Notes on Inference and Learning in HMMs

Xun Zheng

February 17, 2019

1 Problem Setup

Consider an HMM with T time steps, M discrete states, and K-dimensional observations as in Figure 1, where $\mathbf{z}_t \in \{0,1\}^M$, $\|\mathbf{z}_t\| = 1$, $\mathbf{x}_t \in \mathbb{R}^K$ for $t \in [T]$.

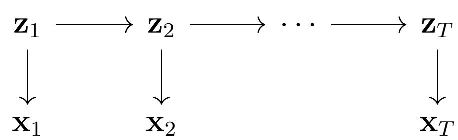

Figure 1: A hidden Markov model.

The joint distribution factorizes over the graph:

$$p(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}) = p(\mathbf{z}_1) \prod_{t=2}^{T} p(\mathbf{z}_t | \mathbf{z}_{t-1}) \prod_{t=1}^{T} p(\mathbf{x}_t | \mathbf{z}_t). \tag{1}$$

Now consider the parameterization of CPDs. Let $\boldsymbol{\pi} \in \mathbb{R}^M$ be the initial state distribution and $A \in \mathbb{R}^{M \times M}$ be the transition matrix. The emission density $f(\cdot)$ is parameterized by $\boldsymbol{\phi}_i$ at state i. In other words,

$$p(z_{1i} = 1) = \pi_i, \qquad p(\mathbf{z}_1) = \prod_{i=1}^{M} \pi_i^{z_{1i}}, \tag{2}$$

$$p(z_{tj} = 1 | z_{t-1,i} = 1) = a_{ij}, \qquad p(\mathbf{z}_t | \mathbf{z}_{t-1}) = \prod_{i=1}^{M} \prod_{j=1}^{M} a_{ij}^{z_{t-1,i} z_{tj}}, \qquad t = 2, \ldots, T \tag{3}$$

$$p(\mathbf{x}_t | z_{ti} = 1) = f(\mathbf{x}_t; \boldsymbol{\phi}_i), \qquad p(\mathbf{x}_t | \mathbf{z}_t) = \prod_{i=1}^{M} f(\mathbf{x}_t; \boldsymbol{\phi}_i)^{z_{ti}}, \qquad t = 1, \ldots, T. \tag{4}$$

Define $\theta = (\boldsymbol{\pi}, A, \{\boldsymbol{\phi}_i\}_{i=1}^{M})$ to be the set of parameters of the HMM.

2 The Baum-Welch algorithm

Let $\widehat{p}$ be the empirical distribution of $\mathbf{x}_{1:T}$. We would like to find MLE of θ by solving the following problem:

$$\max_{\theta} \ \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\log p_\theta(\mathbf{x}_{1:T})\right]. \tag{5}$$

However the marginal likelihood is intractable due to summation over M^T terms:

$$p_\theta(\mathbf{x}_{1:T}) = \sum_{\mathbf{z}_{1:T}} p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T}). \tag{6}$$

A variational distribution $q(\mathbf{z}_{1:T})$ can be introduced to derive a lower bound of the marginal likelihood:

$$\begin{aligned} L(\mathbf{x}_{1:T}; \theta, q) &\coloneqq \log p_\theta(\mathbf{x}_{1:T}) - \underbrace{\mathrm{KL}\left[q(\mathbf{z}_{1:T}) \| p_\theta(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\right]}_{\geq 0} && (7) \\ &= \underbrace{\mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[\log p_\theta(\mathbf{x}_{1:T}, \mathbf{z}_{1:T})\right]}_{=:F(\mathbf{x}_{1:T};\theta)} + \mathrm{H}\left[q(\mathbf{z}_{1:T})\right]. && (8) \end{aligned}$$

The EM algorithm maximizes the lower bound as a surrogate:

$$\max_{\theta, q} \ \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[L(\mathbf{x}_{1:T}; \theta, q)\right]. \tag{9}$$

Alternatively maximizing (9) w.r.t. (θ, q) results in the following updates:

- (E-step) Maximize (7) w.r.t. q:

$$\begin{aligned} q^*(\mathbf{z}_{1:T}) &= \operatorname*{argmin}_{q(\mathbf{z}_{1:T})} \ \mathrm{KL}\left[q(\mathbf{z}_{1:T}) \| p_\theta(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})\right] && (10) \\ &= p_\theta(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}). && (11) \end{aligned}$$

  The optimal q^* is the posterior parameterized by the current θ.

