CHAPTER 5

Discriminative Learning Algorithm for HMM

In this chapter, we extend the growth transformation (GT)-based approach to the discriminative parameter estimation problem from the stationary probability model characterized by the exponential-family distributions discussed in Chapter 4 to the nonstationary probability model. The nonstationarity discussed here is characterized by the Markov chain that underlies the hidden Markov model (HMM), where the emission probabilities in the HMM can be represented by any member in the exponential-family distributions as well as by the mixture Gaussian distribution. Specifically, in the algorithm derivation discussed in this chapter, we only use the example for the Gaussian HMM. It can be easily generalized to other types of continuous-density HMMs (CDHMMs) by using some of the derivation steps discussed in Chapter 4. To make the coverage of HMM complete, we first discuss the discriminative parameter estimation problem in classifier design where each class is characterized by a discrete HMM.

5.1 ESTIMATION FORMULAS FOR DISCRETE HMM

In this section, we derive the GT estimation formulas for the discrete HMM's parameters — $\Lambda = \{\{a_{i,j}\}, \{b_i(k)\}\}$ for transition probabilities and emitting probabilities. The formula "grows" the unified discriminative training criterion $O(\Lambda)$. In the next section, we will present the derivation for the CDHMM. In both cases, $O(\Lambda)$ is difficult to optimize directly but because it is a rational function as expressed in (3.2), we can construct the auxiliary functions of (1) F and then (2) V based on F. Optimizing $V(\Lambda; \Lambda')$ becomes a relatively easier problem and it leads to the final GT formulas for all types of discriminative criteria unified by (3.2).

For the discrete HMM, $X = X_1, \ldots, X_R$ is used to denote observation sequences with discrete indexes. That is, at time t for token r, the observation $x_{r,t}$ is an index belonging to the set $[1, 2, \ldots, K]$, where K is the size of the index set.

5.1.1 Constructing Auxiliary Function $F(\Lambda;\Lambda')$

Starting from (4.11), we have the generic auxiliary function of

$$F(\Lambda;\Lambda') = \sum_s p\,(X,s|\Lambda)\,\left[C(s) - O(\Lambda')\right] + D$$

which, for an HMM, becomes

$$F(\Lambda;\Lambda') \;=\; \sum_s\sum_q p\,\left(X,q,s|\Lambda\right)\,\left[C(s) - O(\Lambda')\right] + D \qquad (5.1)$$

where q is the HMM state sequence, and $s = s_1, \ldots, s_R$ is the label (e.g., word or phone) sequence for all R training tokens (including correct or incorrect sentences). The same interpretation as in Chapter 4 can be given to the main terms in the auxiliary function $F(\Lambda;\Lambda')$ above as the average deviation of the accuracy count.

Because $p(s)$ depends on the language model, it is irrelevant for optimizing Λ. We therefore can have the following decomposition: $p(X,q,s|\Lambda) = p(s)\cdot p(X, q|s, \Lambda)$. Hence,

$$\begin{aligned} F(\Lambda;\Lambda') &= \sum_s\sum_q [C(s) - O(\Lambda')]p(s)p(X,q|s,\Lambda) + D \\ &= \sum_s\sum_q\sum_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right]p(\chi,q|s,\Lambda) \qquad (5.2) \end{aligned}$$

where

$$\Gamma(\Lambda') = \delta(\chi,X)p(s)\left[C(s) - O(\Lambda')\right] \qquad (5.3)$$

As before, $D = \sum_s \mathrm{d}(s)$ is a quantity independent of the parameter set Λ. In (5.3), $\delta(\chi, X)$ is the Kronecker delta function, in which χ represents the entire data space where X is in. The summation over this data space is introduced here again for accommodating the parameter-independent constant D, that is, $\sum_s\sum_g\sum_\chi \mathrm{d}(s)p(\chi, q \mid s, \Lambda) = \sum_s \mathrm{d}(s) = D$ is a Λ-independent constant (an idea originally proposed in [17] for HMM).

5.1.2. Constructing Auxiliary Function $V(\Lambda;\Lambda')$

We now desire to construct the new auxiliary function $V(\Lambda; \Lambda')$ using (5.2), in the same way as we did in Chapter 4. We first identify

$$f\,(\chi,q,s,\Lambda) = \left[\Gamma(\Lambda') + \mathrm{d}(s)\right]p(\chi,q|s,\Lambda)$$

as before. Again, to ensure that $f\,(\chi, q, s, \Lambda)$ above is positive, $\mathrm{d}(s)$ should be selected to be sufficiently large so that $\Gamma(\Lambda') + \mathrm{d}(s) > 0$ (note $p\,(\chi, q\,|\,s, \Lambda)$ in (5.2) is nonnegative). Then, using (4.4) again, we have

$$\begin{aligned} V(\Lambda;\Lambda') &= \sum_{q}\sum_{s}\sum_{\chi}\left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p\left(\chi,q|s,\Lambda'\right)\ \log\left\{\underbrace{\left[\Gamma(\Lambda') + \mathrm{d}(s)\right]}_{\text{optimization - indept}} p\left(\chi,q|s,\Lambda\right)\right\} \\ &= \sum_{q}\sum_{s}\sum_{\chi}\left[\Gamma(\Lambda') + d(s)\right] p(\chi,q|s,\Lambda')\log\ \left(\chi,q|s,\Lambda\right) + \text{Const.} \\ &= \sum_{q}\sum_{s} p(X,q,s|\Lambda')\left(C(s) - O(\Lambda')\right)\log p\left(X,q|s,\Lambda\right) \\ &\quad + \sum_{q}\sum_{s}\sum_{\chi}\mathrm{d}(s) p\left(\chi,q|s,\Lambda'\right)\ \log p\left(\chi,q|s,\Lambda\right) + \text{Const.} \end{aligned} \tag{5.4}$$

