Machine Learning Methods for Neural Data Analysis

Fitting (Switching) Linear Dynamical Systems

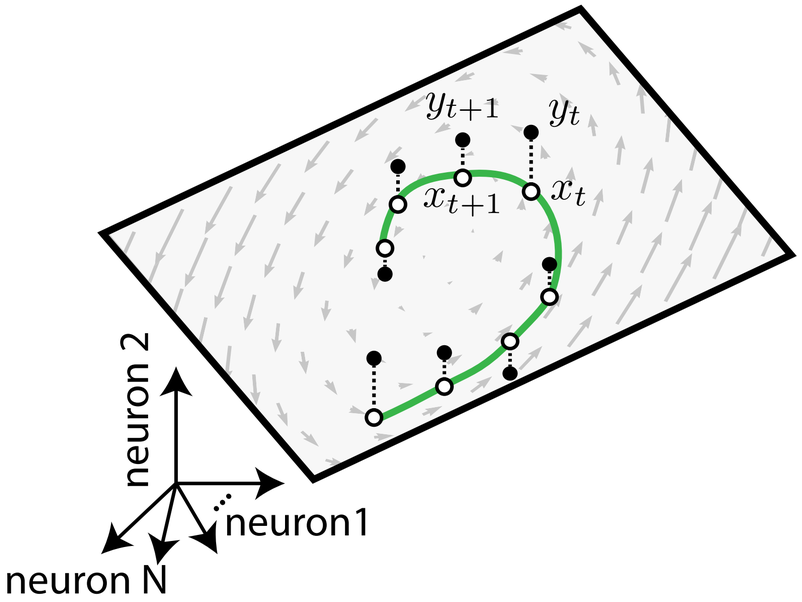
Scott Linderman

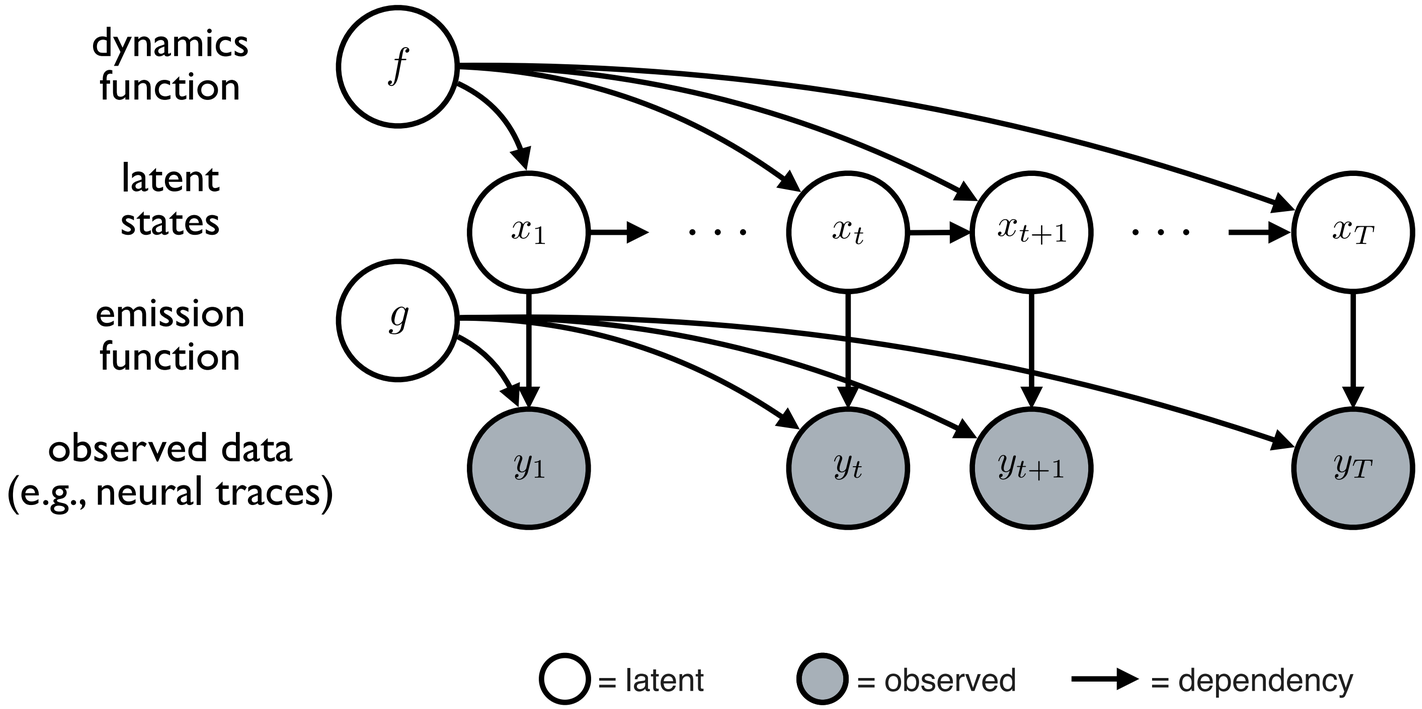
STATS 220/320 (NBIO220, CS339N).

