDITION is false}) \\ 1 & (\text{CONDITION is true}) \end{cases} && (52)
\end{aligned}
$$

$$
\boldsymbol{\alpha_k^*} = \boldsymbol{m} + \boldsymbol{\alpha_0} \quad (53)
$$

Here, Mult represents multinomial distribution.

2.4 Posterior distribution of θ_k

The posterior distribution of θ_k is represented as same as GMM. So, I do not introduce in here. Please see GMM.

Appendix

A Blocked Gibbs sampling

In Gibbs sampling, all parameters are sampled by the conditional distribution each other. For example, the parameters are $\{ \theta_1, \theta_2, \theta_3 \}$, we can get the values which sampled by joint distribution $p(\theta_1, \theta_2, \theta_3)$ using each conditional distribution $p(\theta_1|\theta_2, \theta_3)$, $p(\theta_2|\theta_1, \theta_3)$, and $p(\theta_3|\theta_1, \theta_2)$.

1: initialize the parameters θ_1, θ_2, and θ_3.
2: **while** burn-in loops **do**
3: $\theta_1 \sim p(\theta_1|\theta_2, \theta_3)$
4: $\theta_2 \sim p(\theta_2|\theta_1, \theta_3)$
5: $\theta_3 \sim p(\theta_3|\theta_1, \theta_2)$
6: **end while**
7: $(\theta_1, \theta_2, \theta_3)$ are become to values which sampled by $p(\theta_1, \theta_2, \theta_3)$.

Then, in the blocked Gibbs sampling, some grouped parameters are sampled by conditional joint distribution. For example, the θ_1 and θ_2 are grouped, the procedure of Gibbs sampling are as follows.

1: initialize the parameters θ_1, θ_2, and θ_3.
2: **while** burn-in loops **do**
3: $(\theta_1, \theta_2) \sim p(\theta_1, \theta_2|\theta_3)$
4: $\theta_3 \sim p(\theta_3|\theta_1, \theta_2)$
5: **end while**
6: $(\theta_1, \theta_2, \theta_3)$ are become to values which sampled by $p(\theta_1, \theta_2, \theta_3)$.

In addition, the conditional joint distribution $p(\theta_1, \theta_2|\theta_3)$ are divided to as follows:

$$p(\theta_1, \theta_2|\theta_3) = p(\theta_1|\theta_3)p(\theta_2|\theta_1, \theta_3). \tag{54}$$

And, the marginal distribution $p(\theta_1|\theta_3)$ are calculated by

$$p(\theta_1|\theta_3) = \int p(\theta_1, \theta_2|\theta_3)d\theta_2 \tag{55}$$

B Scaling the backward message

In the Gibbs sampling of HMM, we calculated the backward message as follows.

$$\beta_t(i) = \sum_j p(x_{t+1}|z_{t+1} = j)p(z_{t+1} = j|z_t = i)\beta_{t+1}(j) \tag{56}$$

$$\beta_T(i) = 1 \tag{57}$$

And the backward message was used to sampling the hidden state sequence as follows.

$$p(z_t = i|z_{1:t-1}, x_{1:T}) \overset{z_t}{\propto} p(x_t|z_t = i)p(z_t = i|z_{t-1})\beta_t(i) \tag{58}$$

Here, the backward message will be becoming to very small values, because, the backward message was calculated by many products of probability which value is less than 1. The case of the sequence length is very long, the value of the

backward message might not be able to calculate because of underflow. So, we need to scaling the backward message. In this section, I will be introducing the 2 way to scaling backward message.

