Linear Dynamical Systems and State Space Models

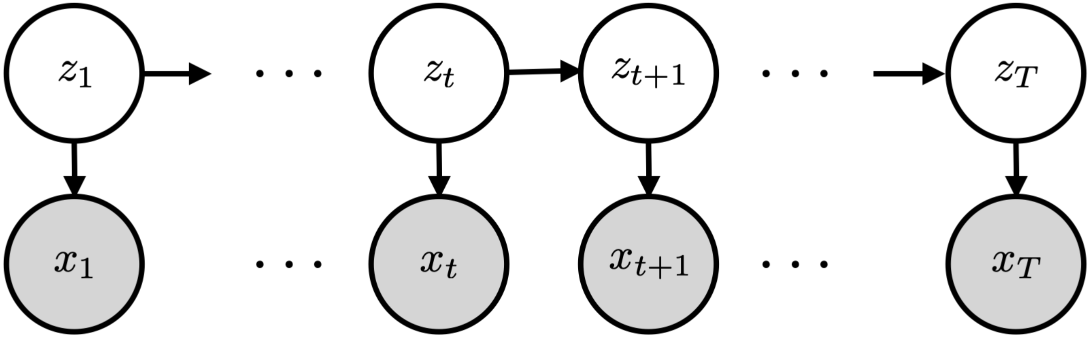
STATS 305C: Applied Statistics

Scott Linderman

May 11, 2022

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 are the *predictive distributions* of $p(z_{t+1} \mid \boldsymbol{x}_{1:t})$?
- 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* $z^{\star}_{1:T} = \arg\max p(z_{1:T} \mid \boldsymbol{x}_{1:T})$?
- How can we *sample the posterior* $p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T})$ of an HMM?
- What is the *marginal likelihood* $p(\boldsymbol{x}_{1:T})$?
- How can we *learn the parameters* of an HMM?

Question: Why might these sound like hard problems?

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 $\boldsymbol{z}_t \in \mathbb{R}^D$ 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 Gaussian **linear dynamical system** (LDS).

Stability of Gaussian linear dynamical systems

Question: What is the asymptotic mean of a Gaussian LDS, $\lim_{t \to \infty} \mathbb{E}[\boldsymbol{z}_t]$?

Question: When is a Gaussian LDS stable? I.e. when is the asymptotic mean finite?

$\mu_\infty = \lim_{t \to \infty} \mathbb{E}_t [z_t]$

$\mu_\infty = A \mu_\infty + b$

$\mu_\infty = (I - A)^{-1} b$

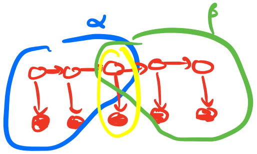
$|\mathrm{eig}(A)| < 1$

EE263: LDS

Message passing in HMMs

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

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



where the *forward and backward messages* are defined recursively

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

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,

$$p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T}) \propto \int \mathrm{d}\boldsymbol{z}_1 \cdots \int \mathrm{d}\boldsymbol{z}_{t-1} \int \mathrm{d}\boldsymbol{z}_{t+1} \cdots \int \mathrm{d}\boldsymbol{z}_T \, p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}) \tag{11}$$

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

where the *forward and backward messages* are defined recursively

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

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

The initial conditions are $\alpha_1(\boldsymbol{z}_1) = p(\boldsymbol{z}_1)$ and $\beta_T(\boldsymbol{z}_T) \propto 1$.