Agenda

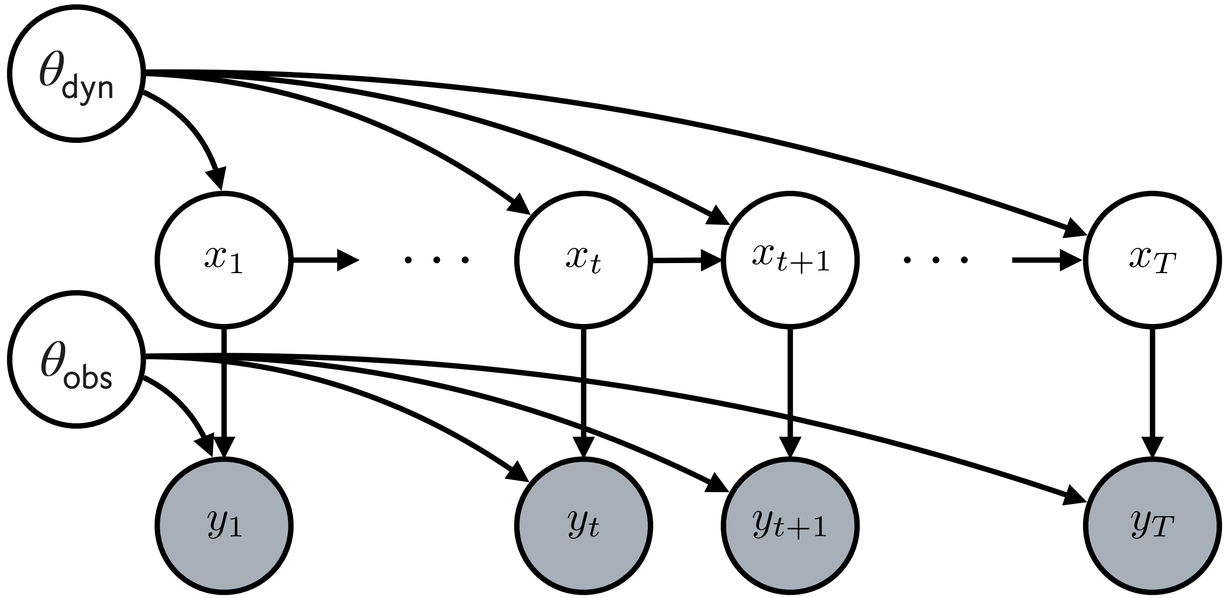
- EM for Gaussian LDS
- The Kalman filter and smoother
- Variational EM (vEM) for SLDS
- Coordinate Ascent VI (CAVI)

Probabilistic state space models





Gaussian Linear Dynamical Systems (LDS)



A Gaussian LDS has **linear Gaussian dynamics**,

$$x_{t+1} \sim \mathcal{N}(Ax_t + b, Q)$$

parameterized by $\theta_{\text{dyn}} = (A, b, Q)$,

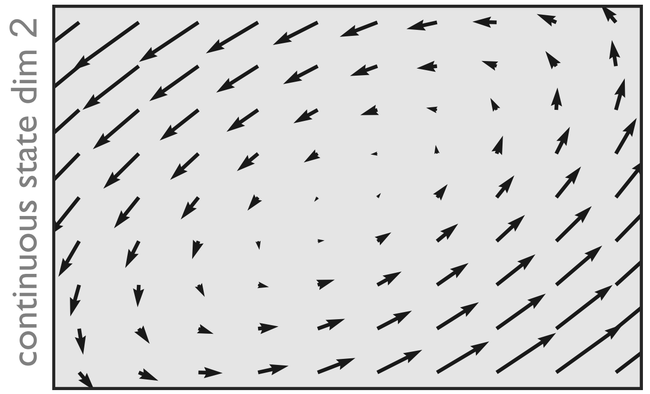
and **linear Gaussian emissions,**

$$y_t \sim \mathcal{N}(Cx_t + d, R)$$

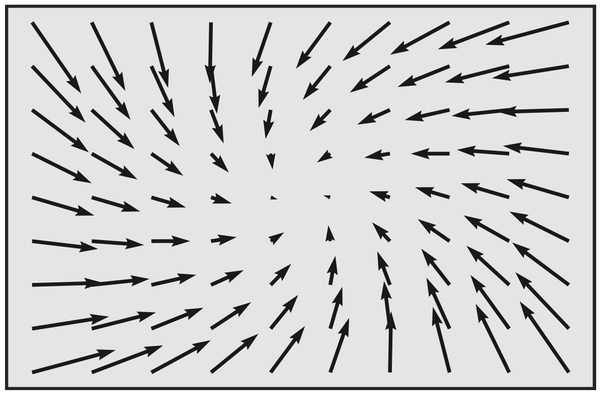
parameterized by $\theta_{\text{obs}} = (C, d, R)$.

What can linear models do?

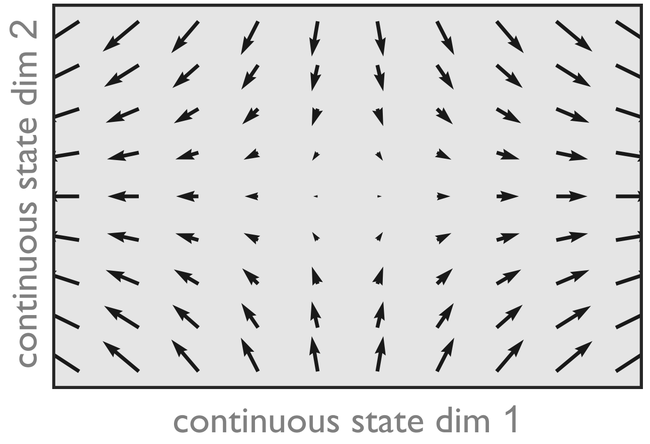
rotational dynamics (e.g., motor control)



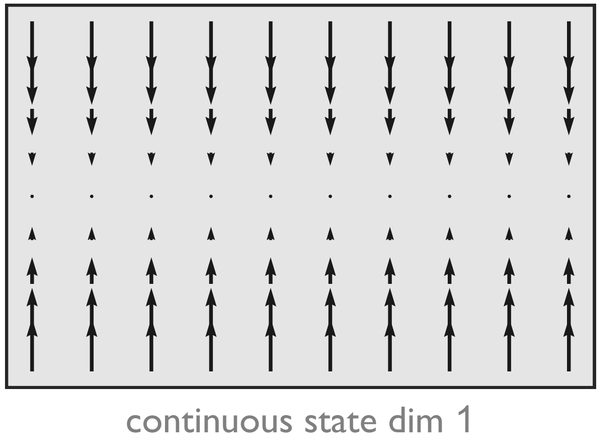
point attractor (e.g., memory)



saddle point (e.g., winner-take-all)



line attractor (e.g., integration)