- (M-step) Maximize (8) w.r.t. θ:

$$\theta^* = \operatorname*{argmax}_{\theta} \ \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[F(\mathbf{x}_{1:T}; \theta)\right] \tag{12}$$

  The optimal θ^* is the MLE of a fully observed model, where the "observed" hidden variables $\mathbf{z}_{1:T}$ follow q^*, the posterior parameterized by the current θ.

3 The M-step objective

The factorization (1) allows decomposition of expected joint likelihood:

$$F(\mathbf{x}_{1:T};\theta) = \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[\log p_\theta(\mathbf{x}_{1:T},\mathbf{z}_{1:T})\right] \tag{13}$$

$$= \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[\log p(\mathbf{z}_1) + \sum_{t=2}^{T}\log p(\mathbf{z}_t|\mathbf{z}_{t-1}) + \sum_{t=1}^{T}\log p(\mathbf{x}_t|\mathbf{z}_t)\right] \tag{14}$$

$$= \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[\sum_{i=1}^{M} z_{1i}\log\pi_i\right] + \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[\sum_{t=2}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M} z_{t-1,i}z_{tj}\log a_{ij}\right] \tag{15}$$

$$+ \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[\sum_{t=1}^{T}\sum_{i=1}^{M} z_{ti}\log f(\mathbf{x}_t;\boldsymbol{\phi}_i)\right]. \tag{16}$$

Define shorthands γ and ξ for the posterior expectations:

$$\gamma(z_{ti}) \coloneqq \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[z_{ti}\right], \quad t = 1,\ldots,T \tag{17}$$

$$\xi(z_{t-1,i},z_{tj}) \coloneqq \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[z_{t-1,i}z_{tj}\right]. \quad t = 2,\ldots,T \tag{18}$$

Then

$$F(\mathbf{x}_{1:T};\theta) = \sum_{i=1}^{M}\gamma(z_{1i})\log\pi_i + \sum_{t=2}^{T}\sum_{i=1}^{M}\sum_{j=1}^{M}\xi(z_{t-1,i},z_{tj})\log a_{ij} \tag{19}$$

$$+ \sum_{t=1}^{T}\sum_{i=1}^{M}\gamma(z_{ti})\log f(\mathbf{x}_t;\boldsymbol{\phi}_i). \tag{20}$$

4 Parameter estimation given $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$

Suppose γ and ξ are given. The MLE (12) has closed form for $\boldsymbol{\pi}$ and A:

$$\max_{\boldsymbol{\pi}\in\Delta}\ \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{i=1}^{M}\gamma(z_{1i})\log\pi_i\right] \implies \pi_i^* \propto \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\gamma(z_{1i})\right], \tag{21}$$

$$\max_{\boldsymbol{a}_j\in\Delta}\ \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{t=2}^{T}\sum_{i=1}^{M}\xi(z_{t-1,i},z_{tj})\log a_{ij}\right] \implies a_{ij}^* \propto \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{t=2}^{T}\xi(z_{t-1,i},z_{tj})\right]. \tag{22}$$