Forward pass in a linear dynamical system

Consider an linear dynamical system (LDS) with Gaussian emissions,

$$p(\boldsymbol{x}_{1:T}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{z}_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) \quad \prod_{t=1}^{T} p(x_t \mid z_t) \tag{15}$$

$$= \mathcal{N}(\boldsymbol{z}_1 \mid \boldsymbol{b}_1, \boldsymbol{Q}_1) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{b}, \boldsymbol{Q}) \prod_{t=1}^{\mathcal{N}} (\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R}) \tag{16}$$

Let's derive the forward message $\alpha_{t+1}(\boldsymbol{z}_{t+1})$. Assume $\alpha_t(\boldsymbol{z}_t) \propto \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \Sigma_{t|t-1})$. [Induction]

$$\alpha_{t+1}(\boldsymbol{z}_{t+1}) = \int p(z_{t+1} \mid z_t)\, p(\boldsymbol{x}_t \mid z_t)\, \alpha_t(z_t)\, \mathrm{d}z_t \tag{17}$$

$$= \int \overset{\text{predict}}{\mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{b}, \boldsymbol{Q})}\, \overset{\text{update}}{\mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \Sigma_{t|t-1})}\, \mathrm{d}\boldsymbol{z}_t \tag{18}$$

The update step

The first step is the **update step**, where we **condition on** the emission $\boldsymbol{x}_t$,

Exercise: Expand the densities, collect terms, and complete the square to compute the mean $\boldsymbol{\mu}_{t|t}$ and covariance $\Sigma_{t|t}$ after the update step,

$$\mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \Sigma_{t|t-1}) \propto \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t}, \Sigma_{t|t}). \tag{19}$$

$$\propto \exp\left\{-\frac{1}{2}(x_t - Cz_t - d)^T R^{-1}(x_t - Cz_t - d) - \frac{1}{2}(z_t - \mu_{t|t-1})^T \Sigma_{t|t-1}^{-1}(z_t - \mu_{t|t-1})\right\}$$

$$\propto \exp\left\{-\frac{1}{2} z_t^T J_{t|t} z_t + h_{t|t}^T z_t\right\}$$

$$\mu_{t|t} = \mu_{t|t-1} + \Sigma_{t|t-1} C^T (C \Sigma_{t|t-1} C^T + R)^{-1} C \Sigma_{t|t-1} (x_t - C\mu_{t|t-1} - d)$$

$$J_{t|t} = C^T R^{-1} C + \Sigma_{t|t-1}^{-1}$$

$$h_{t|t} = C^T R^{-1}(x_t - d) + \Sigma_{t|t-1}^{-1}\mu_{t|t-1}$$

$$\mu_{t|t} = J_{t|t}^{-1} h_{t|t}$$

$$= (C^T R^{-1} C + \Sigma_{t|t-1})^{-1}\left[C^T R^{-1}(x_t - d) + \Sigma_{t|t-1}^{-1}\mu_{t|t-1}\right]$$

The update step II

Write the joint distribution,

$$p(\boldsymbol{z}_t, \boldsymbol{x}_t \mid \boldsymbol{x}_{1:t-1}) = \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1}) \tag{20}$$

$$= \mathcal{N}\left(\begin{bmatrix} \boldsymbol{z}_t \\ \boldsymbol{x}_t \end{bmatrix} \,\middle|\, \begin{bmatrix} \boldsymbol{\mu}_{t|t-1} \\ \boldsymbol{C}\boldsymbol{\mu}_{t|t-1} + \boldsymbol{d} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{t|t-1} & \boldsymbol{\Sigma}_{t|t-1}\boldsymbol{C}^\top \\ \boldsymbol{C}\boldsymbol{\Sigma}_{t|t-1} & \boldsymbol{C}\boldsymbol{\Sigma}_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R} \end{bmatrix} \right) \tag{21}$$

Since $(\boldsymbol{z}_t, \boldsymbol{x}_t)$ are jointly Gaussian, $\boldsymbol{z}_t$ must be conditionally Gaussian as well,

$$p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t}) = \mathcal{N}(\boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t}). \tag{22}$$