5.1.3 Simplifying Auxiliary Function $V(\Lambda;\Lambda')$

After ignoring optimization-independent constant in (5.4), we divide $V(\Lambda;\Lambda')$ by $p(X|\Lambda')$ to convert the joint probability $p(X, q, s\,|\,\Lambda')$ to the posterior probability $p(q, s\,|\,X, \Lambda') = p(s\,|\,X, \Lambda')p(q\,|\,X, \Lambda')$. The equivalent auxiliary function then becomes

$$\begin{aligned} U(\Lambda;\Lambda') =& \sum_{q}\sum_{s} p\left(s|X,\Lambda'\right) p\left(q|X,s,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\log p\left(X,q|s,\Lambda\right) \\ &+ \sum_{q}\sum_{s}\sum_{\chi}\mathrm{d}'(s) p\left(\chi,q|s,\Lambda'\right)\log p\left(\chi,q|s,\Lambda\right) \end{aligned} \tag{5.5}$$

where

$$\mathrm{d}'(s) = \mathrm{d}(s)/p\left(X|\Lambda'\right) \tag{5.6}$$

Because X depends only on the HMM state sequence q, we have $p(X, q\,|\,s, \Lambda) = p(q\,|\,s, \Lambda)\,.\,p(X|\,q, \Lambda)$. Therefore, $U(\Lambda;\Lambda')$ can be further decomposed to four terms below:

$$\begin{aligned} U(\Lambda;\Lambda') =& \overbrace{\sum_{q}\sum_{s} p\left(s|X,\Lambda'\right) p\left(q|X,s,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\ \log p\left(X|q,\Lambda\right)}^{\text{term - I}} \\ &\underbrace{+ \sum_{q}\sum_{s}\sum_{\chi}\mathrm{d}'(s) p\left(\chi,q|s,\Lambda'\right)\ \log p\left(\chi|q,\Lambda\right)}_{\text{term - II}} \\ &+ \overbrace{\sum_{q}\sum_{s} p\left(s|X,\Lambda'\right) p\left(q|X,s,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\ \log p\left(q|s,\Lambda\right)}^{\text{term - III}} \\ &\underbrace{+ \sum_{q}\sum_{s}\sum_{\chi}\mathrm{d}'(s) p\left(\chi,q|s,\Lambda'\right)\ \log p\left(q|s,\Lambda\right)}_{\text{term - IV}} \end{aligned} \tag{5.7}$$

Note the two new terms above compared with the corresponding auxiliary function (4.14) in Chapter 4. Here, $X = X_1, \ldots, X_R$ is a large aggregate of all training data with R independent sentence tokens. For each token $X_r = x_{r,1}, \ldots, x_{r,Tr}$, the observation vector $x_{r,t}$ is independent of each other and it depends only on the HMM state at time t. Hence, $\log p(X|q, \Lambda)$ can be decomposed, enabling simplification of both term-I and term-II in (5.7). To simplify term-III and term-IV in (5.7), we decompose $\log p(q\,|s, \Lambda)$ based on the property of the first-order HMM — the state at time t depends only on state at time $t-1$. We now elaborate on the simplification of all four terms in (5.7).

For term-I, we first define

$$\gamma_{i,r,s_r}(t) \triangleq \sum_{q, q_{r,t}=i} p\left(q|X,s,\Lambda'\right) = p\left(q_{r,t}=i|X,s,\Lambda'\right) = p\left(q_{r,t}=i|X_r,s_r,\Lambda'\right) \qquad (5.8)$$

The last equality comes from the fact that the sentence tokens in the training set are independent of each other. $\gamma_{i,r,s_r}(t)$ is the occupation probability of state i at time t, given the label sequence s_r and observation sequence X_r, and an efficient forward–backward algorithm exists to compute it [43]. Using the definition of (5.8) and assuming that the HMM state index is from 1 to I, we have

$$\text{term - I} = \sum_s p\left(s|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_q p\left(q|X,s,\Lambda'\right)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\log p\left(x_{r,t}\middle|q_{r,t},\Lambda\right)$$

$$= \sum_s p\left(s|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\sum_{q, q_{r,t}=i} p\left(q|X,s,\Lambda'\right)\log p\left(x_{r,t}\middle|q_{r,t}=i,\Lambda\right)$$

$$= \sum_s p\left(s|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\gamma_{i,r,s_r}(t)\log p\left(x_{r,t}\middle|q_{r,t}=i,\Lambda\right) \qquad (5.9)$$

The simplification process for the second term in (5.7) is as follows. Using the notations $\tilde{q} = q_{1,1}, \ldots, q_{r,t-1}, q_{r,t+1}, \ldots, q_{R,T_R}$, $\tilde{\chi} = \chi_{1,1}, \ldots, \chi_{r,t-1}, \chi_{r,t+1}, \ldots, \chi_{R,T_R}$, we have