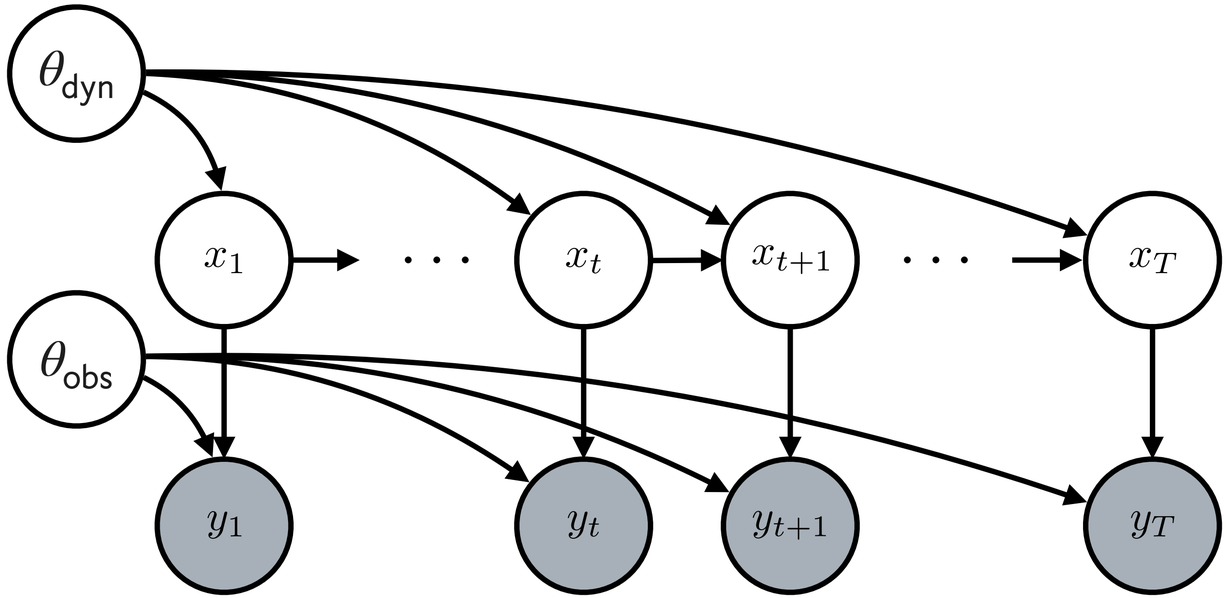
A lot! E.g., the motifs from before were all linear models, $f(x) = Ax + b$.

Moreover, linear systems are interpretable.

We can find analytical solutions for:

- fixed points and stability
- dynamics along eigenmodes
- posterior distribution over latent states (with the Kalman filter/smoother)
- optimal control (with dynamic programming)

Expectation-Maximization for Gaussian LDS



How can we "fit" an LDS? Like other latent variable models, we can use EM!

E-step: Compute the posterior distribution over latent states,

$$q(x_{1:T}) \leftarrow p(x_{1:T} \mid y_{1:T}; \theta).$$

(Really, we only need expected sufficient statistics.)

M-step: Update the parameters,

$$\theta \leftarrow \arg\max_{\theta} \mathbb{E}_q \left[\log p(x_{1:T}, y_{1:T}; \theta)\right],$$

by maximizing the expected log probability.

Message passing in probabilistic state space models

Recall our **message passing** algorithm for **hidden Markov models.** The same recursive algorithm applies (in theory) to **any state space model**, but the **sums are replaced with integrals**,

$$p(x_t \mid y_{1:T}) \propto \int \mathrm{d}x_1 \cdots \int \mathrm{d}x_{t-1} \int \mathrm{d}x_{t+1} \cdots \int \mathrm{d}x_T \, p(x_{1:T}, y_{1:T})$$
$$= \alpha_t(x_t)\, p(y_t \mid x_t)\, \beta_t(x_t)$$

where the **forward and backward messages** are defined recursively

$$\alpha_t(x_t) = \int p(x_t \mid x_{t-1})\, p(y_{t-1} \mid x_{t-1})\, \alpha_{t-1}(x_{t-1}) \mathrm{d}x_{t-1}$$

$$\beta_t(x_t) = \int p(x_{t+1} \mid x_t)\, p(y_{t+1} \mid x_{t+1})\, \beta_{t+1}(x_{t+1}) \mathrm{d}x_{t+1}$$

The **initial conditions** are $\alpha_1(x_1) = p(x_1)$ and $\beta_T(x_T) \propto 1$.

Computing the forward messages with the Kalman filter

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

$$p(y_{1:T}, x_{1:T}) = \mathcal{N}(x_1 \mid m, Q) \prod_{t=2}^{T} \mathcal{N}(x_t \mid Ax_{t-1} + b, Q) \prod_{t=1}^{T} \mathcal{N}(y_t \mid Cx_t + d, R)$$

To derive the forward message, make the **inductive hypothesis** that $\alpha_t(x_t) \propto \mathcal{N}(x_t \mid \mu_{t|t-1}, \Sigma_{t|t-1})$. Then,

$$\begin{aligned}
\alpha_{t+1}(x_{t+1}) &= \int p(x_{t+1} \mid x_t)\, p(y_t \mid x_t)\, \alpha_t(x_t)\, \mathrm{d}x_t \\
&= \int \mathcal{N}(x_{t+1} \mid Ax_t + b, Q)\, \mathcal{N}(y_t \mid Cx_t + d, R)\, \mathcal{N}(x_t \mid \mu_{t|t-1}, \Sigma_{t|t-1})\, \mathrm{d}x_t
\end{aligned}$$

Since the integrand is a **product of linear Gaussian terms**, the result is a Gaussian of the hypothesized form!

This algorithm is called the **Kalman filter**, and it is one of the most important algorithms in signal processing.

Computing expected sufficient statistics under the posterior

The **Kalman smoother** combines the forward messages with a backward pass to compute the posterior marginals.

Since the model is constructed with **linear Gaussian dependencies**, the posterior marginals are all Gaussian distributions.