Exercise: Now use the **Schur complement** from Week 1 to solve for $\boldsymbol{\mu}_{t|t}$ and $\boldsymbol{\Sigma}_{t|t}$

matrix inversion lemma (2x)

$$\mu_{t|t} = \mu_{t|t-1} + \Sigma_{t|t-1} C^T (C \Sigma_{t|t-1} C^T + R)^{-1} (x_t - C\mu_{t|t-1} - d)$$

$$\mu_{t|t} = (C^T R^{-1} C + \Sigma_{t|t-1}^{-1})^{-1} \left[C^T R^{-1} (x_t - d) + \Sigma_{t|t-1}^{-1} \mu_{t|t-1} \right]$$

$$\Sigma_{t|t} = \Sigma_{t|t-1} - \Sigma_{t|t-1} C^T (C \Sigma_{t|t-1} C^T + R)^{-1} C \Sigma_{t|t-1}$$

The update step III

Exercise: Write $\mu_{t|t}$ and $\Sigma_{t|t}$ in terms of the **Kalman gain**,

$$\boldsymbol{K}_t = \Sigma_{t|t-1}\boldsymbol{C}^\top(\boldsymbol{C}\Sigma_{t|t-1}\boldsymbol{C}^\top + \boldsymbol{R})^{-1} \qquad \hat{x}_t = C\mu_{t|t-1} + d \tag{23}$$

What is the Kalman gain doing?

$$\mu_{t|t} = \mu_{t|t-1} + \Sigma_{t|t-1} C^T (C\Sigma_{t|t-1} C^T + R)^{-1} (x_t - C\mu_{t|t-1} - d)$$

$$= \underbrace{\mu_{t|t-1}}_{\text{predicted mean}} + K_t \underbrace{(x_t - \hat{x}_t)}_{\text{residual}}$$

The predict step

The predict step yields $p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t}) = \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})$. To complete the forward pass, we need the **predict step**,

$$\alpha_{t+1}(\boldsymbol{z}_{t+1}) = \int p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)\, p(\boldsymbol{x}_t \mid \boldsymbol{z}_t)\, \alpha_t(\boldsymbol{z}_t)\, \mathrm{d}\boldsymbol{z}_t \tag{24}$$

$$= \int \mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{A}\boldsymbol{z}_t + \boldsymbol{b}, \boldsymbol{Q})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t}, \boldsymbol{\Sigma}_{t|t})\, \mathrm{d}\boldsymbol{z}_t \tag{25}$$

$$= \mathcal{N}(\boldsymbol{z}_{t+1} \mid \boldsymbol{\mu}_{t+1|t}, \boldsymbol{\Sigma}_{t+1|t}) \tag{26}$$

Exercise: Solve for the mean $\boldsymbol{\mu}_{t+1|t}$ and covariance $\boldsymbol{\Sigma}_{t+1|t}$ after the predict step.

$\mu_{t+1|t} = A\mu_{t|t} + b$

$\Sigma_{t+1|t} = Q + A\Sigma_{t|t}A^T$

Completing the recursions

That wraps up the recursions! All that's left is the base case, which comes from the initial state distribution,

$$\boldsymbol{\mu}_{1|0} = \boldsymbol{b}_1 \quad \text{and} \quad \boldsymbol{\Sigma}_{1|0} = \boldsymbol{Q}_1. \tag{27}$$

$$p(x_1) = \mathcal{N}(x_1 \mid b_1, Q_1)$$

Computing the marginal likelihood

Like in the discrete HMM, we can compute the marginal likelihood along the forward pass.

$$p(\boldsymbol{x}_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid \boldsymbol{x}_{1:t-1}) \tag{28}$$

$$= \prod_{t=1}^{T} \int p(\boldsymbol{x}_t \mid \boldsymbol{z}_t)\, p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t-1})\, \mathrm{d}\boldsymbol{z}_t \tag{29}$$

$$= \prod_{t=1}^{T} \int \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R})\, \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{\mu}_{t|t-1}, \boldsymbol{\Sigma}_{t|t-1})\, \mathrm{d}\boldsymbol{z}_t \tag{30}$$

