A) Linear Dynamical Systems (LDS)

At timestep $t=1, \ldots, T$, let:

$z_t \in \mathbb{R}^m$ be the state variable

(e.g. arm position and velocity)

$x_t \in \mathbb{R}^d$ be the observation

(e.g. spike count vector)
State model:

\[
\begin{align*}
Z_t \mid Z_{t-1} &\sim N(A Z_{t-1}, Q) \quad (1) \\
Z_1 &\sim N(\pi, V)
\end{align*}
\]

Observation model:

\[
X_t \mid Z_t \sim N(C Z_t, R) \quad (2)
\]

Model parameters are \( \theta = \{A, Q, \pi, V, C, R\} \)

What are the dimensions of each of these parameters?

Graphical model:
State model describes how state evolves over time.

Observation model describes how observation relates to the state.

State model uses a Markov assumption:

\[
P(\mathbf{z}_1, \ldots, \mathbf{z}_T) = P(\mathbf{z}_1) P(\mathbf{z}_2 | \mathbf{z}_1) \cdots P(\mathbf{z}_T | \mathbf{z}_1, \ldots, \mathbf{z}_{T-1})
\]

\[
= P(\mathbf{z}_1) \prod_{t=2}^{T} P(\mathbf{z}_t | \mathbf{z}_{t-1}) \quad \text{Markov assumption}
\]

B) Training phase

Goal: Estimate the model parameters \( \theta = \{ A, Q, \pi, V, C, R \} \) from the training data.

If the values of the state variables \( \mathbf{z}_t \) are unknown during training, use EM algorithm (unsupervised learning).

Maximize \( P(\{\mathbf{x}\} | \theta) \) w.r.t. \( \theta \).
Here, we will consider the simpler case, where the $z_t$ are known during training (supervised learning).

Maximize $P(\{x_t, \{z_t\} | \theta\}$ w.r.t. $\theta$.

For decoding arm trajectories from neural activity, the $z_t$ (arm states) are typically known during training.

$$P(\{x_t, \{z_t\} | \theta\) = P(z_1) \prod_{t=2}^{T} P(z_t | z_{t-1}) \prod_{t=1}^{T} P(x_t | z_t)$$

Let

$$\mathcal{L}(\theta) = \log P(\{x_t, \{z_t\} | \theta\)$$

$$= \log P(z_1) + \sum_{t=2}^{T} \log P(z_t | z_{t-1}) + \sum_{t=1}^{T} \log P(x_t | z_t)$$

$$= -\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |V| - \frac{1}{2} (z_1 - \Pi)^T V^{-1} (z_1 - \Pi)$$

$$+ \sum_{t=2}^{T} \left( -\frac{M}{2} \log(2\pi) - \frac{1}{2} \log |Q| - \frac{1}{2} (z_t - A z_{t-1})^T Q^{-1} (z_t - A z_{t-1}) \right)$$

$$+ \sum_{t=1}^{T} \left( -\frac{D}{2} \log(2\pi) - \frac{1}{2} \log |R| - \frac{1}{2} (x_t - C z_t)^T R^{-1} (x_t - C z_t) \right)$$
\[
\frac{\partial \mathcal{L}(\theta)}{\partial A} = \frac{\partial}{\partial A} \left\{ \sum_{t=2}^{T} \left( -z_{t-1}^T A^T Q^{-1} z_{t-1} - z_{t-1}^T Q^{-1} A z_{t-1} + z_{t-1}^T A^T Q^{-1} A z_{t-1} \right) \right\}
\]
\[
= \frac{\partial}{\partial A} \left\{ -\text{Tr} \left( A^T Q^{-1} \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right) - \text{Tr} \left( A \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right) Q^{-1} \right) \right. \\
+ \left. \text{Tr} \left( Q^{-1} A \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right) A^T \right) \right\}
\]
\[
= -Q^{-1} \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right) - Q^{-1} \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right) \\
+ Q^{-1} A \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right) + Q^{-1} A \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right)
\]
\[
= [0]
\]

\[
A = \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right) \left( \sum_{t=2}^{T} z_{t-1} z_{t-1}^T \right)^{-1}
\]

\[
\frac{\partial \mathcal{L}(\theta)}{\partial Q} = \frac{\partial}{\partial Q} \left\{ \frac{-(T-1)}{2} \log |Q| - \frac{1}{2} \text{Tr} \left( Q^{-1} \sum_{n=2}^{T} (z_{n-1} - A z_{n-1}) (z_{n-1} - A z_{n-1})^T \right) \right\}
\]
\[
= -\frac{(T-1)}{2} Q^{-1} - \frac{1}{2} \left( -Q^{-1} \sum_{n=2}^{T} (z_{n-1} - A z_{n-1}) (z_{n-1} - A z_{n-1})^T Q^{-1} \right)
\]
\[
= [0]
\]
\[ Q = \frac{1}{T-1} \sum_{t=2}^{T} (z_t - A z_{t-1})(z_t - A z_{t-1})^T \]  

Note: The expressions for A and Q are entirely analogous to those in linear regression, where A is the "slope" and Q is the "minimum mean squared error". (see Appendix ii)