$$
\begin{aligned}
\text{term - II} &= \sum_{s} \mathrm{d}'(s) \sum_{q_{1,1},\ldots,q_{R,T_R}} \sum_{\chi_{1,1},\ldots,\chi_{R,T_R}} p\left(\chi_{1,1},\ldots,\chi_{R,T_R}, q_{1,1},\ldots,q_{R,T_R} \middle| s,\Lambda'\right) \sum_{r=1}^{R} \sum_{t=1}^{T_r} \log p\left(\chi_{r,t} \middle| q_{r,t},\Lambda\right) \\
&= \sum_{s} \mathrm{d}'(s) \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{q_{r,t}} \sum_{\chi_{r,t}} p\left(\chi_{r,t}, q_{r,t} \middle| s,\Lambda'\right) \underbrace{\sum_{\tilde{q}} \sum_{\tilde{\chi}} p(\tilde{\chi},\tilde{q} | \chi_{r,t}, q_{r,t}, s, \Lambda')}_{=1} \log p\left(\chi_{r,t} \middle| q_{r,t},\Lambda\right) \\
&= \sum_{s} \mathrm{d}'(s) \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{q_{r,t}} \sum_{\chi_{r,t}} p\left(\chi_{r,t}, q_{r,t} \middle| s,\Lambda'\right) \log p\left(\chi_{r,t} \middle| q_{r,t},\Lambda\right) \\
&= \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{i=1}^{I} \sum_{\chi_{r,t}} \sum_{s} d'(s) p(q_{r,t} = i | s,\Lambda') p(\chi_{r,t} | q_{r,t} = i,\Lambda') \log\ p(\chi_{r,t} | q_{r,t} = i,\Lambda) \\
&= \sum_{r=1}^{R} \sum_{t=1}^{T_r} \sum_{i=1}^{I} \mathrm{d}(r,t,i) \sum_{\chi_{r,t}} p\left(\chi_{r,t} \middle| q_{r,t} = i;\Lambda'\right) \log p\left(\chi_{r,t} \middle| q_{r,t} = i;\Lambda\right)
\end{aligned}
\qquad (5.10)
$$

where

$$
\mathrm{d}(r,t,i) = \sum_{s} \mathrm{d}'(s) p\left(q_{r,t} = i \middle| s,\Lambda'\right) \qquad (5.11)
$$

To simplify term-III in (5.7), we first define

$$
\begin{aligned}
\xi_{i,j,r,s_r}(t) \triangleq \sum_{q : q_{r,t-1}=i, q_{r,t}=\mathrm{j}} p(q | X,s,\Lambda') &= p\left(q_{r,t-1} = i, q_{r,t} = \mathrm{j} \middle| X,s,\Lambda'\right) \\
&= p\left(q_{r,t-1} = i, q_{r,t} = \mathrm{j} \middle| X_r,s_r,\Lambda'\right)
\end{aligned}
\qquad (5.12)
$$

which is the posterior probability of staying at state i at time $t - 1$ and staying at state j at time t, given the label sequence s_r and observation sequence X_r. An efficient forward–backward algorithm exists to compute this posterior probability in a standard way [43]. Then, we decompose $p(q| s, \Lambda)$ as follows:

$$
p(q|s,\Lambda) = \prod_{r=1}^{R} p\left(q_{r,1},\ldots,q_{r,T_r} | s_r,\Lambda\right) = \prod_{r=1}^{R} \prod_{t=1}^{T_r} a_{q_{r,t-1},q_{r,t}}
$$

This leads to the following simplifications:

$$\begin{aligned}
\text{term - III} &= \sum_s p\left(s \middle| X, \Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_q p\left(q \middle| X,\ s, \Lambda'\right)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\log a_{q_{r,t-1},q_{r,t}} \\
&= \sum_s p\left(s \middle| X, \Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{q,q_{r,t-1}=i,q_{r,t}=\mathrm{j}} p\left(q \middle| X, s, \Lambda'\right)\log a_{i,j} \\
&= \sum_s p\left(s \middle| X, \Lambda'\right)\left(C(s) - O(\Lambda')\right)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\sum_{j=1}^{I}\xi_{i,j,r,s_r}(t)\log a_{i,j} \qquad (5.13)
\end{aligned}$$

and

$$\begin{aligned}
\text{term - IV} &= \sum_s \mathrm{d}'(s)\sum_q\sum_\chi p\left(\chi, q \middle| s, \Lambda'\right)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\log a_{q_{r,t-1},q_{r,t}} \\
&= \sum_s \mathrm{d}'(s)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_q\sum_\chi p\left(\chi, q \middle| s, \Lambda'\right)\log a_{q_{r,t-1},q_{r,t}} \\
&= \sum_s \mathrm{d}'(s)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{q_{r,t-1}}\sum_{q_{r,t}} p\left(q_{r,t-1}, q_{r,t} \middle| s, \Lambda'\right)\log a_{q_{r,t-1},q_{r,t}} \\
&= \sum_s \mathrm{d}'(s)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\sum_{j=1}^{I} p\left(q_{r,t-1} = i \middle| s, \Lambda'\right) p\left(q_{r,t} = j \middle| q_{r,t-1} = i, s, \Lambda'\right)\log a_{i,j} \\
&= \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\mathrm{d}(r, t - 1, i)\sum_{j=1}^{I} a'_{i,j}\log a_{i,j} \qquad (5.14)
\end{aligned}$$

where $a'_{i,j} = p(q_{r,t} = j | q_{r,t-1} = i,\ s,\ \Lambda')$ is the transition probability from the previous GT iteration.

Substituting (5.19), (5.10), (5.13), and (5.14) into (5.7), and denoting the emitting probability by $b_i(x_{r,t}) = p(x_{r,t} | q_{r,t} = i,\ \Lambda)$ and $b'_i(x_{r,t}) = p(x_{r,t} | q_{r,t} = i,\ \Lambda')$, we obtain the decomposed and simplified objective function:

$$U(\Lambda;\Lambda') = U_1(\Lambda;\Lambda') + U_2(\Lambda;\Lambda') \qquad (5.15)$$

where

$$\begin{aligned}
U_1(\Lambda;\Lambda') = &\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\sum_s \left(s \middle| X, \Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t)\log b_i(x_{r,t}) \\
&+ \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I} d(r,t,i)\sum_{\chi_{r,t}} b'_i(\chi_{r,t})\log b_i(\chi_{r,t}) \qquad (5.16)
\end{aligned}$$

$$U_2(\Lambda;\Lambda') = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\sum_{j=1}^{I}\sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\xi_{i,j,r,s_r}(t)\log a_{i,j}$$
$$+\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I} d(r,t-1,i)\sum_{j=1}^{I} a'_{i,j}\log a_{i,j} \qquad (5.17)$$

In (5.15), $U_1(\Lambda;\Lambda')$ is relevant to optimizing the emitting probability $b_i(k)$, and $U_2(\Lambda;\Lambda')$ is relevant to optimizing the transition probability $a_{i,j}$.