The MLE of $\boldsymbol{\phi}$ has closed form depending on the choice of $f(\cdot)$. For instance, when emission is isotropic Gaussian,

$$f(\mathbf{x}_t;\boldsymbol{\phi}_i) = \mathsf{N}(\mathbf{x}_t;\boldsymbol{\mu}_i,\sigma_i^2 I), \tag{23}$$

whose log-density is

$$\log f(\mathbf{x}_t;\boldsymbol{\phi}_i) = -\frac{K}{2}\log\sigma_i^2 - \frac{1}{2\sigma_i^2}\left\|\mathbf{x}_t-\boldsymbol{\mu}_i\right\|_2^2 + \text{constant}, \tag{24}$$

then the corresponding MLE problem

$$\max_{\boldsymbol{\mu}_i,\sigma_i^2} \ \mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{t=1}^{T}\gamma(z_{ti})\log f(\mathbf{x}_t;\boldsymbol{\phi}_i)\right] \tag{25}$$

has closed form

$$\mu_{ik}^* = \frac{\mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{t=1}^{T}\gamma(z_{ti})\mathbf{x}_t\right]}{\mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{t=1}^{T}\gamma(z_{ti})\right]}, \quad \sigma_i^{2*} = \frac{\mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{t=1}^{T}\gamma(z_{ti})\,\|\mathbf{x}_t-\boldsymbol{\mu}_i\|_2^2\right]}{\mathbb{E}_{\mathbf{x}_{1:T}\sim\widehat{p}}\left[\sum_{t=1}^{T}\gamma(z_{ti})K\right]}. \tag{26}$$

5 Exact inference for $\boldsymbol{\gamma}$ and $\boldsymbol{\xi}$

Recall in (17) and (18) the expectation is taken w.r.t. the posterior parameterized by the current estimate $\hat{\theta}$:

$$q(\mathbf{z}_{1:T}) = p_{\hat{\theta}}(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}), \tag{27}$$

which means γ and ξ are in fact unary and pairwise posterior marginals:

$$\gamma(z_{ti}) = \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[z_{ti}\right] = p_{\hat{\theta}}(z_{ti}=1|\mathbf{x}_{1:T}), \tag{28}$$

$$\xi(z_{t-1,i}, z_{tj}) = \mathbb{E}_{\mathbf{z}_{1:T}\sim q}\left[z_{t-1,i}z_{tj}\right] = p_{\hat{\theta}}(z_{t-1,i}z_{tj}=1|\mathbf{x}_{1:T}). \tag{29}$$

The goal of this section is to perform inference for *all* such marginal queries:

$$\gamma(\mathbf{z}_t) = p_{\hat{\theta}}(\mathbf{z}_t|\mathbf{x}_{1:T}), \quad t=1,\ldots,T \tag{30}$$

$$\xi(\mathbf{z}_{t-1},\mathbf{z}_t) = p_{\hat{\theta}}(\mathbf{z}_{t-1},\mathbf{z}_t|\mathbf{x}_{1:T}). \quad t=2,\ldots,T \tag{31}$$

For convenience, the notation $\hat{\theta}$ will be omitted from now on.

Belief propagation provides an efficient way to perform exact inference on tree-structured graphs such as HMM. First recall that a Bayesian network conditioned on evidence induces a Gibbs distribution defined over *reduced* factors. In the case of posterior inference in HMM, the graph reduced by the evidence $\mathbf{x}_{1:T}$ is simply a chain:

$$\mathbf{z}_1 \ \text{———}\ \mathbf{z}_2 \ \text{———}\ \cdots \ \text{———}\ \mathbf{z}_T$$

where the factors, *i.e.*, initial clique potentials are defined as

$$\psi_1(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1), \tag{32}$$

$$\psi_t(\mathbf{z}_{t-1},\mathbf{z}_t) = p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t), \quad t=2,\ldots,T \tag{33}$$

$$\psi_{T+1}(\mathbf{z}_T) = 1, \tag{34}$$

so that the posterior is the following Gibbs distribution:

$$p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T}) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot \tilde{p}(\mathbf{z}_{1:T}), \tag{35}$$

$$\tilde{p}(\mathbf{z}_{1:T}) = \psi_1(\mathbf{z}_1) \cdot \prod_{t=2}^{T} \psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t) \cdot \psi_{T+1}(\mathbf{z}_T), \tag{36}$$

$$Z(\mathbf{x}_{1:T}) = \sum_{\mathbf{z}_{1:T}} \tilde{p}(\mathbf{z}_{1:T}). \tag{37}$$

