Notes: Entropically Constrained HMM Parameter Estimation

Graham Grindlay
grindlay@soe.ucsc.edu

June 8, 2004

1 Preliminaries

- We are considering absorbing, discrete-observation hidden Markov models.
- We refer to the parameters generically as θ_i (an entry in a multinomial).
- Trying to update parameter estimates in the model given a batch of data sequences Ω, where each sequence $X \in \Omega$ is comprised of a set of observations such that $X = \{x_1, x_2, ..., x_T\}$.

2 The Joint-Entropy Update (Singer & Warmuth, 1996)

- Goal is to reestimate parameters by maximizing a weighted combination of log-likelihood and model divergence:

$$U(\tilde{\theta}) = \mathcal{LL}(\tilde{\theta}|\Omega) - \frac{1}{\eta}\Delta(\tilde{\theta}, \theta) \tag{1}$$

- Intuitively, want to stay close to our old parameters which encapsulate all that we have learned so far.
- $\mathcal{LL}(\tilde{\theta}|\Omega)$ depends on usage statistics from the model whose parameters we are trying to solve for!
- Approximate with first-order Taylor expansion:

$$U(\tilde{\theta}) = \Big(\mathcal{LL}(\theta|\Omega) + (\tilde{\theta} - \theta)\nabla_\theta(\mathcal{LL}(\theta|\Omega)\Big) - \frac{1}{\eta}\Delta(\tilde{\theta}, \theta)$$

- $\Delta(\tilde{\theta}, \theta)$ is a relative entropy measuring the divergence between the distributions induced by two HMMs over all possible data and hidden state sequences. This also depends on usage statistics and so is approximated as well:

$$\hat{\tilde{\Delta}}(\tilde{\theta}, \theta) = \sum_i \hat{n}_{q_i}(\theta) \sum_{j \in q_i} \tilde{\theta}_j \log \frac{\tilde{\theta}_j}{\theta_j}$$

  where $\hat{n}_{q_i}(\theta)$ is the expected usage of state q_i.
- After setting derivatives to 0 and solving, we arrive at the batch JE batch update:

$$\tilde{\theta}_i = \frac{\theta_i \exp\left(\frac{\eta}{\hat{n}_{q_i}(\theta)} \frac{\sum_{X \in \Omega} \hat{n}_{\theta i}(X|\theta)}{|\Omega|\theta_i}\right)}{\sum_{j \in \theta^i} \theta_j \exp\left(\frac{\eta}{\hat{n}_{q_j}(\theta)} \frac{\sum_{X \in \Omega} \hat{n}_{\theta j}(X|\theta)}{|\Omega|\theta_j}\right)} \tag{2}$$

3 The Entropic MAP Estimator (Brand, 1998)

- Very similar form to JE framework, but formulated in a Bayesian fashion where the objective function, $U(\tilde{\theta})$ is interpreted as a log-posterior.
- Given some evidence, ω (in our case usage counts), construct the following posterior:

$$P(\tilde{\theta}|\omega) = \frac{P(\tilde{\theta})\mathcal{L}(\tilde{\theta}|\omega)}{P(\omega)} \tag{3}$$

$$= \frac{e^{-\frac{1}{\eta}\Delta(\tilde{\theta},\theta)}\mathcal{L}(\tilde{\theta}|\omega)}{P(\omega)} \tag{4}$$

$$\propto e^{-\frac{1}{\eta}\Delta(\tilde{\theta},\theta)}P(\omega|\tilde{\theta}) \tag{5}$$

$$\log(P(\tilde{\theta}|\omega)) \propto \mathcal{LL}(\tilde{\theta}|\omega) - \frac{1}{\eta}\Delta(\tilde{\theta},\theta) \tag{6}$$

- Brand's formulation uses the Shannon entropy ($e^{-H(\tilde{\theta})}$) as the prior, thus biasing updates towards sparse distributions.
- We use the relative entropy, which can include Shannon entropy as a special case (using the uniform distribution).
- The expected complete-data log-likelihood (like the E step in EM) is used to form $\hat{\mathcal{LL}}$ and it's derivative:

$$\hat{\mathcal{LL}}(\tilde{\theta}|\Omega) = \frac{1}{|\Omega|}\sum_{X\in\Omega}\log(\prod_i \tilde{\theta}_i^{\,\hat{n}_{\theta i}(X|\theta)}) \tag{7}$$

$$\frac{\partial}{\partial\theta_i}\hat{\mathcal{LL}}(\tilde{\theta}|\Omega) = \frac{\sum_{X\in\Omega}\hat{n}_{\theta i}(X|\theta)}{|\Omega|\tilde{\theta}_i} \tag{8}$$

  where $\hat{n}_{\theta_i}(X|\theta)$ is the expected number of times that parameter θ_i was used over all state sequences that produce X in the HMM with parameters, θ.
- Now define the evidence to be:

$$\omega_i = \frac{\sum_{X\in\Omega}\hat{n}_{\theta i}(X|\theta)}{|\Omega|} \tag{9}$$

- In this form, the log-likelihood depends only on the evidence term rather than the data and hidden variables. In the case of HMMs, this effectively decouples the model parameters so that we can decompose the model into a set of independent multinomial distributions and solve each independently.

