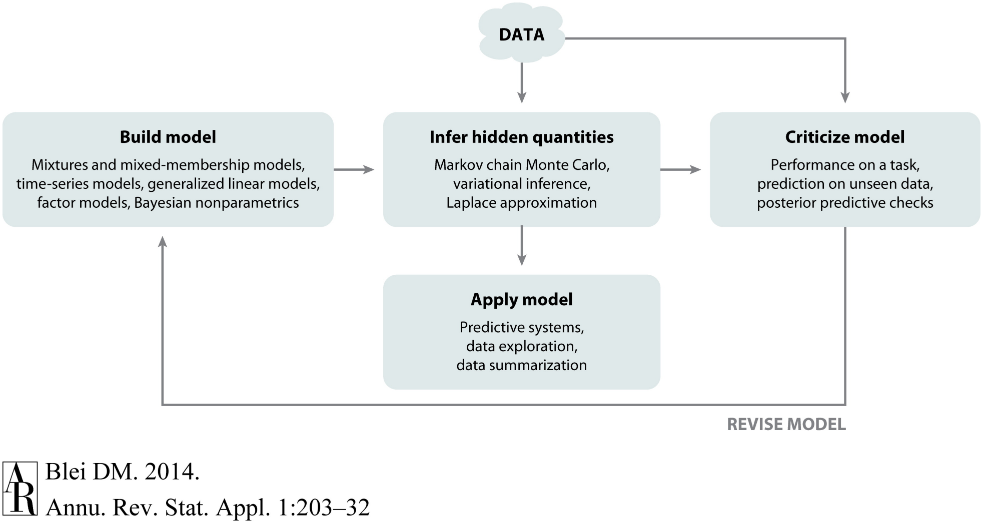
STATS271/371: Applied Bayesian Statistics

Hidden Markov Models (HMMs) and Message Passing Algorithms (Part II)

Scott Linderman

May 17, 2021

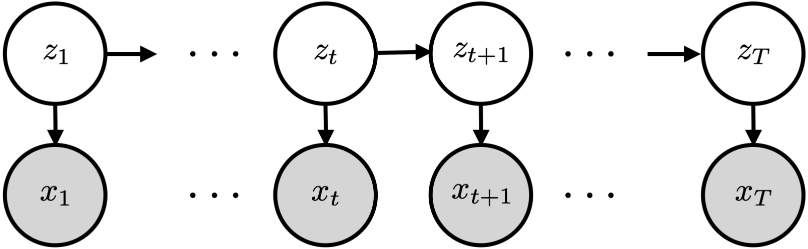
Box's Loop



Blei, *Ann. Rev. Stat. App.* 2014.

Hidden Markov Models

Hidden Markov Models (HMMs) assume a particular factorization of the joint distribution on latent states (z_t) and observations ($\boldsymbol{x}_t$). The graphical model looks like this:



This graphical model says that the joint distribution factors as,

$$p(z_{1:T}, \boldsymbol{x}_{1:T}) = p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid z_t). \tag{1}$$

We call this an HMM because $p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1})$ is a Markov chain.

Hidden Markov Models II

We are interested in questions like:

- What is the *posterior marginal* distribution $p(z_t \mid \boldsymbol{x}_{1:T})$?
- What is the *posterior pairwise marginal* distribution $p(z_t, z_{t+1} \mid \boldsymbol{x}_{1:T})$?
- What is the *posterior mode* $z^\star_{1:T} = \arg\max p(z_{1:T} \mid \boldsymbol{x}_{1:T})$?
- What is the *predictive distribution* of $p(z_{T+1} \mid \boldsymbol{x}_{1:T})$?

State space models

Note that nothing above assumes that z_t is a discrete random variable!

HMM's are a special case of more general **state space models** with discrete states.

State space models assume the same graphical model but allow for arbitrary types of latent states.

For example, suppose that $\boldsymbol{z}_t \in \mathbb{R}^P$ are continuous valued latent states and that,

$$p(\boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) \tag{2}$$

$$= \mathcal{N}(\boldsymbol{z}_1 \mid \boldsymbol{b}_1, \boldsymbol{Q}_1) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{b}, \boldsymbol{Q}) \tag{3}$$

This is called a **linear dynamical system** with Gaussian noise.