B.1 Log scaling

Using log scale is the very popular way to keep the small value without underflow as the classical method. After taking the log scale of backward message is as follows.

$$\begin{aligned} E_t(j) &= \log p(x_t|z_t = j) && (59)\\ A_t(i,j) &= \log p(z_t = j|z_{t-1} = i) && (60)\\ B_t(i) &= \log \beta_t(i) && (61)\end{aligned}$$

$$\begin{aligned} B_t(i) &= \log\left[\sum_j \exp\left\{E_{t+1}(j) + A_{t+1}(i,j) + B_{t+1}(i)\right\}\right] && (62)\\ B_T(i) &= 0 && (63)\end{aligned}$$

In addition, the format of calculating formula $B(z_t)$ is called "LogSumExp." The LogSumExp is calculated by as follows.

$$\begin{aligned} y &= \log\left\{\exp(x_1) + \ldots + \exp(x_n)\right\} && (64)\\ &= \log\left\{\exp(x^*)\left(\exp(x_1 - x^*) + \ldots + \exp(x_n - x^*)\right)\right\} && (65)\\ &= x^* + \log\left\{\exp(x_1 - x^*) + \ldots + \exp(x_n - x^*)\right\} && (66)\\ x^* &= \max\{x_1, \ldots, x_n\} && (67)\end{aligned}$$

By this calculating, the risk of underflow is low, because, the $|x_i - x^*|$ is small than $|x_i|$. In addition, it needs to take exp when sampling the hidden state z_t, but, it needs only the information about proportional. Therefore, the posterior distribution of the hidden state z_t is represented as follows.

$$\begin{aligned} p(z_t = i|z_{1:t-1}, x_{1:T}) &\overset{z_t}{\propto} p(x_t|z_t = i)p(z_t = i|z_{t-1})\exp(B_t(i)) && (68)\\ &\overset{z_t}{\propto} p(x_t|z_t = i)p(z_t = i|z_{t-1})\exp(B_t(i) - B_{MAX}) && (69)\end{aligned}$$

$$B_{MAX} = \max\{B_t(1), B_t(2), \ldots, B_t(K)\} \quad (70)$$

B.2 Normalizing

In normalizing, calculating the proportional value of a backward message as follows.

$$\begin{aligned} \hat{\beta}_t(i) &= \frac{1}{c_t}\beta_t(i) && (71)\\ &\overset{z_t}{\propto} \beta_t(i) && (72)\end{aligned}$$

Here, c_t is a normalizing constant. The c_t is calculated by some restriction. In this section, we set the restriction to the max value of the $\hat{\beta}_t(i)$ is 1.

$$c_t = \max\{\beta_t(1), \beta_t(2), \ldots, \beta_t(K)\} \quad (73)$$

After this representation, the backward message is as follows.

$$\hat{\beta}_t(i) = \frac{c_{t+1}}{c_t} \sum_j p(x_{t+1}|z_{t+1}=j) p(z_{t+1}=j|z_t=i) \hat{\beta}_{t+1}(j) \tag{74}$$

$$\hat{\beta}_T(i) = 1 \tag{75}$$

Here, we need to calculate the normalization term $\frac{c_{t+1}}{c_t}$. And the normalization term is normalizing the max value to 1. Therefore, the value of $\frac{c_{t+1}}{c_t}$ is as follows.

$$\hat{\beta}_t^-(i) = \sum_j p(x_{t+1}|z_{t+1}=j) p(z_{t+1}=j|z_t=i) \hat{\beta}_{t+1}(j) \tag{76}$$

$$\left(\frac{c_{t+1}}{c_t}\right)^{-1} = \max\left\{\hat{\beta}_t^-(1), \hat{\beta}_t^-(2), \ldots, \hat{\beta}_t^-(K)\right\} \tag{77}$$

Finally, the posterior distribution of z_t is as follows.

$$p(z_t=i|z_{1:t-1}, x_{1:T}) \overset{z_t}{\propto} p(x_t|z_t=i) p(z_t=i|z_{t-1}) \beta_t(i) \tag{78}$$

$$\overset{z_t}{\propto} p(x_t|z_t) p(z_t|z_{t-1}) \hat{\beta}_t(i) \tag{79}$$

If you want to use another restriction, you have to care the initial value of backward message. E.g.,

$$c_t = \sum_{k=1}^{K} \beta_t(k) \tag{80}$$

$$\hat{\beta}_T(i) = \frac{1}{K} \tag{81}$$

$$\left(\frac{c_{t+1}}{c_t}\right)^{-1} = \sum_{k=1}^{K} \hat{\beta}_t^-(k) \tag{82}$$