3.1 Multinomial Distributions

- Simplifies the problem considerably.
- Can now use exact relative entropy:

$$\Delta(\tilde{\theta}, \theta) = \sum_i \tilde{\theta}_i \log \frac{\tilde{\theta}_i}{\theta_i} \tag{10}$$

- We can write the posterior as follows:

$$\begin{aligned} P(\tilde{\theta}|\omega) &\propto \mathcal{L}(\tilde{\theta}|\omega)\ P(\tilde{\theta}) \\ &\propto \prod_i {\tilde{\theta}_i}^{\omega_i} \prod_i \left(\frac{\tilde{\theta}_i}{\theta_i}\right)^{-\left(\frac{\tilde{\theta}_i}{\eta}\right)} \\ &\propto \prod_i {\tilde{\theta}_i}^{\omega_i} \left(\frac{\tilde{\theta}_i}{\theta_i}\right)^{-\left(\frac{\tilde{\theta}_i}{\eta}\right)} \end{aligned} \tag{11}$$

- The posterior has an interpretation as a minimization of entropies:

$$\begin{aligned} -\max_{\tilde{\theta}} \log(P(\tilde{\theta}|\omega)) &= \min_{\tilde{\theta}} -\log\left(\prod_i {\tilde{\theta}_i}^{\omega_i} \left(\frac{\tilde{\theta}_i}{\theta_i}\right)^{-\left(\frac{\tilde{\theta}_i}{\eta}\right)}\right) \\ &= \min_{\tilde{\theta}} \left(\Delta(\omega, \tilde{\theta}) + H(\omega) + \frac{1}{\eta}\Delta(\tilde{\theta}, \theta)\right) \end{aligned} \tag{12}$$

- To solve for new parameters, set the log-posterior to 0, and add a Lagrange multiplier to ensure that the parameters sum to 1:

$$\begin{aligned} 0 &= \frac{\partial}{\partial \tilde{\theta}_i}\left[\log\left(\prod_i {\tilde{\theta}_i}^{\omega_i} \left(\frac{\tilde{\theta}_i}{\theta_i}\right)^{-\left(\frac{\tilde{\theta}_i}{\eta}\right)}\right) + \lambda\left(\sum_i \tilde{\theta}_i - 1\right)\right] \\ &= \frac{\omega_i}{\tilde{\theta}_i} - \frac{1}{\eta}\log\frac{\tilde{\theta}_i}{\theta_i} - \frac{1}{\eta} + \lambda \end{aligned} \tag{13}$$

- We can easily solve for λ, but $\tilde{\theta}_i$ is a bit more tricky due to the mixed polynomial and logarithmic terms.
- We can use the Lambert W function: $W(y)e^{W(y)} = y$. This gives:

$$\tilde{\theta}_i = \frac{\eta\omega_i}{W\left(\frac{\eta\omega_i}{\theta_i} e^{1-\eta\lambda}\right)} \tag{14}$$

- $\tilde{\theta}_i$ and λ form a fix-point solution for λ and therefore $\tilde{\theta}_i$. We can iteratively update the *current update's* estimate of $\tilde{\theta}$ by first solving for $\tilde{\theta}$ given λ, normalizing $\tilde{\theta}$, and then calculating λ given $\tilde{\theta}$. Convergence is fast (2-5 iterations).

4 The Joint-Entropy MAP Estimator

- Hybrid of the Joint-Entropy update and entropic MAP estimator.
- Use expected complete-data log-likelihood of new parameters.
- Use approximated (probably can get around this) relative entropy with respect to the entire HMM.
- Setup up and solve log-posterior in similar fashion to multinomial case.
- In the following derivations, superscripted variables denote the state or distribution of which the indexing parameter is a member:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \theta_i} \hat{\mathcal{L}}\mathcal{L}(\tilde{\theta}|\omega) - \frac{\partial}{\partial \tilde{\theta}_i} \frac{1}{\eta} \hat{\tilde{\Delta}}(\tilde{\theta}, \theta) + \frac{\partial}{\partial \tilde{\theta}_i} \lambda_{\tilde{\theta}^i} \left(\sum_i^{|\tilde{\theta}^i|} \tilde{\theta}_i - 1 \right) \\
&= \frac{\omega_i}{\tilde{\theta}_i} - \frac{\hat{n}_{q^i}(\theta)}{\eta} \log \frac{\tilde{\theta}_i}{\theta_i} - \frac{\hat{n}_{q^i}(\theta)}{\eta} + \lambda_{\tilde{\theta}^i} \qquad (15) \\
& \qquad (16)
\end{aligned}
$$