The junction tree of the reduced graph is again a chain with clique size at most two:

$$\mathbf{z}_1 \overset{\mathbf{z}_1}{\text{———}} \mathbf{z}_1\mathbf{z}_2 \overset{\mathbf{z}_2}{\text{———}} \mathbf{z}_2\mathbf{z}_3 \overset{\mathbf{z}_3}{\text{———}} \cdots \overset{\mathbf{z}_{T-1}}{\text{———}} \mathbf{z}_{T-1}\mathbf{z}_T \overset{\mathbf{z}_T}{\text{———}} \mathbf{z}_T$$

The chain structure makes message passing particularly straightforward: there are only two types of messages, forward and backward.

The forward sum-product messages are

$$\alpha(\mathbf{z}_1) = \psi_1(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1), \tag{38}$$

$$\alpha(\mathbf{z}_t) = \sum_{\mathbf{z}_{t-1}} \psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t)\alpha(\mathbf{z}_{t-1}) \tag{39}$$

$$= p(\mathbf{x}_t|\mathbf{z}_t) \sum_{\mathbf{z}_{t-1}} p(\mathbf{z}_t|\mathbf{z}_{t-1})\alpha(\mathbf{z}_{t-1}). \quad t = 2, \ldots, T \tag{40}$$

The backward sum-product messages are

$$\beta(\mathbf{z}_{t-1}) = \sum_{\mathbf{z}_t} \psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t)\beta(\mathbf{z}_t) \tag{41}$$

$$= \sum_{\mathbf{z}_t} p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t)\beta(\mathbf{z}_t), \quad t = 2, \ldots, T \tag{42}$$

$$\beta(\mathbf{z}_T) = \psi_{T+1}(\mathbf{z}_T) = 1. \tag{43}$$

Clique beliefs are product of initial clique potential and incoming messages:

$$c(\mathbf{z}_1) = \psi_1(\mathbf{z}_1)\beta(\mathbf{z}_1) = \alpha(\mathbf{z}_1)\beta(\mathbf{z}_1), \tag{44}$$

$$c(\mathbf{z}_{t-1}, \mathbf{z}_t) = \psi_t(\mathbf{z}_{t-1}, \mathbf{z}_t)\alpha(\mathbf{z}_{t-1})\beta(\mathbf{z}_t) \tag{45}$$

$$= p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t)\alpha(\mathbf{z}_{t-1})\beta(\mathbf{z}_t), \quad t = 2, \ldots, T \tag{46}$$

$$c(\mathbf{z}_T) = \psi_{T+1}(\mathbf{z}_T)\alpha(\mathbf{z}_T) = \alpha(\mathbf{z}_T). \tag{47}$$

Sepset beliefs are product of corresponding messages:

$$s(\mathbf{z}_t) = \alpha(\mathbf{z}_t)\beta(\mathbf{z}_t). \quad t = 1, \ldots, T \tag{48}$$

At calibration, the beliefs represent unnormalized marginals:

$$c(\mathbf{z}_1) = \tilde{p}(\mathbf{z}_1), \tag{49}$$
$$c(\mathbf{z}_{t-1}, \mathbf{z}_t) = \tilde{p}(\mathbf{z}_{t-1}, \mathbf{z}_t), \quad t = 2, \dots, T \tag{50}$$
$$c(\mathbf{z}_T) = \tilde{p}(\mathbf{z}_T), \tag{51}$$
$$s(\mathbf{z}_t) = \tilde{p}(\mathbf{z}_t), \quad t = 1, \dots, T \tag{52}$$

which means the partition function $Z(\mathbf{x}_{1:T})$ can be computed by summing any of the beliefs:

$$\sum_{\mathbf{z}_1} c(\mathbf{z}_1) = \sum_{\mathbf{z}_{t-1}, \mathbf{z}_t} c(\mathbf{z}_{t-1}, \mathbf{z}_t) = \sum_{\mathbf{z}_T} c(\mathbf{z}_T) = \sum_{\mathbf{z}_t} s(\mathbf{z}_t) = \sum_{\mathbf{z}_{1:T}} \tilde{p}(\mathbf{z}_{1:T}) = Z(\mathbf{x}_{1:T}). \tag{53}$$