5.1.4 GT by Optimizing Auxiliary Function $U(\Lambda;\Lambda')$

To optimize the discrete distribution $b_i(k) = p(x_{r,t} = k\,|\,q_{r,t} = i, \Lambda)$, $k = 1, 2, \ldots, K$, where the constraint $\sum_{k=1}^{K} b_i(k) = 1$ is imposed, we apply the Lagrange multiplier method by constructing

$$W_1(\Lambda;\Lambda') = U_1(\Lambda;\Lambda') + \sum_{i=1}^{I}\lambda_i\left(\sum_{k=1}^{K} b_i(k) - 1\right)$$

Setting $\dfrac{\partial W_1(\Lambda;\Lambda')}{\partial\lambda_i} = 0$ and $\dfrac{\partial W_1(\Lambda;\Lambda')}{\partial b_i(k)} = 0$, $k = 1, \ldots, K$, we have the following $K + 1$ equations:

$$\sum_{k=1}^{K} b_i(k) - 1 = 0$$

$$0 = \lambda_i b_i(k) + \sum_{r=1}^{R}\sum_{\substack{t=1\\ s.t. x_{r,t}=k}}^{T_r}\overbrace{\sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t)}^{\Delta\gamma(i,r,t)}\cdot$$
$$+\sum_{r=1}^{R}\sum_{t=1}^{T_r}\mathrm{d}(r,t,i)\, b'_i(k),\ k = 1,\ldots,K$$

where $b_i(k)$ is multiplied on both sides. Solving for $b_i(k)$, we obtain the reestimation formula:

$$b_i(k) = \frac{\displaystyle\sum_{r=1}^{R}\sum_{\substack{t=1\\ s.t. x_{r,t}=k}}^{T_r}\sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t) + b'_i(k)\sum_{r=1}^{R}\sum_{t=1}^{T_r}\mathrm{d}(r,t,i)}{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t) + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\mathrm{d}(r,t,i)} \qquad (5.18)$$

We now define

$$D_i = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\mathrm{d}(r,t,i) \qquad (5.19)$$

$$\Delta\gamma(i,r,t) = \sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t) \tag{5.20}$$

and can rewrite (5.18) as

$$b_i(k) = \frac{\displaystyle\sum_{r=1}^{R}\sum_{\substack{t=1\\ s.t.x_{r,t}=k}}^{T_r} \Delta\gamma\left(i,r,t\right) + b_i'(k)D_i}{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}\Delta\gamma\left(i,r,t\right) + D_i} \tag{5.21}$$

To optimize transition probabilities $a_{i,j}$, with constraint $\sum_{j=1}^{I} a_{i,j} = 1$, we apply Lagrange multiplier method by constructing

$$W_2(\Lambda;\Lambda') = U_2(\Lambda;\Lambda') + \sum_{i=1}^{I}\lambda_i\left(\sum_{j=1}^{I} a_{i,j} - 1\right) \tag{5.22}$$

Setting $\frac{\partial W_2(\Lambda;\Lambda')}{\partial \lambda_i} = 0$ and $\frac{\partial W_2(\Lambda;\Lambda')}{\partial a_{i,j}} = 0$ and, $j = 1, \ldots, I$, we have the following $I + 1$ equations:

$$\sum_{j=1}^{I} a_{i,j} - 1 = 0$$

$$0 = \lambda_i a_{i,j} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\overbrace{\sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\xi_{i,j,r,s_r}(t)}^{\Delta\xi(i,j,r,t)} + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\mathrm{d}(r,t-1,i)a_{i,j}', j = 1,\ldots,I$$

Note that $\sum_{j=1}^{I}\xi_{i,j,r,sr}(t) = \gamma_{i,r,sr}(t)$. Solving for $a_{i,j}$, we obtain the reestimation formula with a standard procedure (used for deriving the expectation–maximization estimate of transition probabilities [10]):

$$a_{i,j} = \frac{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\xi_{i,j,r,s_r}(t) + a_{i,j}'\sum_{r=1}^{R}\sum_{t=1}^{T_r}\mathrm{d}(r,t-1,i)}{\displaystyle\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{s} p\left(s\middle|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t) + \sum_{r=1}^{R}\sum_{t=1}^{T_r}\mathrm{d}(r,t-1,i)} \tag{5.23}$$

Now we define

$$\tilde{D}_i = \sum_{r=1}^{R} \sum_{t=1}^{T_r} \mathrm{d}(r, t-1, i) \tag{5.24}$$

$$\Delta\xi\left(i, j, r, t\right) = \sum_{s} p\left(s \,\middle|\, X, \Lambda'\right) \left(C(s) - O(\Lambda')\right) \xi_{i,j,r,s_r}(t) \tag{5.25}$$

and together with (5.20), we rewrite (5.23) as

$$a_{i,j} = \frac{\displaystyle\sum_{r=1}^{R} \sum_{t=1}^{T_r} \Delta\xi\left(i, j, r, t\right) + a'_{i,j}\tilde{D}_i}{\displaystyle\sum_{r=1}^{R} \sum_{t=1}^{T_r} \Delta\gamma(i, r, t) + \tilde{D}_i} \tag{5.26}$$

The parameter reestimation formulas (5.18) and (5.26) for discrete HMMs are unified across maximum mutual information (MMI), minimum classification error (MCE), and minimum phone error/minimum word error (MPE/MWE). What distinguishes among MMI, MCE, and MPE/MWE is the different weighing term $\Delta\gamma(i, r, t)$ in (5.20) and $\Delta\xi\,(i, j, r, t)$ in (5.25) due to the different $C(s)$ contained in the unified objective function. Details for computing $\Delta\gamma(i, r, t)$ for MMI, and MCE, and MPE/MWE can be found in Chapter 6.