- Easy to solve for $\lambda_{\tilde{\theta}^i}$:

$$
\lambda_{\tilde{\theta}^i} = -\frac{\omega_i}{\tilde{\theta}_i} + \frac{\hat{n}_{q^i}(\theta)}{\eta} \log \frac{\tilde{\theta}_i}{\theta_i} + \frac{\hat{n}_{q^i}(\theta)}{\eta} \qquad (17)
$$

- Now let $a = \frac{\hat{n}_{q^i}(\theta)}{\eta}$ to simplify the expression a bit.

$$
0 = a \left[\frac{\omega_i}{a\tilde{\theta}_i} - \log \frac{\tilde{\theta}_i}{\theta_i} - 1 + \frac{\lambda_{\tilde{\theta}^i}}{a} \right] \qquad (18)
$$

- Solve for $\tilde{\theta}_i$ by working backwards from the W function to the bracketed expression in (18). First, note that the W function can be re-written as: $W(y) + \log(W(y)) = \log(y)$. Now let $y = e^m$.

$$
\begin{aligned}
0 &= W(e^m) + \log(W(e^m)) - m \\
&= \frac{z}{z/W(e^m)} + \log(W(e^m)) - m + \log z - \log z \\
&= \frac{z}{z/W(e^m)} + \log \left(\frac{W(e^m)}{z} \right) - m + \log z \qquad (19)
\end{aligned}
$$

Now, let $m = 1 - \frac{\lambda_{\tilde{\theta}^i}}{a} + \log z - \log \theta_i$. Substituting in m and continuing, we have:

$$
\begin{aligned}
0 &= \frac{z}{z/W\left(e^{1-\frac{\lambda}{a}+\log z-\log\theta_i}\right)} + \log\left(\frac{W\left(e^{1-\frac{\lambda}{a}+\log z-\log\theta_i}\right)}{z}\right) - 1 + \frac{\lambda}{a} + \log\theta_i \\
&= \frac{z}{z/W\left(\frac{z}{\theta_i}e^{1-\frac{\lambda}{a}}\right)} + \log\left(\frac{W\left(\frac{z}{\theta_i}e^{1-\frac{\lambda}{a}}\right)}{z}\right) - 1 + \frac{\lambda}{a} + \log\theta_i \\
&= \frac{z}{z/W\left(\frac{z}{\theta_i}e^{1-\frac{\lambda}{a}}\right)} - \log\left(\frac{z}{\theta_i\, W\left(\frac{z}{\theta_i}e^{1-\frac{\lambda}{a}}\right)}\right) - 1 + \frac{\lambda}{a} \\
&= \frac{\frac{\omega_i}{a}}{\frac{\omega_i}{a}/W\left(\frac{\omega_i}{a\theta_i}e^{1-\frac{\lambda}{a}}\right)} - \log\left(\frac{\frac{\omega_i}{a}}{\theta_i\, W\left(\frac{\omega_i}{a\theta_i}e^{1-\frac{\lambda}{a}}\right)}\right) - 1 + \frac{\lambda}{a}
\end{aligned} \tag{20}
$$

Where we have let $z = \frac{\omega_i}{a}$. This implies that:

$$
\tilde{\theta}_i = \frac{\omega_i/a}{W\left(\frac{\omega_i}{a\theta_i}e^{1-\frac{\lambda}{a}}\right)} \tag{21}
$$

Substituting (21) into (20), we get:

$$
0 = \frac{\omega_i/a}{\tilde{\theta}_i} - \log\left(\frac{\tilde{\theta}_i}{\theta_i}\right) - 1 + \frac{\lambda}{a} \tag{22}
$$

Which is the derivative of the log-posterior divided by the constant a (ie. the bracketed term in (18)). Multiplying by a we get:

$$
\begin{aligned}
0 &= \frac{\omega_i}{\tilde{\theta}_i} - a\log\left(\frac{\tilde{\theta}_i}{\theta_i}\right) - a + \lambda \\
&= \frac{\omega_i}{\tilde{\theta}_i} - \frac{\hat{n}_{q^i}(\theta)}{\eta}\log\left(\frac{\tilde{\theta}_i}{\theta_i}\right) - \frac{\hat{n}_{q^i}(\theta)}{\eta} + \lambda_{\tilde{\theta}^i}
\end{aligned} \tag{23}
$$

- We have arrived back at (15) and therefore derived an expression for $\tilde{\theta}_i$. Now we can substitute back in for ω_i and a:

$$
\begin{aligned}
\tilde{\theta}_i &= \frac{\omega_i/a}{W\left(\frac{\omega_i}{a\theta_i}e^{1-\frac{\lambda}{a}}\right)} \\
&= \frac{\frac{\eta\sum_{X\in\Omega}\hat{n}_{\theta i}(X|\theta)}{|\Omega|\hat{n}_{q^i}(\theta)}}{W\left(\frac{\eta\sum_{X\in\Omega}\hat{n}_{\theta i}(X|\theta)}{\theta_i|\Omega|\hat{n}_{q^i}(\theta)}\exp(1-\frac{\eta\lambda}{\hat{n}_{q^i}(\theta)})\right)}
\end{aligned} \tag{24}
$$