Similarly,

\[ \frac{\partial \theta}{\partial C} = \begin{bmatrix} 0 \end{bmatrix} \]  

yields

\[ C = \left( \sum_{t=1}^{T} x_t z_t^T \right) \left( \sum_{t=1}^{T} z_t z_t^T \right)^{-1} \]  

\[ \frac{\partial \theta}{\partial R} = \begin{bmatrix} 0 \end{bmatrix} \]  

yields

\[ R = \frac{1}{T} \sum_{t=1}^{T} (x_t - C z_t)(x_t - C z_t)^T \]  

use the C found in (5).
For notational simplicity, we consider only one sequence \( (\tilde{X}_t, \ldots, \tilde{X}_T) \) here. In general, there may be multiple sequences, each with a different number of time steps.

Let \( \{X_n\}, \{\tilde{X}_n\} \) represent the \( n \)th sequence \( (n = 1, \ldots, N) \).

Now, the goal is to maximize \( \prod_{n=1}^{N} P(\{X_n\}, \{\tilde{X}_n\} \mid \theta) \) w.r.t. \( \theta \).

The resulting expressions for (3) through (6) have the same form, but each summation sums over more elements.

\( \overline{X} \) and \( V \) are the sample mean and covariance, respectively, of the \( N \) instances of \( \tilde{X}_1 \).
C) Test phase: Decoding arm trajectories from neural activity

Goal: To compute \( P(\mathbf{z}_t \mid \mathbf{x}_1, \ldots, \mathbf{x}_t) \)

for \( t = 1, \ldots, T \).

We will use the shorthand \( \{\mathbf{x}_3^t\} \).

The variables \( \mathbf{z}_1, \ldots, \mathbf{z}_T, \mathbf{x}_1, \ldots, \mathbf{x}_T \) are jointly Gaussian, so \( P(\mathbf{z}_t \mid \{\mathbf{x}_3^t\}) \) is Gaussian.

Thus, we need only find its mean and covariance.

We can compute \( P(\mathbf{z}_t \mid \{\mathbf{x}_3^t\}) \) recursively starting at \( t = 1 \):

- One-step prediction

\[
P(\mathbf{z}_t \mid \{\mathbf{x}_3^{t-1}\}) = \int P(\mathbf{z}_t \mid \mathbf{z}_{t-1}) P(\mathbf{z}_{t-1} \mid \{\mathbf{x}_3^{t-1}\}) d\mathbf{z}_{t-1}
\]

(7)

- Measurement update

\[
P(\mathbf{z}_t \mid \{\mathbf{x}_3^t\}) = \frac{P(\mathbf{x}_t \mid \mathbf{z}_t) P(\mathbf{z}_t \mid \{\mathbf{x}_3^{t-1}\})}{P(\mathbf{x}_t \mid \{\mathbf{x}_3^{t-1}\})}
\]

(8)
Let
\[ \mu_t^t = E[Z_t | \{X_j^t\}_1^t] \]
\[ \Sigma_t^t = \text{cov}(Z_t | \{X_j^t\}_1^t) \]

We want to express (7) and (8) in terms of the model parameters.

Plugging the state and observation models into (7) and (8), then simplifying, is a method that will always work.

Here, we will recognize that all the distributions in (7) and (8) are Gaussian, and just solve for means and covariances.

- One-step prediction

\[ Z_t | \{X_j^t\}_1^t \sim N(\mu_t^{t-1}, \Sigma_t^{t-1}) \]

Find \( \mu_t^{t-1} \) and \( \Sigma_t^{t-1} \).

An equivalent way of writing (1) is

\[ Z_t = A Z_{t-1} + V_t, \quad V_t \sim N(0, Q) \]

\[ \mu_t^{t-1} = E[Z_t | \{X_j^t\}_1^t] \]

\[ = A E[Z_{t-1} | \{X_j^t\}_1^t] + E[V_t | \{X_j^t\}_1^t] \]

\[ \mu_t^t = A \mu_t^{t-1} \]
\[
\Sigma_t^{t-1} = \text{cov} \left( z_t \mid \{x_i\}_{i=1}^{t-1} \right) \\
= A \text{cov} \left( z_{t-1} \mid \{x_i\}_{i=1}^{t-1} \right) A^T + \text{cov} \left( r_t \mid \{x_i\}_{i=1}^{t-1} \right) \\
\Sigma_t = A \Sigma_t^{t-1} A^T + Q 
\] (10)

**Measurement update**

\[
z_t \mid \{x_i\}_{i=1}^{t} \sim N \left( \mu_t^*, \Sigma_t^* \right) \\
\text{Find } \mu_t^* \text{ and } \Sigma_t^* 
\]

Recognize that (8) is Bayes rule for \( z_t \) and \( x_t \), with all terms conditioned on \( \{x_i\}_{i=1}^{t-1} \).