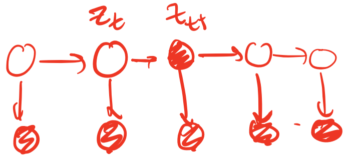
Exercise: Obtain a closed form expression for the integrals.

$$= \prod_t \mathcal{N}(x_t \mid C\mu_{t|t-1} + d,\ R + C\Sigma_{t|t-1}C^T)$$

Computing the smoothing distributions

- The forward pass gives us the filtering distributions $p(\mathbf{z}_t \mid \mathbf{x}_{1:t})$. How can we compute the smoothing distributions, $p(\mathbf{z}_t \mid \mathbf{x}_{1:T})$?
- In the discrete HMM we essentially ran the *same algorithm in reverse* to get the backward messages, starting from $\beta_T(\mathbf{z}_T) \propto 1$.
- We can do the same sort of thing here, but it's a bit funky because we need to start with an improper Gaussian distribution $\beta_T(\mathbf{z}_T) \propto \mathcal{N}(\mathbf{0}, \infty \mathbf{I})$.
- Instead, we'll derive an alternative way of computing the smoothing distributions.

Bayesian Smoothing



Note: $\boldsymbol{z}_t$ is conditionally independent of $\boldsymbol{x}_{t+1:T}$ given $\boldsymbol{z}_{t+1}$, so

$p(z_t | x_{1:T})$ $\neq$

$$
\begin{aligned}
p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:T}) &= p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:t}) && (31)\\
&= \frac{p(\boldsymbol{z}_t, \boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})}{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})} && (32)\\
&= \frac{p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})\, p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)}{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})} && (33)
\end{aligned}
$$

Question: what rules did we apply in each of these steps?

Bayesian Smoothing II

Now we can write the joint distribution as,

$$p(\boldsymbol{z}_t, \boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T}) = p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t+1}, \boldsymbol{x}_{1:T}) p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T}) \tag{34}$$

$$= \frac{p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t})\, p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)\, p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T})}{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})}. \tag{35}$$

Marginalizing over $\boldsymbol{z}_{t+1}$ gives us,

$$p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T}) = p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:t}) \int \frac{\overbrace{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{z}_t)}^{\text{model}}\, \overbrace{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:T})}^{\text{smoothing}}}{\underbrace{p(\boldsymbol{z}_{t+1} \mid \boldsymbol{x}_{1:t})}_{\text{filtering}}} \mathrm{d}\boldsymbol{z}_{t+1} \tag{36}$$

Question: Can we compute each of these terms?

The Rauch-Tung-Striebel Smoother, aka Kalman Smoother

These recursions apply to any Markovian state space model. For the special case of a Gaussian linear dynamical system, the smoothing distributions are again Gaussians,

$$p(\boldsymbol{z}_t \mid \boldsymbol{x}_{1:T}) = \mathcal{N}(\boldsymbol{z}_t \mid \underline{\underline{\boldsymbol{\mu}_{t|T}}}, \underline{\underline{\boldsymbol{\Sigma}_{t|T}}}) \tag{37}$$

where

$$\boldsymbol{\mu}_{t|T} = \boldsymbol{\mu}_{t|t} + \boldsymbol{G}_t(\boldsymbol{\mu}_{t+1|T} - \boldsymbol{\mu}_{t+1|t}) \tag{38}$$

$$\boldsymbol{\Sigma}_{t|T} = \boldsymbol{\Sigma}_{t|t} + \boldsymbol{G}_t(\boldsymbol{\Sigma}_{t+1|T} - \boldsymbol{\Sigma}_{t+1|t})\boldsymbol{G}_t^\top \tag{39}$$

$$\boldsymbol{G}_t \triangleq \boldsymbol{\Sigma}_{t|t}\boldsymbol{A}^\top \boldsymbol{\Sigma}_{t+1|t}^{-1}. \tag{40}$$

This is called the **Rauch-Tung-Striebel (RTS) smoother** or the **Kalman smoother**.