The Kalman smoother returns several **expected sufficient statistics** under the posterior:

- Posterior means, $\mathbb{E}_q[x_t]$
- Posterior covariances, $\mathrm{Cov}_q[x_t]$
- Posterior cross-covariances, $\mathbb{E}_q[x_t x_{t+1}^\top]$
- From the first two, we can also compute $\mathbb{E}_q[x_t x_t^\top]$

where $q(x_{1:T}) = p(x_{1:T} \mid y_{1:T}; \theta)$ is the posterior distribution.

The M-step for a Gaussian LDS

The **M-step** for a Gaussian LDS solves for parameters that maximize the expected log joint probability,

$$\theta \leftarrow \arg\max_{\theta} \mathbb{E}_q \left[\log p(x_{1:T}, y_{1:T}; \theta)\right]$$

where $q(x_{1:T}) = p(x_{1:T} \mid y_{1:T}; \theta_{\text{old}})$. These updates only require the expected sufficient statistics.

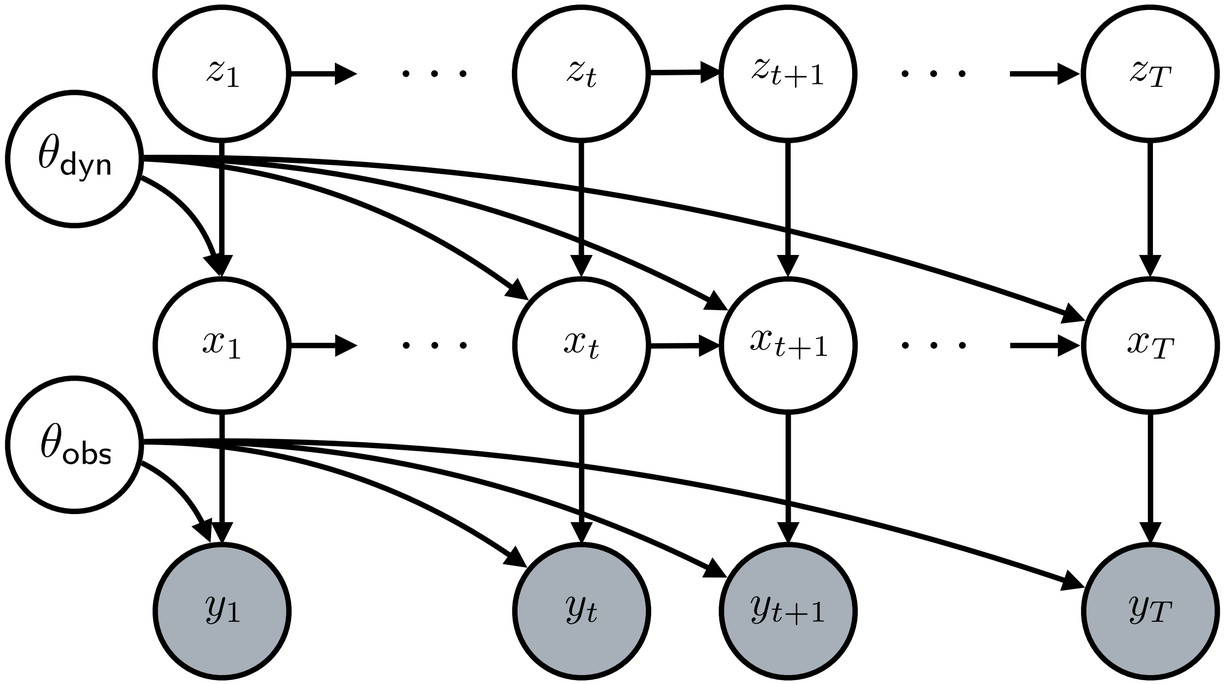
For example, to update the **dynamics matrix** (assuming $b = 0$ for simplicity) we find,

$$\begin{aligned} A &\leftarrow \arg\max_{A} \mathbb{E}_q \left[\sum_{t=1}^{T-1} \log \mathcal{N}(x_{t+1} \mid Ax_t, Q)\right] \\ &= \arg\min_{A} \sum_{t=1}^{T-1} \mathbb{E}_q \left[(x_{t+1} - Ax_t)^\top Q^{-1}(x_{t+1} - Ax_t)\right] \\ &= \left(\sum_{t=1}^{T-1} \mathbb{E}_q[x_{t+1}x_t^\top]\right)\left(\sum_{t=1}^{T-1} \mathbb{E}_q[x_t x_t^\top]\right)^{-1}. \end{aligned}$$

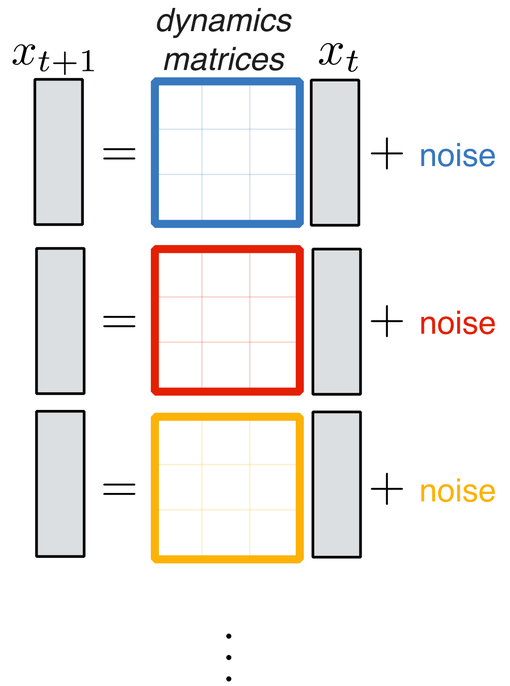
Agenda

- EM for Gaussian LDS
- The Kalman filter and smoother
- **Variational EM (vEM) for SLDS**
- Coordinate Ascent VI (CAVI)

Switching linear dynamical systems (SLDS)