5.2 ESTIMATION FORMULAS FOR CDHMM

For CDHMMs, $X = X_1, \ldots, X_R$, are continuous random variables. The previous objective functions for discrete HMMs still hold, except that χ is a continuous variable and hence the summation over domain χ is changed to integration over χ. Thus, we have the objective function:

$$V(\Lambda; \Lambda') = \sum_{s} \sum_{q} \int_{\chi} f(\chi, q, s, \Lambda') \log f(\chi, q, s, \Lambda) \mathrm{d}\chi \tag{5.27}$$

where the integrand $f(\chi, q, s, \Lambda)$ is defined by

$$F(\Lambda; \Lambda') = \sum_{s} \sum_{q} \int_{\chi} f(\chi, q, s, \Lambda) \mathrm{d}\chi \tag{5.28}$$

Correspondingly, we have

$$\begin{aligned} F(\Lambda;\Lambda') &= \sum_{s}\sum_{q}[C(s) - O(\Lambda')]p(s)p\left(X,q|s,\Lambda\right) + D \\ &= \sum_{s}\sum_{q}\int\limits_{\chi}\left[\Gamma(\Lambda') + \mathrm{d}(s)\right]p\left(\chi,q|s,\Lambda\right)\mathrm{d}\chi \end{aligned} \tag{5.29}$$

where

$$f(\chi,q,s,\Lambda) = \left[\Gamma(\Lambda') + \mathrm{d}(s)\right]p(\chi,q|s,\Lambda) \tag{5.30}$$

and

$$\Gamma(\Lambda') = \delta(\chi,X)p(s)\left[C(s) - O(\Lambda')\right] \tag{5.31}$$

with $\delta(\chi, X)$ in (5.31) being the Dirac delta function. After a similar derivation to that in the preceding section, it can be shown that the transition probability estimation formula (5.26) stays the same as for the discrete HMM. But for the emitting probability, (5.16) is changed to

$$\begin{aligned} U_1(\Lambda;\Lambda') = &\sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\sum_{s}p\left(s|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t)\log b_i(x_{r,t}) \\ &+ \sum_{r=1}^{R}\sum_{t=1}^{T_r}\sum_{i=1}^{I}\mathrm{d}(r,t,i)\int\limits_{\chi_{r,t}} b_i'(\chi_{r,t})\log b_i(\chi_{r,t})\mathrm{d}\chi_{r,t} \end{aligned} \tag{5.32}$$

As the most common member of the CDHMM, we use a Gaussian HMM to derive its parameters' GT formulas (the results for CDHMMs with exponential-family emitting probabilities can be derived similarly). For the Gaussian HMM, $b_i(x_{r,t})$ in (5.32) is a Gaussian distribution:

$$b_i(x_{r,t}) \propto \frac{1}{|\Sigma_i|^{1/2}}\exp\left[-\frac{1}{2}(x_{r,t} - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}(x_{r,t} - \mu_i)\right].$$

where $\{\mu_i, \Sigma_i\}$, $i = 1, 2, \ldots, I$ are the Gaussian mean vectors and covariance matrices.

To solve for μ_i and Σ_i, based on (5.32), for the Gaussian at HMM's state i, we set

$$\frac{\partial U_1(\Lambda;\Lambda')}{\partial \mu_i} = 0; \quad \text{and} \quad \frac{\partial U_1(\Lambda;\Lambda')}{\partial \Sigma_i} = 0.$$

This gives:

$$0 = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\overbrace{\sum_{s} p\left(s\,|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t)}^{\Delta\gamma(i,r,t)}\Sigma_i^{-1}(x_{r,t} - \mu_i)$$
$$+ \sum_{r=1}^{R}\sum_{t=1}^{T_r} \mathrm{d}(r,t,i)\Sigma_i^{-1}\int\limits_{\chi_{r,t}} b_i'(\chi_{r,t})(\chi_{r,t} - \mu_i)\mathrm{d}\chi_{r,t} \tag{5.33}$$

$$0 = \sum_{r=1}^{R}\sum_{t=1}^{T_r}\overbrace{\sum_{s} p\left(s\,|X,\Lambda'\right)\left(C(s) - O(\Lambda')\right)\gamma_{i,r,s_r}(t)}^{\Delta\gamma(i,r,t)}\left[\Sigma_i^{-1} - \Sigma_i^{-1}(x_{r,t} - \mu_i)(x_{r,t} - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}\right]$$
$$+ \sum_{r=1}^{R}\sum_{t=1}^{T_r} \mathrm{d}(r,t,i)\int\limits_{\chi_{r,t}} b_i'(\chi_{r,t})\left[\Sigma_i^{-1} - \Sigma_i^{-1}(\chi_{r,t} - \mu_i)(\chi_{r,t} - \mu_i)^{\mathrm{T}}\Sigma_i^{-1}\right]\mathrm{d}\chi_{r,t} \tag{5.34}$$