Message passing in HMMs

In the HMM with discrete states, we showed how to compute posterior marginal distributions using message passing,

$$
\begin{aligned}
p(z_t \mid \boldsymbol{x}_{1:T}) &\propto \sum_{z_1} \cdots \sum_{z_{t-1}} \sum_{z_{t+1}} \cdots \sum_{z_T} p(z_{1:T}, \boldsymbol{x}_{1:T}) && (4) \\
&= \alpha_t(z_t)\, p(\boldsymbol{x}_t \mid z_t)\, \beta_t(z_t) && (5)
\end{aligned}
$$

where the *forward and backward messages* are defined recursively

$$
\begin{aligned}
\alpha_t(z_t) &= \sum_{z_{t-1}} p(z_t \mid z_{t-1})\, p(\boldsymbol{x}_{t-1} \mid z_{t-1})\, \alpha_{t-1}(z_{t-1}) && (6) \\
\beta_t(z_t) &= \sum_{z_{t+1}} p(z_{t+1} \mid z_t)\, p(\boldsymbol{x}_{t+1} \mid z_{t+1})\, \beta_{t+1}(z_{t+1}) && (7)
\end{aligned}
$$

The initial conditions are $\alpha_1(z_1) = p(z_1)$ and $\beta_T(z_T) = 1$.

What do the forward messages tell us?

The forward messages are equivalent to,

$$\alpha_t(z_t) = \sum_{z_1} \cdots \sum_{z_{t-1}} p(z_{1:t}, \boldsymbol{x}_{1:t-1}) \tag{8}$$

$$p(z_t, \boldsymbol{x}_{1:t-1}). \tag{9}$$

The normalized message is the *predictive distribution*,

$$\frac{\alpha_t(z_t)}{\sum_{z_t'} \alpha_t(z_t')} = \frac{p(z_t, \boldsymbol{x}_{1:t-1})}{\sum_{z_t'} p(z_t', \boldsymbol{x}_{1:t-1})} = \frac{p(z_t, \boldsymbol{x}_{1:t-1})}{p(\boldsymbol{x}_{1:t-1})} = p(z_t \mid \boldsymbol{x}_{1:t-1}), \tag{10}$$

The final normalizing constant yields the marginal likelihood, $\sum_{z_T} \alpha_T(z_T) = p(\boldsymbol{x}_{1:T})$.

Message passing in state space models

The same recursive algorithm applies (in theory) to any state space model with the same factorization, but the sums are replaced with integrals,

$$p(z_t \mid \boldsymbol{x}_{1:T}) \propto \int \mathrm{d}z_1 \cdots \int \mathrm{d}z_{t-1} \int \mathrm{d}z_{t+1} \cdots \int \mathrm{d}z_T \, p(z_{1:T}, \boldsymbol{x}_{1:T}) \tag{11}$$

$$= \alpha_t(z_t) \, p(\boldsymbol{x}_t \mid z_t) \, \beta_t(z_t) \tag{12}$$

where the *forward and backward messages* are defined recursively

$$\alpha_t(z_t) = \int p(z_t \mid z_{t-1}) \, p(\boldsymbol{x}_{t-1} \mid z_{t-1}) \, \alpha_{t-1}(z_{t-1}) \, \mathrm{d}z_{t-1} \tag{13}$$

$$\beta_t(z_t) = \int p(z_{t+1} \mid z_t) \, p(\boldsymbol{x}_{t+1} \mid z_{t+1}) \, \beta_{t+1}(z_{t+1}) \, \mathrm{d}z_{t+1} \tag{14}$$

The initial conditions are $\alpha_1(z_1) = p(z_1)$ and $\beta_T(z_T) = 1$.

Message passing in a linear dynamical system

Exercise: Consider an LDS with Gaussian noise and assume that $p(\boldsymbol{x}_t \mid \boldsymbol{z}_t) = \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})$. Derive the forward message $\alpha_t(\boldsymbol{z}_t)$ under the inductive hypothesis that $\alpha_{t-1}(\boldsymbol{z}_{t-1}) \propto \mathcal{N}(\boldsymbol{z}_{t-1} \mid \boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1})$.

Message passing in nonlinear dynamical systems

Question: What if $p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(z_t \mid f(\boldsymbol{z}_{t-1}), \boldsymbol{Q})$ for some nonlinear function f?