$\mu_{t|t} \longrightarrow \mu_{T|T}$

Kalman smoothing in information form

So far we've worked with the *mean parameters* $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, but working with *natural parameters* $\boldsymbol{J}$ and $\boldsymbol{h}$ offers another perspective.

Let's go back to the basics,

$$p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}) \propto p(\boldsymbol{z}_{1:T}, \boldsymbol{x}_{1:T}) \tag{41}$$

$$= p(\boldsymbol{z}_1) \prod_{t=2}^{T} p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) \prod_{t=1}^{T} p(\boldsymbol{x}_t \mid \boldsymbol{z}_t) \tag{42}$$

$$= \mathcal{N}(\boldsymbol{z}_1 \mid \boldsymbol{b}_1, \boldsymbol{Q}_1) \prod_{t=2}^{T} \mathcal{N}(\boldsymbol{z}_t \mid \boldsymbol{A}\boldsymbol{z}_{t-1} + \boldsymbol{b}, \boldsymbol{Q}) \prod_{t=1}^{T} \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{C}\boldsymbol{z}_t + \boldsymbol{d}, \boldsymbol{R}) \tag{43}$$

Kalman smoothing in information form II

Expand the Gaussian densities,

$$p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}) \propto \exp\Big\{ -\frac{1}{2}(\boldsymbol{z}_1 - \boldsymbol{b}_1)^\top \boldsymbol{Q}_1^{-1}(\boldsymbol{z}_1 - \boldsymbol{b}_1) \tag{44}$$

$$-\frac{1}{2}\sum_{t=2}^{T}(\boldsymbol{z}_t - \boldsymbol{A}\boldsymbol{z}_{t-1} - \boldsymbol{b})^\top \boldsymbol{Q}^{-1}(\boldsymbol{z}_t - \boldsymbol{A}\boldsymbol{z}_{t-1} - \boldsymbol{b}) \tag{45}$$

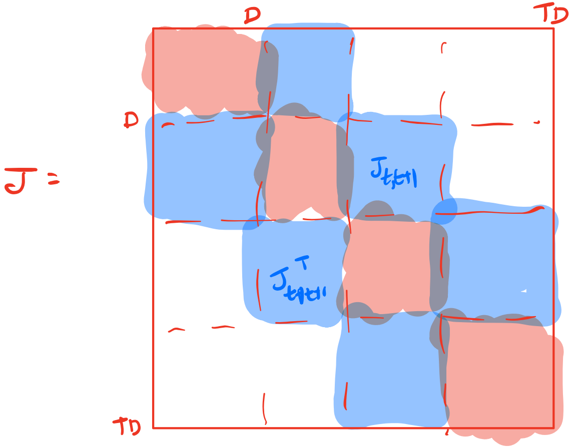
$$-\frac{1}{2}\sum_{t=1}^{T}(\boldsymbol{x}_t - \boldsymbol{C}\boldsymbol{z}_t - \boldsymbol{d})^\top \boldsymbol{R}^{-1}(\boldsymbol{x}_t - \boldsymbol{C}\boldsymbol{z}_t - \boldsymbol{d})\Big\} \tag{46}$$

This is a giant quadratic expression in $\boldsymbol{z}_{1:T}$; i.e. a multivariate normal distribution on $\mathbb{R}^{TD}$.

We can write it in terms of its natural parameters $\boldsymbol{J} \in \mathbb{R}^{TD \times TD}$ and $\boldsymbol{h} \in \mathbb{R}^{TD}$

Kalman smoothing in information form III

Question: Which entries in J are nonzero?



block tridiagonal

$T(D^2) + (T-1)D^2$

Duality between message passing and sparse linear algebra

Recall that to get mean from the natural parameters we have,

$$p(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}_{1:T}) = \mathcal{N}(\boldsymbol{z}_{1:T} \mid \boldsymbol{J}^{-1}\boldsymbol{h}, \boldsymbol{J}^{-1}). \tag{47}$$

In other words, the posterior mean is the solution of a linear system $\boldsymbol{J}^{-1}\boldsymbol{h}$.