For the Gaussian distribution $b'_i(\chi_{r,t}) = p(\chi_{r,t}|q_{r,t} = i;\, \Lambda')$, we have

$$\int\limits_{\chi_{r,t}} b_i'(\chi_{r,t})\mathrm{d}\chi_{r,t} = 1,$$

$$\int\limits_{\chi_{r,t}} \chi_{r,t}\cdot b_i'(\chi_{r,t})\mathrm{d}\chi_{r,t} = \mu_i',$$

$$\int\limits_{\chi_{r,t}} (\chi_{r,t} - \mu_i')(\chi_{r,t} - \mu_i')^{\mathrm{T}}\cdot b_i'(\chi_{r,t})\mathrm{d}\chi_{r,t} = \Sigma_i'.$$

Hence, the integrals in (5.33) and (5.34) give closed-form results. Next, we multiply both sides of (5.33) by Σ_i at the left end, and multiply both sides of (5.34) by Σ_i at both left and right ends, respectively. Finally, solving μ_i and Σ_i gives the GT formulas of

$$\mu_i = \frac{\sum\limits_{r=1}^{R}\sum\limits_{t=1}^{T_r}\Delta\gamma(i,r,t)x_t + D_i\mu_i'}{\sum\limits_{r=1}^{R}\sum\limits_{t=1}^{T_r}\Delta\gamma(i,r,t) + D_i} \tag{5.35}$$

$$\Sigma_i = \frac{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\left[\Delta\gamma(i,r,t)(x_t-\mu_i)(x_t-\mu_i)^{\mathrm{T}}\right] + D_i\Sigma_i' + D_i(\mu_i-\mu_i')(\mu_i-\mu_i')^{\mathrm{T}}}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\Delta\gamma(i,r,t) + D_i} \tag{5.36}$$

where $\Delta\gamma(i, r, t)$ is defined in (5.20) and D_i defined in (5.19).

Just as for the discrete HMM presented in the preceding section, (5.35) and (5.36) are unified parameter estimation formulas for MMI, MCE, and MPE/MWE. And $\Delta\gamma(i, r, t)$ in (5.35) and (5.36) as defined in (5.20) is in the same way as for the discrete distribution case — differing in $C(s)$ for MMI, and MCE, and MPE/MWE, respectively, as will be detailed in Chapter 6.

5.3 RELATIONSHIP WITH GRADIENT-BASED METHODS

The relation between the GT method and gradient-based search method has been studied in the literature (e.g., [3, 46]). It can be shown that, with carefully selected, parameter-dependent step sizes, these two methods can be made identical. However, as was shown in [3] for MMI, the GT-based updating formula (5.35) is best viewed not as a simple gradient ascent but as an approximation to a quadratic Newton update; that is, it can be formulated as a gradient ascent with the step size that approximates inverse Hessian matrix H of the objective function. In the following, we will show the similar relationship for all MMI, MCE, and MPE/MWE enabled by our unifying framework.

Let us use *mean* vector estimation as an example. The gradient of $O(\Lambda)$ with respect to μ_i can be shown as:

$$\nabla_{\mu_i} O(\Lambda)|_{\Lambda=\Lambda'} = \Sigma_i'^{-1}\sum_{r=1}^{R}\sum_{t=1}^{T_r}\Delta\gamma(i,r,t)(x_t-\mu_i') \tag{5.37}$$

On the other hand, we can rewrite the GT formula of (5.35) into the following equivalent form

$$\begin{aligned}\mu_i &= \mu_i' + \frac{1}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\Delta\gamma(i,r,t)+D_i}\cdot\sum_{r=1}^{R}\sum_{t=1}^{T_r}\Delta\gamma(i,r,t)(x_t-\mu_i')\\ &= \mu_i' + \frac{1}{\sum_{r=1}^{R}\sum_{t=1}^{T_r}\Delta\gamma(i,r,t)+D_i}\Sigma_i'\cdot\nabla_{\mu_i}O(\Lambda)|_{\Lambda=\Lambda'}\end{aligned} \tag{5.38}$$

Consider the quadratic Newton update, where the Hessian H_i for μ_i can be approximated by the following equation after dropping the dependency of μ_i with $\Delta\gamma(i,r,t)$.

$$H_i = \nabla^2_{\mu_i} O(\Lambda)|_{\Lambda=\Lambda'} \approx -\Sigma_i'^{-1} \sum_{r=1}^{R} \sum_{t=1}^{T_r} \Delta\gamma(i,r,t)$$

Therefore, the updating formula of GT in (5.35) can be further rewritten to

$$\mu_i \approx \mu_i' - \underbrace{\frac{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \Delta\gamma(i,r,t)}{\sum_{r=1}^{R} \sum_{t=1}^{T_r} \Delta\gamma(i,r,t) + D_i}}_{\varepsilon_i} H_i^{-1} \nabla_{\mu_i} O(\Lambda)|_{\Lambda=\Lambda'} = \mu_i' + \varepsilon_i \nabla_{\mu_i} O(\Lambda)|_{\Lambda=\Lambda'} \qquad (5.39)$$

Compared with simple gradient ascent optimization that the model parameters Λ is updated by

$$\Lambda = \Lambda' + \varepsilon \cdot \nabla O(\Lambda)|_{\Lambda = \Lambda'}$$

where the step size ε is a single, global constant independent of the parameters. Equation (5.39) can be viewed either as a generalization of the simple gradient ascent where the global step size ε is replaced by the Gaussian-dependent step size ε_i, or as a generalization of the quadratic Newton update $\mu_i = \mu_i' - \alpha \cdot H_i^{-1} \nabla_{\mu i} O(\Lambda)|_{\Lambda = \Lambda'}$. Thus, the GT formula of (5.35) leads to more rapid convergence than the simple gradient-based search.