Different **linear dynamics** in each discrete state



Exact EM for a Gaussian SLDS

E-step: Update the posterior over latent variables,

$$q(z, x) \leftarrow p(z, x \mid y, \theta) = \frac{p(z, x, y \mid \theta)}{p(y \mid \theta)}.$$

M-step: Update the parameters,

$$\theta \leftarrow \arg\max \mathbb{E}_{q(z,x)} \left[\log p(z, x, y \mid \theta)\right]$$

As before, we **only need certain expectations** under q,

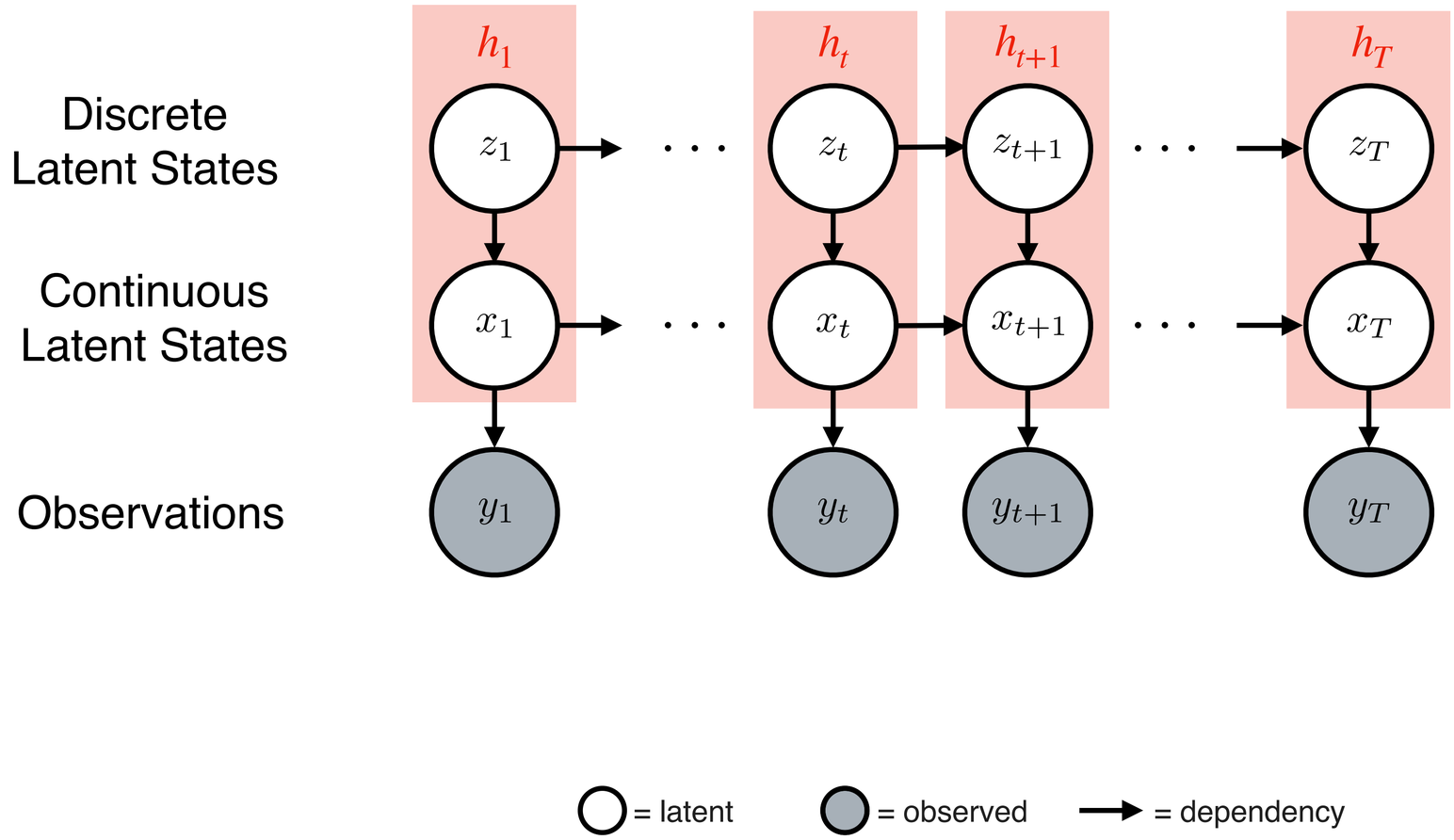
$$\mathbb{E}_{q(z,x)} \left[\mathbb{I}[z_t = k]\right], \quad \mathbb{E}_{q(z,x)} \left[\mathbb{I}[z_t = k] x_t\right], \quad \mathbb{E}_{q(z,x)} \left[\mathbb{I}[z_t = k] x_t x_t^\top\right], \quad \mathbb{E}_{q(z,x)} \left[\mathbb{I}[z_t = k] x_t x_{t+1}^\top\right],$$

Unfortunately, **computing the posterior expectations is a lot harder now!**

Naively applying the message passing recursions for and SLDS

We can think of the SLDS as a **hybrid** state space model.

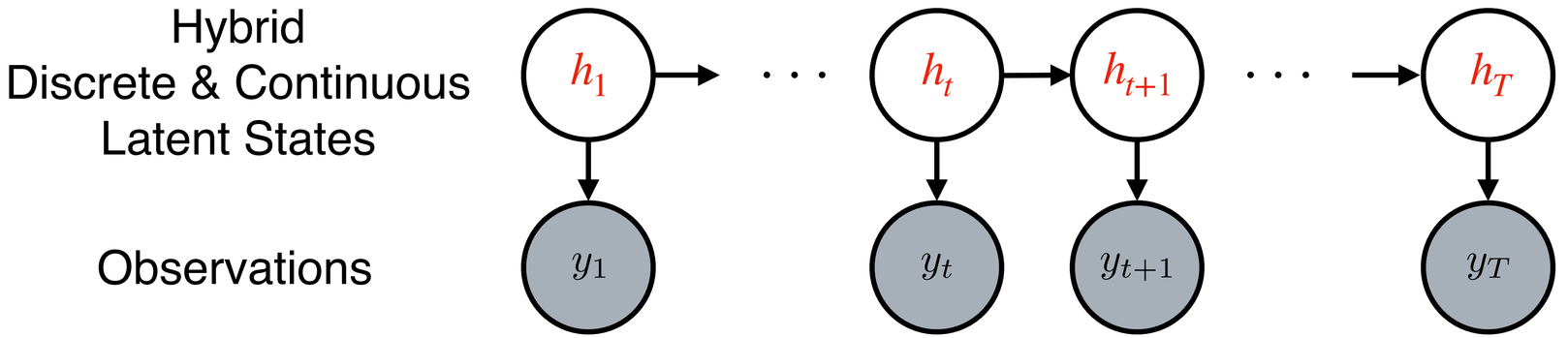
Let $h_t = (z_t, x_t)$ denote the hybrid discrete & continuous latent state.