Thus, we will first find the joint probability of \( z_t \) and \( x_t \) given \( \{x_i\}_{i=1}^{t-1} \) then apply the results of conditioning for jointly Gaussian random vars.

\[
\begin{bmatrix}
X_t \\
Z_t
\end{bmatrix} \mid \{x_i\}_{i=1}^{t-1} \sim N \left( \begin{bmatrix}
\mu_t^{t-1} \\
\mu_t^{t-1}
\end{bmatrix}, \begin{bmatrix}
\Sigma_t^{t-1} & C \Sigma_t^{t-1} \\
C \Sigma_t^{t-1} & \Sigma_t + R
\end{bmatrix} \right) 
\] (11)
An equivalent way of writing (2) is

$$X_t = CZ_t + W_t, \quad W_t \sim N(0, R).$$

a) $$E[X_t | \{X_{s}\}_{s=1}^{t-1}] = CE[Z_t | \{X_{s}\}_{s=1}^{t-1}] + E[W_t | \{X_{s}\}_{s=1}^{t-1}]$$

$$= C \mu_t^{t-1}$$

$$\text{cov}(X_t | \{X_{s}\}_{s=1}^{t-1}) = C \text{cov}(Z_t | \{X_{s}\}_{s=1}^{t-1}) C^T + \text{cov}(W_t | \{X_{s}\}_{s=1}^{t-1})$$

$$= C \sum_{t-s}^{t-1} C^T + R$$

b) $$E[X_tZ_T^T | \{X_{s}\}_{s=1}^{t-1}] = E[X_t | \{X_{s}\}_{s=1}^{t-1}] E[Z_T | \{X_{s}\}_{s=1}^{t-1}]^T$$

$$= E[CZ_tZ_T^T + W_tZ_T^T | \{X_{s}\}_{s=1}^{t-1}] - C \mu_t^{t-1} \mu_T^{t-1T}$$

$$= C \left( \sum_{t-s}^{t-1} \mu_T^{t-1} \mu_T^{t-1T} \right) + E[W_t | \{X_{s}\}_{s=1}^{t-1}] E[Z_T | \{X_{s}\}_{s=1}^{t-1}]^T$$

$$- C \mu_t^{t-1} \mu_T^{t-1T}$$

$$= C \sum_{t}^{t-1}$$
Applying the results of conditioning for jointly Gaussian random variables to (11),
(see Appendix iii)

\[ \mu_t^t = E[ \xi_t \mid x_t, \{x_i \}_{i=1}^{t-1}] \]

\[ = \mu_t^{t-1} + \Sigma_t^{t-1} C^T (C \Sigma_t^{t-1} C^T + R)^{-1} (x_t - C \mu_t^{t-1}) \]

(all this \( K_t \), the "Kalman gain"

 Rewriting,

\[ \mu_t^t = \mu_t^{t-1} + K_t (x_t - C \mu_t^{t-1}) \]  
\[ \text{(12)} \]

\[ \Sigma_t^t = \text{cov} (\xi_t \mid x_t, \{x_i \}_{i=1}^{t-1}) \]

\[ = \Sigma_t^{t-1} - \Sigma_t^{t-1} C^T (C \Sigma_t^{t-1} C^T + R)^{-1} C \Sigma_t^{t-1} \]

\[ \Sigma_t^t = \Sigma_t^{t-1} - K_t C \Sigma_t^{t-1} \]  
\[ \text{(13)} \]
Taking the recursions defined by (9), (10), (12), (13), we obtain $\mu_t$ and $\Sigma_t$ for $t=1, \ldots, T$.

- $\mu_t$ is the arm state estimate at time $t$.
- $\Sigma_t$ is our uncertainty around that estimate at time $t$.

Initialize recursions with $\mu_1^0 = \Pi$

$$\Sigma_1^0 = \Lambda.$$
Appendix

i) Useful matrix properties

\[ \frac{d}{dX} \text{Tr}(X A^T) = \frac{d}{dX} \text{Tr}(X^T A) = A \]

\[ \frac{d}{dX} \text{Tr}(A X B X^T C) = A^T C^T X B^T + C A X B \]

\[ \frac{d}{dX} \log |X| = X^{-T} \]

\[ \frac{d}{dX} \text{Tr}(X^{-T} A) = -X^{-T} A^T X^{-T} \]

ii) Linear regression

\[ X_n = W Z_n + \mu, \quad n = 1, \ldots, N \]

Minimizing mean squared error with respect to \( W \),

\[ W^* = \frac{\sum_{n=1}^{N} (X_n - \mu) Z_n}{\sum_{n=1}^{N} Z_n^2} \]

Using this \( W^* \), minimum mean squared error is

\[ \frac{1}{N} \sum_{n=1}^{N} (X_n - W^* Z_n - \mu)^2 \]
iii) Gaussian conditioning

\[ y = \begin{bmatrix} x_a \\ x_b \end{bmatrix} \sim N \left( \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix}, \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix} \right) \]

\( P(x_a | x_b) \) is Gaussian with the following mean and covariance:

\[ \mathbb{E} \left[ x_a | x_b \right] = \mu_a + \Sigma_{ab} \Sigma_{bb}^{-1} (x_b - \mu_b) \]

\[ \text{cov} (x_a | x_b) = \Sigma_{aa} - \Sigma_{ab} \Sigma_{bb}^{-1} \Sigma_{ba} \]