5.4 SETTING CONSTANT *D* FOR GT-BASED OPTIMIZATION

Based on Jensen's inequality, the theoretical basis for setting D_i is the requirement described in (5.2). That is, d(s) must be sufficiently large to ensure that for any string s and any observation sequence χ, $\Gamma(\Lambda') + \mathrm{d}(s) > 0$, where $\Gamma(\Lambda') = \delta(\chi, X) p(s)[C(s) - O(\Lambda')]$ from (5.3). However, for the CDHMM, $\delta(\chi, X)$ becomes the Dirac delta function, which is unbounded at the Center point. That is, $\delta(\chi, X) = +\infty$ when $\chi = X$. Therefore, for the string s that gives $C(\mathrm{s}) - O(\Lambda') < 0$, $\Gamma(\Lambda')|_{\chi = X} = -\infty$. Under this condition, it is impossible to find a bounded d(s) that ensures $\Gamma(\Lambda') + \mathrm{d}(s) > 0$ and hence Jensen's inequality may not apply. (Note that the discrete HMM does not encounter such a difficulty because $\delta(\chi, X)$ takes final values of 0 or 1.)

The above difficulty for CDHMMs can be overcome if it can be shown that there exists a sufficiently large but still bounded constant D so that $V(\Lambda; \Lambda')$ of (5.27), with the integrand defined by (5.30) is still a valid auxiliary function of $F(\Lambda; \Lambda')$; that is, an increase in the value of $V(\Lambda; \Lambda')$

can guarantee the increase in the value of $F(\Lambda; \Lambda')$. Such a proof has indeed been developed in the recent work of Axelrod et al. [3], which we will outline in the later part of this section. (Because $V(\Lambda; \Lambda')$ is still a valid auxiliary function, all derivations from (5.27) to (5.36) are valid.)

Given sufficiently large D_i, the convergence of the model estimation formulas, that is, (5.21), (5.26), (5.35), and (5.36), can be proved. However, the value of D_i that guarantees convergence is usually too large to obtain a reasonable convergence speed. Before further research advance to lower the value of D_i, in practice, D_i is often empirically set to achieve compromised training performance.

Empirical setting of D_i has been extensively studied since GT/EBW was proposed. In the early days, only one global constant D was used for all parameters [14, 34]. Later research discovered on the empirical basis that for CDHMM, a useful lower bound on (nonglobal) D_i is the value satisfying the constraint that the newly estimated variances remain positive [35]. In Refs. [50, 51], this constraint was further explored, leading to quadratic inequalities with which the lower bound of D_i can be solved. Most recently, in [46], constant D_i was further bounded by an extra condition that the denominators in the reestimation formula remain nonsingular.

In [52], use of Gaussian-specific D_i was reported to give further improved convergence speed. For MMI, the Gaussian-specific constant D_i was set empirically to be the maximum of (i) twice the value necessary to ensure positive variance, that is, $2 \cdot D_{\min}$; and (ii) a global constant E multiplied by the denominator occupancy; for example, $E \cdot \gamma_i^{\text{den}}$. Specifically, for MMI in the work of Woodland and Povey [52], $\gamma_i^{\text{den}} = \sum_{r=1}^{R} \sum_{t=1}^{T} \gamma_{i,r}^{\text{den}}(t) = \sum_{r=1}^{R} \sum_{t=1}^{Tr} \sum_{Sr} p(s_r|X_r, \Lambda')\gamma_{i,r,s_r}(t)$. For MPE reported in [38–40], the empirical setting of D_i was the same as MMI, that is, $D_i = \max\{2 \cdot D_{\min}, \mathrm{E} \cdot \gamma_i^{\text{den}}\}$ except that the computation of the denominator occupancy became: $\gamma_i^{\text{den}} = \sum_{r=1}^{R} \sum_{t=1}^{Tr} \max(0, -\Delta\gamma(i, r, t))$. Moreover, the obtained new parameters were smoothed with the ML estimate of parameters (which was called I-smoothing).

For MCE, in our previous experimental work [20, 58], we developed the empirical setting of γ_i^{den} as $\sum_{r=1}^{R} \sum_{t=1}^{Tr} p(S_r|X_r, \Lambda')\gamma_{i,r,s_r}(t)$. It was based on the consideration that MCE and MMI are equivalent in the special case of having one utterance in the training set and hence the parameter estimation formulas of them should be identical in this special case. We tested this setting and obtained strong results as reported in [20, 58]. Further discussions and comparisons of different settings of empirical D_i can be found in [14, 20, 34, 35, 40, 46, 51, 52].

5.4.1. Existence Proof of Finite D in GT Updates for CDHMM

As discussed earlier, optimization based on Jensen's inequality cannot be applied directly to Gaussian CDHMM because the value D in the GT update formulas (5.35) and (5.36) may be infinite, making the algorithm's convergence infinitely slow. In this section, we follow the insight provided in

[3] to prove that there exist finite values of D that make the GT update formulas (5.35) and (5.36) practical for all MMI, MCE, and MPE/MWE.