Typically, this would cost $O((TD)^3)$, but since $\boldsymbol{J}$ is block-tridiagonal (or more generally, banded), we can compute it in only $O(TD^3)$ time.

The algorithm for solving this sparse linear system is essentially the same as the message passing algorithm we derived today.

Message passing in nonlinear dynamical systems

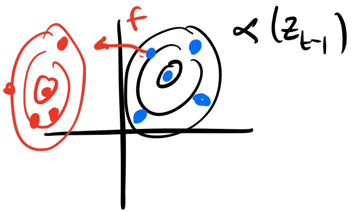
Question: What would you do if you were given a nonlinear model,
$p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) = \mathcal{N}(z_t \mid f(\boldsymbol{z}_{t-1}), \boldsymbol{Q})$?

$$\alpha(z_{t-1}) = \mathcal{N}(z_{t-1} \mid \mu_{t-1|t-2}, \Sigma_{t-1|t-2})$$

$$f(z_{t-1}) \approx f(\mu_{t-1|t-2}) + \nabla f(\mu_{t-1|t-2})(z_{t-1} - \mu_{t-1|t-2}) + O((z_t - \hat{z}_{t-1})^2)$$

$\Rightarrow$ extended Kalman filter

c.f. unscented Kalman filter



- neural network
  - $\rightarrow$ Deep Kalman filter
  - $\rightarrow$ Structured VAEs

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,

$$\alpha_t(\boldsymbol{z}_t) = \int p(\boldsymbol{z}_t \mid z_{t-1})\, p(\boldsymbol{x}_{t-1} \mid \boldsymbol{z}_{t-1})\, \alpha_{t-1}(\boldsymbol{z}_{t-1})\, \mathrm{d}\boldsymbol{z}_{t-1} \tag{48}$$

$$\propto \mathbb{E}_{\boldsymbol{z}_{t-1} \sim p(\boldsymbol{z}_{t-1} \mid \boldsymbol{x}_{1:t-2})} \left[p(\boldsymbol{z}_t \mid z_{t-1})\, p(\boldsymbol{x}_{t-1} \mid \boldsymbol{z}_{t-1})\right] \tag{49}$$

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

$$\hat{\alpha}_t(\boldsymbol{z}_t) \approx \frac{1}{S} \sum_{s=1}^{S} \left[w_{t-1}^{(s)}\, p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}^{(s)}) \right] \tag{50}$$

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,

$$\int \hat{\alpha}_{t-1}(\boldsymbol{z}_{t-1})\,\mathrm{d}\boldsymbol{z}_{t-1} = \frac{1}{S}\sum_{s=1}^{S} w_{t-2}^{(s)} \int p(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_{t-2}^{(s)})\,\mathrm{d}\boldsymbol{z}_{t-1} = \frac{1}{S}\sum_{s=1}^{S} w_{t-2}^{(s)}. \tag{51}$$

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

$$\bar{\alpha}_{t-1}(\boldsymbol{z}_{t-1}) \triangleq \frac{\hat{\alpha}_{t-1}(\boldsymbol{z}_{t-1})}{\int \hat{\alpha}_{t-1}(\boldsymbol{z}'_{t-1})\,\mathrm{d}\boldsymbol{z}'_{t-1}} = \sum_{s=1}^{S} \bar{w}_{t-2}^{(s)}\, p(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_{t-2}^{(s)}) \tag{52}$$

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,

$$w_t^{(s)} = \frac{p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}^{(s)})\, p(\boldsymbol{x}_t \mid \boldsymbol{z}_t)}{r(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}^{(s)})} \tag{53}$$

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

References I