Finally, the marginal queries can be computed by normalizing the beliefs:

$$\gamma(\mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot s(\mathbf{z}_t), \tag{54}$$
$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot c(\mathbf{z}_{t-1}, \mathbf{z}_t), \tag{55}$$

It is not a coincidence that the messages are named α and β: the above belief propagation procedure is precisely the forward-backward algorithm in terms of (α, β)-recursion.

6 Scaling $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ messages

Implemented as presented above, the (α, β)-recursion is likely to encounter numerical instability due to repeated multiplication of small values. One way to mitigate the numerical issue is to scale (α, β) messages at each step t, so that the scaled values are always in some appropriate range, while not affecting the inference result for (γ, ξ).

Recall that the forward message is in fact a joint distribution

$$\alpha(\mathbf{z}_t) = p(\mathbf{x}_{1:t}, \mathbf{z}_t). \tag{56}$$

Define scaled messages by re-normalizing α w.r.t. $\mathbf{z}_t$:

$$\hat{\alpha}(\mathbf{z}_t) \coloneqq \frac{1}{Z(\mathbf{x}_{1:t})} \cdot \alpha(\mathbf{z}_t), \tag{57}$$
$$Z(\mathbf{x}_{1:t}) = \sum_{\mathbf{z}_t} \alpha(\mathbf{z}_t). \tag{58}$$

Furthermore, define

$$r_1 \coloneqq Z(\mathbf{x}_1), \tag{59}$$
$$r_t \coloneqq \frac{Z(\mathbf{x}_{1:t})}{Z(\mathbf{x}_{1:t-1})}. \quad t = 2, \dots, T \tag{60}$$

Notice that $Z(\mathbf{x}_{1:t}) = r_1 \cdots r_t$, hence

$$\hat{\alpha}(\mathbf{z}_t) = \frac{1}{r_1 \cdots r_t} \cdot \alpha(\mathbf{z}_t). \tag{61}$$

Plugging $\hat{\alpha}$ into forward messages, the new $\hat{\alpha}$-recursion is

$$\hat{\alpha}(\mathbf{z}_1) = \frac{1}{r_1} \cdot \underbrace{p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1)}_{\tilde{\alpha}(\mathbf{z}_1)} \tag{62}$$

$$\hat{\alpha}(\mathbf{z}_t) = \frac{1}{r_t} \cdot \underbrace{p(\mathbf{x}_t|\mathbf{z}_t) \sum_{\mathbf{z}_{t-1}} p(\mathbf{z}_t|\mathbf{z}_{t-1})\hat{\alpha}(\mathbf{z}_{t-1})}_{\tilde{\alpha}(\mathbf{z}_t)}. \quad t = 2, \ldots, T \tag{63}$$

Since $\hat{\alpha}$ is normalized, each r_t serves as the normalizing constant:

$$r_t = \sum_{\mathbf{z}_t} \tilde{\alpha}(\mathbf{z}_t). \tag{64}$$

Now switch focus to β. In order to make the inference for (γ, ξ) invariant of scaling, β has to be scaled in a way that counteracts the scaling on α. Plugging $\hat{\alpha}$ into the marginal queries,

$$\gamma(\mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot r_1 \cdots r_t \cdot \hat{\alpha}(\mathbf{z}_t)\beta(\mathbf{z}_t), \tag{65}$$

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{Z(\mathbf{x}_{1:T})} \cdot p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t) \cdot r_1 \cdots r_{t-1} \cdot \hat{\alpha}(\mathbf{z}_{t-1})\beta(\mathbf{z}_t). \tag{66}$$