Naively applying the message passing recursions for and SLDS

We can think of the SLDS as a **hybrid** state space model.

Let $h_t = (z_t, x_t)$ denote the hybrid discrete & continuous latent state.



The messages are **mixtures of Gaussians**,

$$\alpha_t(z_t, x_t) = \sum_{z_{t-1}} \int p(x_t, z_t \mid x_{t-1}, z_{t-1})\, p(y_t \mid x_t, z_t)\, \alpha_{t-1}(z_{t-1}, x_{t-1})\, \mathrm{d}x_{t-1}$$

The last message, $\alpha_T(h_T)$ has K^T mixture components!

Variational Inference

Since the **exact posterior is intractable**, let's try to approximate it instead.

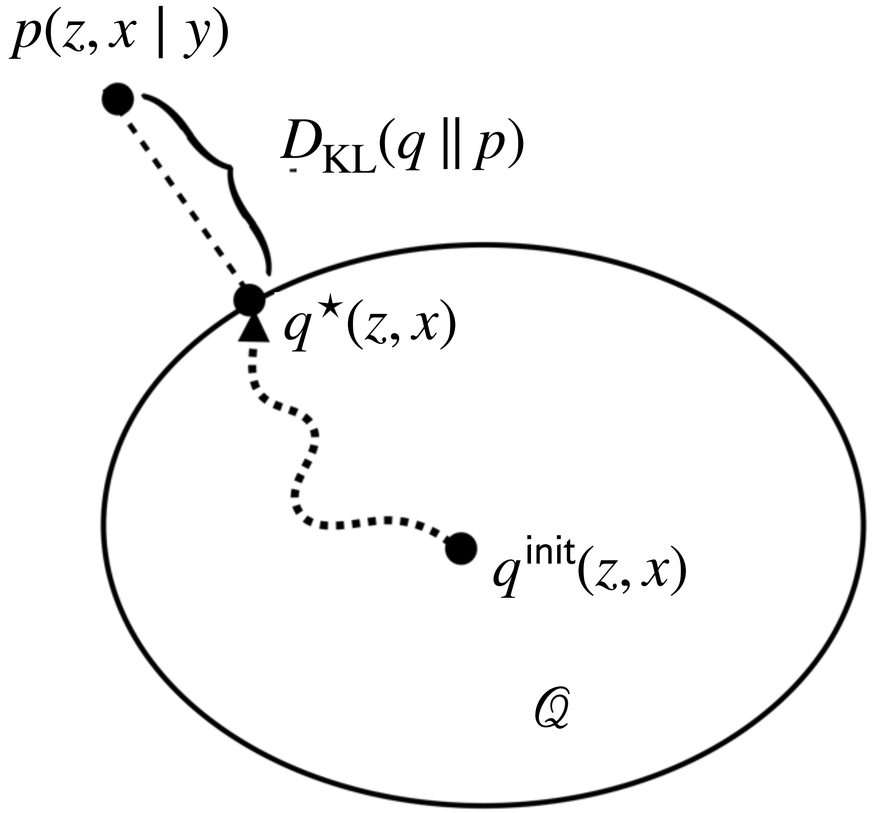
One way to do so is with **variational inference**: *Find an approximate posterior, $q(z, x)$, that is as close as possible to the true posterior, $p(z, x \mid y)$.*

We constrain $q(z, x)$ to belong to a **variational family** of *simple* distributions, $\mathcal{Q}$.

We typically measure closeness with the **Kullback-Leibler (KL) divergence**.

The KL divergence is zero iff $q = p$.

Question: what if $\mathcal{Q}$ is unconstrained, i.e., the set of all distributions?



Kullback-Leibler (KL) Divergence

The **Kullback-Leibler (KL) divergence** is a measure of how dissimilar two distributions are.

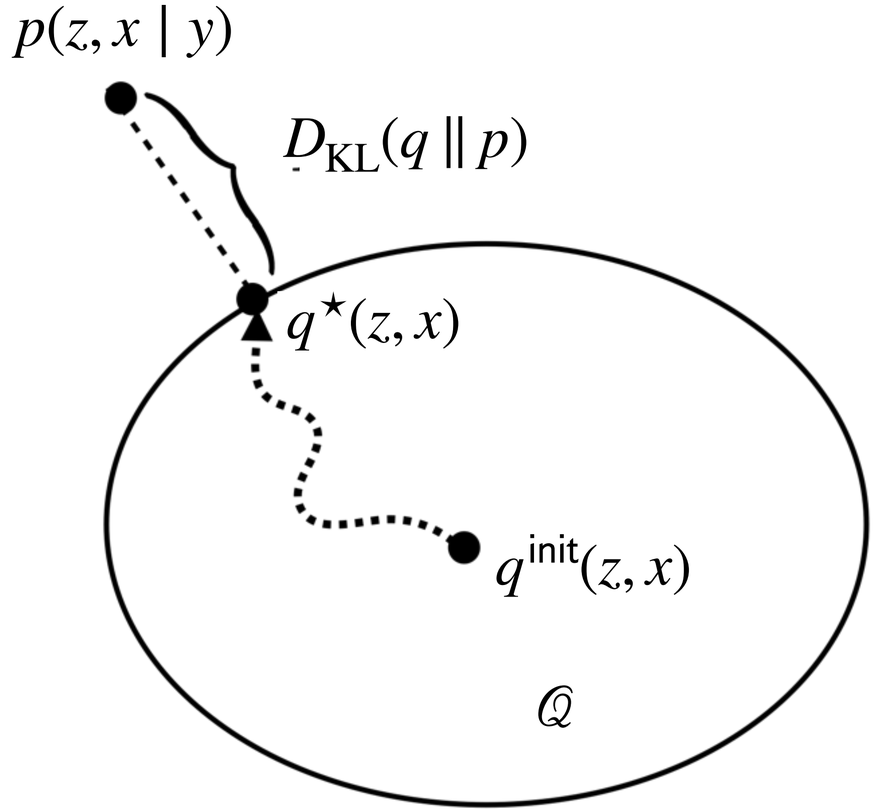
$$D_{\mathrm{KL}}(q \,\|\, p) = \mathbb{E}_{q(x)}\left[\log \frac{q(x)}{p(x)}\right]$$