Sequential Monte Carlo

Recall that the forward messages are proportional to the predictive distributions $p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t-1})$. We can view the forward recursions as an expectation,

$$\alpha_t(\boldsymbol{z}_t) = \int p(\boldsymbol{z}_t \mid z_{t-1})\, p(\boldsymbol{x}_{t-1} \mid \boldsymbol{z}_{t-1})\, \alpha_{t-1}(\boldsymbol{z}_{t-1})\, \mathrm{d}\boldsymbol{z}_{t-1} \tag{15}$$

$$\propto \mathbb{E}_{\boldsymbol{z}_{t-1} \sim p(\boldsymbol{z}_{t-1} \mid \boldsymbol{x}_{1:t-2})} \left[p(\boldsymbol{z}_t \mid z_{t-1})\, p(\boldsymbol{x}_{t-1} \mid \boldsymbol{z}_{t-1}) \right] \tag{16}$$

One natural idea is to approximate this expectation with Monte Carlo,

$$\hat{\alpha}_t(\boldsymbol{z}_t) \approx \frac{1}{S} \sum_{s=1}^{S} \left[w_{t-1}^{(s)}\, p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}^{(s)}) \right] \tag{17}$$

where we have defined the **weights** $w_{t-1}^{(s)} \triangleq p(\boldsymbol{x}_{t-1} \mid \boldsymbol{z}_{t-1}^{(s)})$.

How do we sample $\boldsymbol{z}_{t-1}^{(s)} \overset{\text{iid}}{\sim} p(\boldsymbol{z}_{t-1} \mid \boldsymbol{x}_{1:t-2})$? Let's sample the normalized $\hat{\alpha}_{t-1}(\boldsymbol{z}_{t-1})$ instead!

Sequential Monte Carlo II

The normalizing constant is,

$$\int \hat{\alpha}_{t-1}(\boldsymbol{z}_{t-1})\, \mathrm{d}\boldsymbol{z}_{t-1} = \frac{1}{S}\sum_{s=1}^{S} w_{t-2}^{(s)} \int p(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_{t-2}^{(s)})\, \mathrm{d}\boldsymbol{z}_{t-1} = \frac{1}{S}\sum_{s=1}^{S} w_{t-2}^{(s)}. \tag{18}$$

Use this to define the *normalized forward message* (i.e. the Monte Carlo estimate of the predictive distribution) is,

$$\bar{\alpha}_{t-1}(\boldsymbol{z}_{t-1}) \triangleq \frac{\hat{\alpha}_{t-1}(\boldsymbol{z}_{t-1})}{\int \hat{\alpha}_{t-1}(\boldsymbol{z}'_{t-1})\, \mathrm{d}\boldsymbol{z}'_{t-1}} = \sum_{s=1}^{S} \bar{w}_{t-2}^{(s)}\, p(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_{t-2}^{(s)}) \tag{19}$$

where $\bar{w}_{t-2}^{(s)} = \frac{w_{t-2}^{(s)}}{\sum_{s'} w_{t-2}^{(s')}}$ is the normalized weight of sample $\boldsymbol{z}_{t-2}^{(s)}$.

The normalized forward message is just a mixture distribution with weights $\bar{w}_{t-2}^{(s)}$!

Putting it all together

Combining the above, we have the following algorithm for the forward pass:

1. Let $\bar{\alpha}_1(\boldsymbol{z}_1) = p(z_1)$
2. For $t = 1, \ldots, T$:
   a. Sample $\boldsymbol{z}_t^{(s)} \overset{\text{iid}}{\sim} \bar{\alpha}_t(\boldsymbol{z}_t)$ for $s = 1, \ldots, S$
   b. Compute weights $w_t^{(s)} = p(\boldsymbol{x}_t \mid \boldsymbol{z}_t^{(s)})$ and normalize $\bar{w}_t^{(s)} = w_t^{(s)} / \sum_{s'} w_t^{(s')}$.
   c. Compute normalized forward message $\bar{\alpha}_{t+1}(\boldsymbol{z}_{t+1}) = \sum_{s=1}^{S} \bar{w}_t^{(s)} p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t^{(s)})$.

This is called **sequential Monte Carlo** (SMC) using the model dynamics as the proposal.

Note that Step 2a can **resample** the same $\boldsymbol{z}_{t-1}^{(s)}$ multiple times according to its weight.

Question: How can you approximate the marginal likelihood $p(\boldsymbol{x}_{1:T})$ using the weights? *Hint: look back to Slide 7.*

Generalizations

- Instead of sampling $\bar{\alpha}_t(\boldsymbol{z}_t)$, we could have sampled with a **proposal distribution** $r(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}^{(s)})$ instead and corrected for it by defining the weights to be,

$$w_t^{(s)} = \frac{p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}^{(s)})\, p(\boldsymbol{x}_t \mid \boldsymbol{z}_t)}{r(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}^{(s)})} \tag{20}$$

  Moreover, the proposal distribution can "look ahead" to future data $\boldsymbol{x}_t$.

References I