To proceed, we substitute (5.30) into (5.27) and obtain

$$V(\Lambda;\Lambda') = \sum_q \sum_s \int_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p\left(\chi, q \middle| s, \Lambda'\right) \log p\left(\chi, q \middle| s, \Lambda\right) + \mathrm{Const.} \tag{5.40}$$

We prove below that for a CDHMM, given a sufficiently large but bounded (i.e., finite) constant D,

$$F(\Lambda;\Lambda') - F(\Lambda';\Lambda') \geq V(\Lambda;\Lambda') - V(\Lambda';\Lambda') \tag{5.41}$$

First, we define

$$\Delta_D = \left[F(\Lambda;\Lambda') - F(\Lambda';\Lambda')\right] - \left[V(\Lambda;\Lambda') - V(\Lambda';\Lambda')\right] \tag{5.42}$$

and will show that $\Delta_D \geq 0$ for any parameter set Λ. Substituting (5.29) and (5.40) into (5.42), we obtain

$$\begin{aligned}
\Delta_D &= \left[F(\Lambda;\Lambda') - F(\Lambda';\Lambda')\right] - \left[V(\Lambda;\Lambda') - V(\Lambda';\Lambda')\right] \\
&= \sum_q \sum_s \int_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] \left[p\left(\chi, q \middle| s, \Lambda\right) - p(\chi, q \middle| s, \Lambda'\right] \mathrm{d}\chi \\
&\quad - \sum_q \sum_s \int_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p\left(\chi, q \middle| s, \Lambda'\right) \left[\log p\left(\chi, q \middle| s, \Lambda\right) - \log p\left(\chi, q \middle| s, \Lambda'\right)\right] \mathrm{d}\chi \qquad (5.43) \\
&= \sum_q \sum_s \int_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p\left(\chi, q \middle| s, \Lambda'\right) \left[\frac{p\left(\chi, q \middle| s, \Lambda\right)}{p\left(\chi, q \middle| s, \Lambda'\right)} - 1 - \log \frac{p\left(\chi, q \middle| s, \Lambda\right)}{p\left(\chi, q \middle| s, \Lambda'\right)}\right] \mathrm{d}\chi \\
&= \sum_q \sum_s \int_\chi \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p\left(\chi, q \middle| s, \Lambda'\right) \mathrm{H}\left(\chi, q, s, \Lambda, \Lambda'\right) \mathrm{d}\chi
\end{aligned}$$

where

$$\mathrm{H}\left(\chi, q, s, \Lambda, \Lambda'\right) = \left[\frac{p\left(\chi, q \middle| s, \Lambda\right)}{p\left(\chi, q \middle| s, \Lambda'\right)} - 1\right] - \log\left[\left[\frac{p\left(\chi, q \middle| s, \Lambda\right)}{p\left(\chi, q \middle| s, \Lambda'\right)} - 1\right] + 1\right]$$

Then, we need to show that there exists a bounded $d(s)$ that ensures the summand of Δ_D in (5.43) be nonnegative. To proceed, we expand the summand to

$$\int_{\chi} \left[\Gamma(\Lambda') + \mathrm{d}(s)\right] p\left(\chi, q \middle| s, \Lambda'\right) \mathrm{H}\left(\chi, q, s, \Lambda, \Lambda'\right) \mathrm{d}\chi$$

$$= p(s)\left[C(s) - O(\Lambda')\right] p\left(X, q \middle| s, \Lambda'\right) \mathrm{H}\left(X, q, s, \Lambda, \Lambda'\right) + \mathrm{d}(s) \int_{\chi} p\left(\chi, q \middle| s, \Lambda'\right) \mathrm{H}\left(\chi, q, s, \Lambda, \Lambda'\right) \mathrm{d}\chi \qquad (5.44)$$

We now use the following key theorem from [3]: If $f(X, \Lambda)$ is nonnegative and analytic for $X \in \chi$ and $\Lambda \in \Omega$, where χ and Ω are the data space and model space, respectively, then there is a Λ-independent constant $K > 0$ such that

$$\int_{\chi} f(\chi, \Lambda) \mathrm{d}\chi \geq K f(X, \Lambda) \qquad (5.45)$$

for any valid model Λ. (Readers are referred to [3] for a rigorous proof.)

Define $f(X, \Lambda) = p(X, q|s, \Lambda')\ \mathrm{H}(X, q, s, \Lambda, \Lambda')$. Here, $f(X, \Lambda)$ is nonnegative and analytic because both $p(X, q|s, \Lambda')$ and $\mathrm{H}(X, q, s, \Lambda, \Lambda')$ are nonnegative and analytic (for CDHMM). Using (5.45), we have

$$\int_{\chi} p\left(\chi, q \middle| s, \Lambda'\right) \mathrm{H}\left(\chi, q, s, \Lambda, \Lambda'\right) \mathrm{d}\chi \geq K p\left(X, q \middle| s, \Lambda'\right) \mathrm{H}\left(X, q, s, \Lambda, \Lambda'\right) \qquad (5.46)$$

Now we construct nonnegative $\mathrm{d}(s)$ as follows:

$$\mathrm{d}(s) = \begin{cases} 0 & \text{if } C(s) \geq O(\Lambda') \\ \dfrac{1}{K} p(s)\left(O(\Lambda') - C(s)\right) & \text{if } C(s) < O(\Lambda') \end{cases} \qquad (5.47)$$

Then, (5.46) becomes

$$\mathrm{d}(s) \int_{\chi} p\left(\chi, q \middle| s, \Lambda'\right) \mathrm{H}\left(\chi, q, s, \Lambda, \Lambda'\right) \mathrm{d}\chi > -p(s)\left[C(s) - O(\Lambda')\right] p\left(X, q \middle| s, \Lambda'\right) \mathrm{H}\left(X, q, s, \Lambda, \Lambda'\right)$$

This proves that the summand of Δ_D, $\int_{\chi}[\Gamma(\Lambda') + \mathrm{d}(s)]\, p(\chi, q|s, \Lambda')\ \mathrm{H}(X, q, s, \Lambda, \Lambda')\mathrm{d}\chi$, is nonnegative for any s (according to (5.44)), and therefore $\Delta_D \geq 0$.

• • • •