Since $Z(\mathbf{x}_{1:T}) = r_1 \ldots r_T$, a natural scaling scheme for β is

$$\hat{\beta}(\mathbf{z}_{t-1}) \coloneqq \frac{1}{r_t \cdots r_T} \cdot \beta(\mathbf{z}_{t-1}), \quad t = 2, \ldots, T \tag{67}$$

$$\hat{\beta}(\mathbf{z}_T) \coloneqq \beta(\mathbf{z}_T), \tag{68}$$

which simplifies the expression for marginals (γ, ξ) to

$$\gamma(\mathbf{z}_t) = \hat{\alpha}(\mathbf{z}_t)\hat{\beta}(\mathbf{z}_t), \tag{69}$$

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{r_t} \cdot p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t)\hat{\alpha}(\mathbf{z}_{t-1})\hat{\beta}(\mathbf{z}_t). \tag{70}$$

The new $\hat{\beta}$-recursion can be obtained by plugging $\hat{\beta}$ into backward messages:

$$\hat{\beta}(\mathbf{z}_{t-1}) = \frac{1}{r_t} \cdot \sum_{\mathbf{z}_t} p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t)\hat{\beta}(\mathbf{z}_t), \quad t = 2, \ldots, T \tag{71}$$

$$\hat{\beta}(\mathbf{z}_T) = 1. \tag{72}$$

In other words, $\hat{\beta}(\mathbf{z}_{t-1})$ is scaled by $1/r_t$, the normalizer of $\hat{\alpha}(\mathbf{z}_t)$.

The full algorithm is summarized below.

Algorithm 1 Exact inference for (γ, ξ)

1. Scaled forward message for $t = 1$:

$$\tilde{\alpha}(\mathbf{z}_1) = p(\mathbf{z}_1)p(\mathbf{x}_1|\mathbf{z}_1) \tag{73}$$

$$r_1 = \sum_{\mathbf{z}_1} \tilde{\alpha}(\mathbf{z}_1) \tag{74}$$

$$\hat{\alpha}(\mathbf{z}_1) = \frac{1}{r_1} \cdot \tilde{\alpha}(\mathbf{z}_1) \tag{75}$$

2. Scaled forward message for $t = 2, \ldots, T$:

$$\tilde{\alpha}(\mathbf{z}_t) = p(\mathbf{x}_t|\mathbf{z}_t) \sum_{\mathbf{z}_{t-1}} p(\mathbf{z}_t|\mathbf{z}_{t-1})\hat{\alpha}(\mathbf{z}_{t-1}) \tag{76}$$

$$r_t = \sum_{\mathbf{z}_t} \tilde{\alpha}(\mathbf{z}_t) \tag{77}$$

$$\hat{\alpha}(\mathbf{z}_t) = \frac{1}{r_t} \cdot \tilde{\alpha}(\mathbf{z}_t) \tag{78}$$

3. Scaled backward message for $t = T + 1$:

$$\hat{\beta}(\mathbf{z}_T) = 1 \tag{79}$$

4. Scaled backward message for $t = T, \ldots, 2$:

$$\hat{\beta}(\mathbf{z}_{t-1}) = \frac{1}{r_t} \cdot \sum_{\mathbf{z}_t} p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t)\hat{\beta}(\mathbf{z}_t) \tag{80}$$

5. Unary marginal for $t = 1, \ldots, T$:

$$\gamma(\mathbf{z}_t) = \hat{\alpha}(\mathbf{z}_t)\hat{\beta}(\mathbf{z}_t) \tag{81}$$

6. Pairwise marginal for $t = 2, \ldots, T$:

$$\xi(\mathbf{z}_{t-1}, \mathbf{z}_t) = \frac{1}{r_t} \cdot p(\mathbf{z}_t|\mathbf{z}_{t-1})p(\mathbf{x}_t|\mathbf{z}_t)\hat{\alpha}(\mathbf{z}_{t-1})\hat{\beta}(\mathbf{z}_t) \tag{82}$$