It has several nice properties:

- The KL is divergence is non-negative.
- It is zero iff $q = p$.

But **it is not a distance!** In particular, it is not symmetric,

$$D_{\mathrm{KL}}(q \,\|\, p) \neq D_{\mathrm{KL}}(p \,\|\, q)$$



Kullback-Leibler (KL) Divergence

Example: minimizing the KL divergence between a single Gaussian and a mixture of Gaussians.

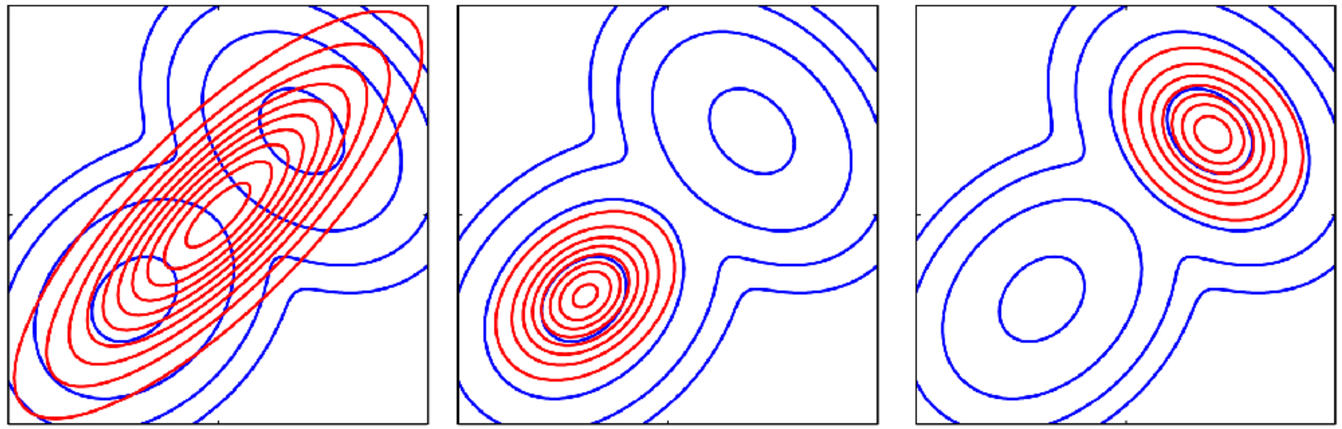


Figure 10.3 Another comparison of the two alternative forms for the Kullback-Leibler divergence. (a) The blue contours show a bimodal distribution $p(\mathbf{Z})$ given by a mixture of two Gaussians, and the red contours correspond to the single Gaussian distribution $q(\mathbf{Z})$ that best approximates $p(\mathbf{Z})$ in the sense of minimizing the Kullback-Leibler divergence $\mathrm{KL}(p\|q)$. (b) As in (a) but now the red contours correspond to a Gaussian distribution $q(\mathbf{Z})$ found by numerical minimization of the Kullback-Leibler divergence $\mathrm{KL}(q\|p)$. (c) As in (b) but showing a different local minimum of the Kullback-Leibler divergence.

Bishop Ch 10.

Coordinate Ascent Variational Inference (CAVI) with a Mean Field Posterior

The **mean-field approximation** to the posterior treats the variables as independent,

$$\mathcal{Q}_{\text{MF}} = \left\{ q : q(z, x) = q(z)\, q(x) \right\}$$

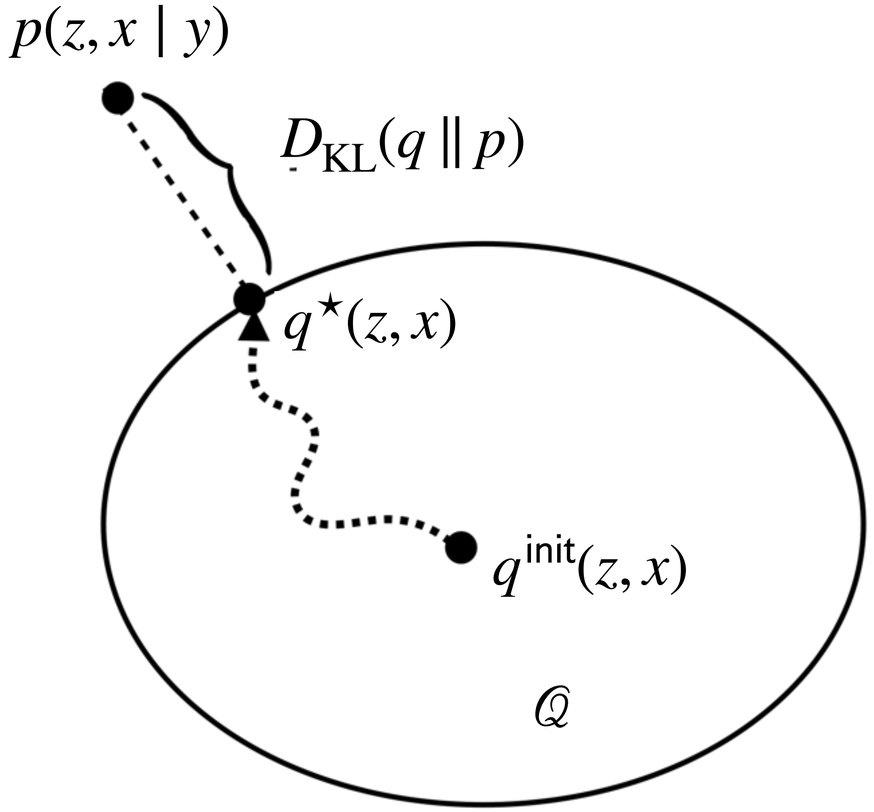
With this variational family, we can solve for $q^\star$ by **coordinate ascent.** For example,

$$q(z) \leftarrow \arg\min \text{KL}(q(z) \,\|\, \widetilde{p}(z))$$

where

$$\widetilde{p}(z) \propto \exp\left\{ \mathbb{E}_{q(x)} \left[\log p(z, x, y; \theta) \right] \right\}.$$

If $q(z)$ is otherwise unconstrained, the optimal coordinate update is to set $q(z) \leftarrow \widetilde{p}(z)$.



Coordinate Ascent Variational Inference (CAVI) for a Gaussian SLDS

For a Gaussian SLDS, $\widetilde{p}(z)$ takes the form of a **posterior distribution under an HMM**,

$$\begin{aligned}
\widetilde{p}(z) &\propto \exp\left\{\mathbb{E}_{q(x)}\left[\log p(z, x, y)\right]\right\} \\
&\propto \exp\left\{\mathbb{E}_{q(x)}\left[\log p(z_1) + \sum_{t=2}^{T} \log p(z_t \mid z_{t-1}) + \log p(x_1 \mid z_1) + \sum_{t=2}^{T} \log p(x_t \mid x_{t-1}, z_t)\right]\right\} \\
&\propto p(z_1) \prod_{t=2}^{T} p(z_t \mid z_{t-1}) \prod_{t=1}^{T} e^{l_t(z_t)}
\end{aligned}$$

where

$$l_t(z_t) = \mathbb{E}_{q(x)}\left[\log p(x_t \mid x_{t-1}, z_t)\right] = \mathbb{E}_{q(x)}\left[\log \mathcal{N}(x_t \mid A_{z_t} x_{t-1} + b_{z_t}, Q_{z_t})\right]$$

is the (expected) log likelihood associated with state z_t.

Constraining the form of the factors

Sometimes, we further **constrain the functional form** of the variational factors.

For example, we could constrain the posterior over z to be a delta function on $z^\star$

$$q(z) = \delta_{z^\star}(z)$$

Then updating $q(z)$ amounts to finding the mode of $\widetilde{p}(z)$,

$$q(z) = \arg\min_q \; D_{\mathrm{KL}}(q(z) \,\|\, \widetilde{p}(z)) \qquad \Longleftrightarrow \qquad z^\star = \arg\max_z \widetilde{p}(z)$$

Since $\widetilde{p}(z)$ is the posterior of an HMM, we can find the mode with the **Viterbi algorithm**.

Coordinate Ascent Variational Inference (CAVI) for a Gaussian SLDS

By symmetry, $\widetilde{p}(x)$ takes the form of a **posterior distribution under an LDS**,

$$\begin{aligned}\widetilde{p}(x) &\propto \exp\left\{\mathbb{E}_{q(z)}\left[\log p(z, x, y)\right]\right\} \\ &\propto \exp\left\{\mathbb{E}_{q(z)}\left[\log p(x_1 \mid z_1) + \sum_{t=2}^{T} \log p(x_t \mid x_{t-1}, z_t) + \sum_{t=1}^{T} \log p(y_t \mid x_t)\right]\right\}\end{aligned}$$

This expression simplifies nicely when you work with the natural parameters of the Gaussian distribution.

Of course, when $q(z) = \delta_{z^\star}(z)$, this expression simplifies even further,

$$\widetilde{p}(x) \propto p(x_1 \mid z_1^\star) \prod_{t=2}^{T} p(x_t \mid x_{t-1}, z_t^\star) \prod_{t=1}^{T} p(y_t \mid x_t).$$

When $q(x)$ is otherwise unconstrained, the optimal update is to set $q(x) \leftarrow \widetilde{p}(x)$.

Variational EM for a Gaussian SLDS

Variational E-step: *Approximate* the posterior over latent variables,

$$q(z, x) \leftarrow \arg\min_{\mathcal{Q}} D_{\mathrm{KL}}(q(z, x) \,\|\, p(z, x \mid y; \theta))$$

using a variational family of our choice (e.g., a mean field family).

Now, the **expectations under q are tractable,**

$$\mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k]\right], \quad \mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k]x_t\right], \quad \mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k]x_t x_t^\top\right], \quad \mathbb{E}_{q(z,x)}\left[\mathbb{I}[z_t = k]x_t x_{t+1}^\top\right],$$

M-step: Update the parameters,

$$\theta \leftarrow \arg\max \mathbb{E}_{q(z,x)}\left[\log p(z, x, y \mid \theta)\right]$$

For a Gaussian SLDS, these updates admit closed form solutions.

Conclusion

- Switching LDS combine ARHMMs and LDS to get the best of both worlds.
- They approximate nonlinear dynamical systems by switching between linear dynamical states.
- However, posterior inference is harder because the posterior has exponentially many modes.
- Variational EM is a natural generalization of EM to more complex latent variable models like this:
    - Simply replace the E-step with a variational approximation.

Further Reading

- Barber, David. 2012. Bayesian Reasoning and Machine Learning. Cambridge University Press. Chapter 25.

- Linderman, Scott W., Matthew J. Johnson, Andrew C. Miller, Ryan P. Adams, David M. Blei, and Liam Paninski. 2017. "Bayesian Learning and Inference in Recurrent Switching Linear Dynamical Systems." In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS).