CS281A/Stat241A Lecture 18

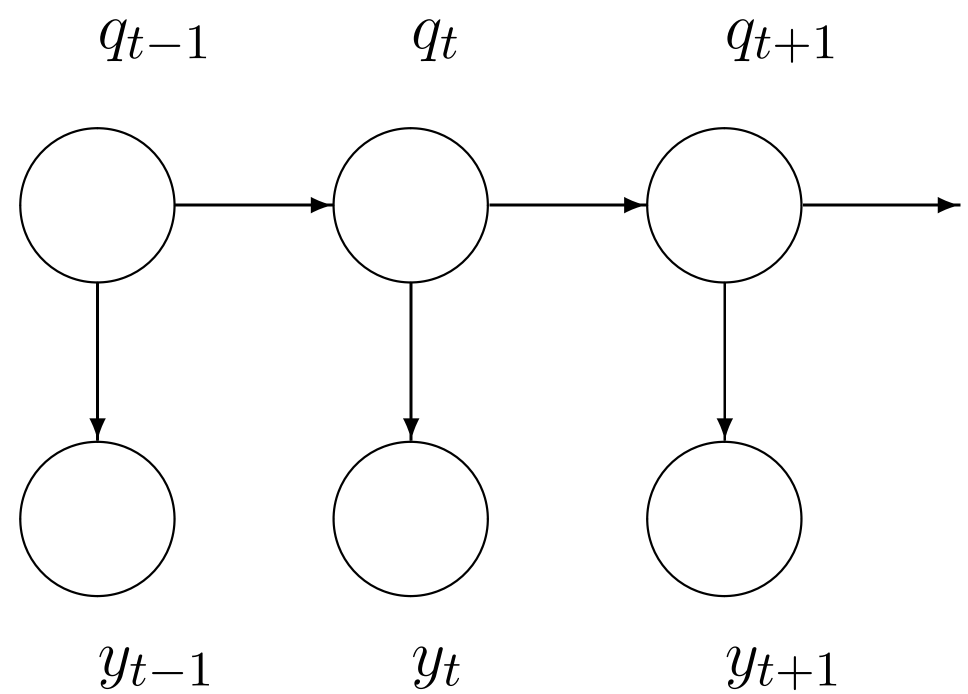
State Space Models

Peter Bartlett

Key ideas of this lecture

- Review: EM in HMMs.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - Recall: All distributions are Gaussian, so parameters suffice.
  - Inference: Kalman filter and smoother.
  - Parameter estimation with EM.
  - Extended Kalman filter.
- Junction Tree Algorithm.

Recall: Hidden Markov Models



Hidden Markov Models

$$
\begin{aligned}
p(q_0^i = 1) &= \pi_i, \\
p(q_t^j = 1 | q_{t-1}^i = 1) &= a_{ij}, \\
p(y_t | q_t^i = 1) &= h(y_t) \exp\left(\eta_i' T(y_t) - A(\eta_i)\right).
\end{aligned}
$$

Hidden Markov Models

EM: We have data $y_0, \ldots, y_T$, and we wish to estimate the parameters of an HMM.

1. Write down the complete log likelihood.
2. **E step:** Calculate the conditional expectation of the complete log likelihood (ie: sufficient statistics).
3. **M step:** Maximize $\mathbb{E}[\ell_c|y]$.

EM in HMMs: ℓ_c

$$\log p(q, y|\theta) = \log \pi_{q_0} + \sum_{t=0}^{T-1} \log a_{q_t, q_{t+1}} + \sum_{t=0}^{T} \log p(y_t|q_t)$$

$$= \sum_i \underbrace{q_0^i}_{SS} \log \pi_i + \sum_{i,j} \underbrace{\sum_{t=0}^{T-1} q_t^i q_{t+1}^j}_{SS} \log a_{i,j}$$

$$+ \sum_i \underbrace{\sum_{t=0}^{T} q_t^i \left(T(y_t)'\right.}_{SS} \left.\eta_i - A(\eta_i)\right)$$

$$+ (T+1) \log h(y_t).$$

EM in HMMs: E step

- Calculate the conditional expectation of the complete log likelihood.
- This corresponds to computing expected sufficient statistics:

$$\begin{aligned}
\mathbb{E}\left[\ell_c(\theta; q, y) | y\right] = \sum_i p(q_0^i = 1|y) \log \pi_i \\
+ \sum_{i,j} \sum_{t=0}^{T-1} p(q_t^i q_{t+1}^j = 1|y) \log a_{i,j} \\
+ \sum_i \sum_{t=0}^{T} p(q_t^i = 1|y) \left(T(y_t)' \eta_i - A(\eta_i)\right) \\
+ (T+1) \log h(y_t).
\end{aligned}$$

EM in HMMs: E step

In the notation of forward-backward algorithm, the expected sufficient statistics are

$$\text{for } \pi_i: \qquad p(q_0^i = 1|y) = \gamma_0^i,$$

$$\text{for } a_{i,j}: \qquad \sum_{t=0}^{T-1} p(q_t^i q_{t+1}^j = 1|y) = \sum_{t=0}^{T-1} \xi_{t,t+1}^{i,j},$$

$$\text{for } \mu_i: \qquad \sum_{t=0}^{T} p(q_t^i = 1|y) T(y_t) = \sum_{t=0}^{T} \gamma_t^i T(y_t).$$

EM in HMMs: M step

Recall: For complete data, the ML estimates are

$$\hat{\pi}_i = q_0^i;$$

$$\hat{a}_{i,j} = \frac{\sum_{t=0}^{T-1} q_t^i q_{t+1}^j}{\sum_{t=0}^{T-1} q_t^i} \qquad \text{prop. of } i \to j$$

$$\hat{\mu}_i = \frac{\sum_{t=0}^{T} q_t^i T(y_t)}{\sum_{t=0}^{T-1} q_t^i} \qquad \text{av. of SS}$$

EM in HMMs: M step

Maximizing $\mathbb{E}[\ell_c|y]$ is the same as in the completely observed case, but the counts are replaced by 'soft' counts:

$$\hat{\pi}_i = \gamma_0^i;$$

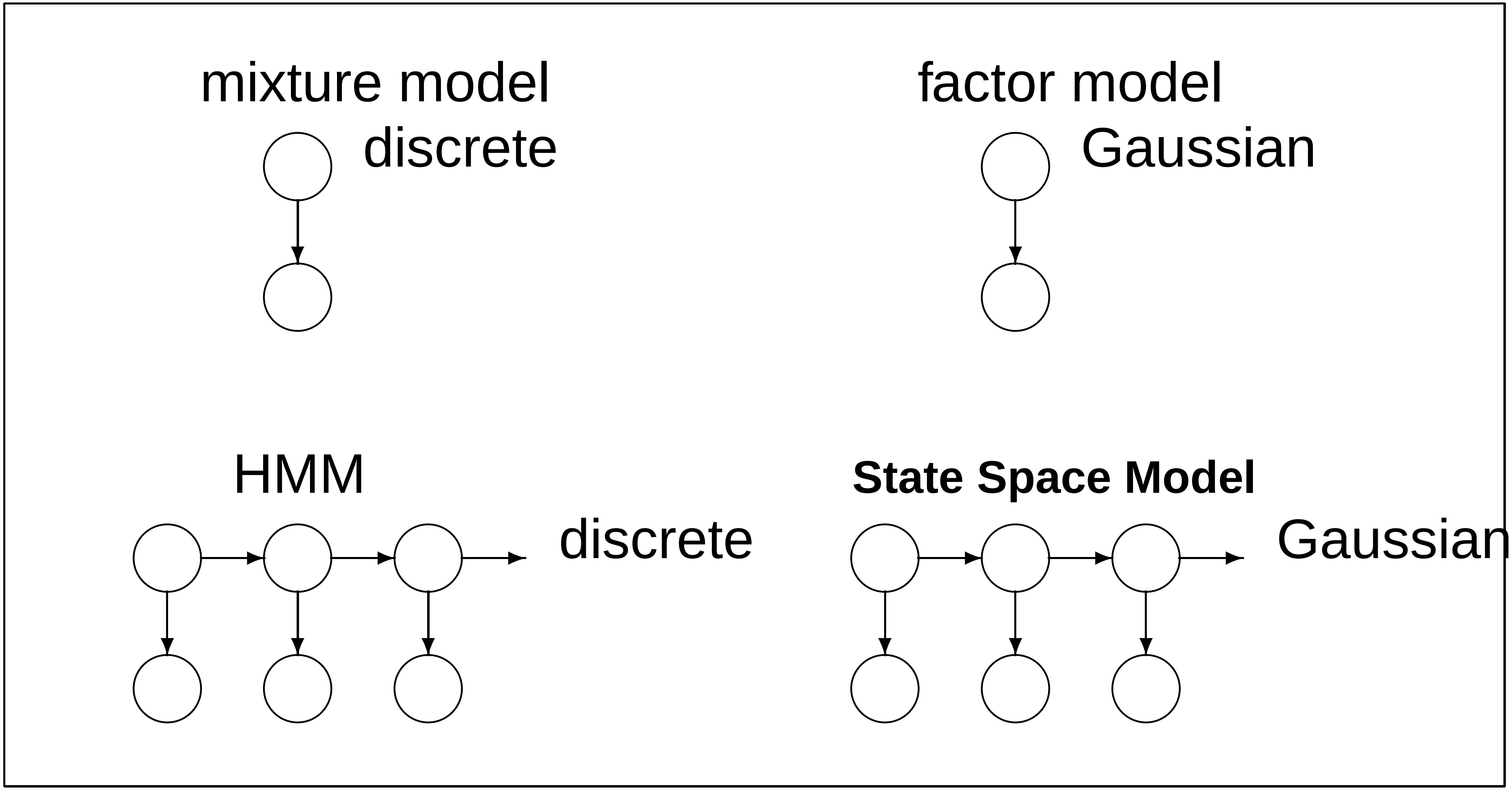
$$\hat{a}_{i,j} = \frac{\sum_{t=0}^{T-1} \xi_{t,t+1}^{i,j}}{\sum_{t=0}^{T-1} \gamma_t^i}$$

$$\hat{\mu}_i = \frac{\sum_{t=0}^{T} \gamma_t^i T(y_t)}{\sum_{t=0}^{T-1} \gamma_t^i}$$

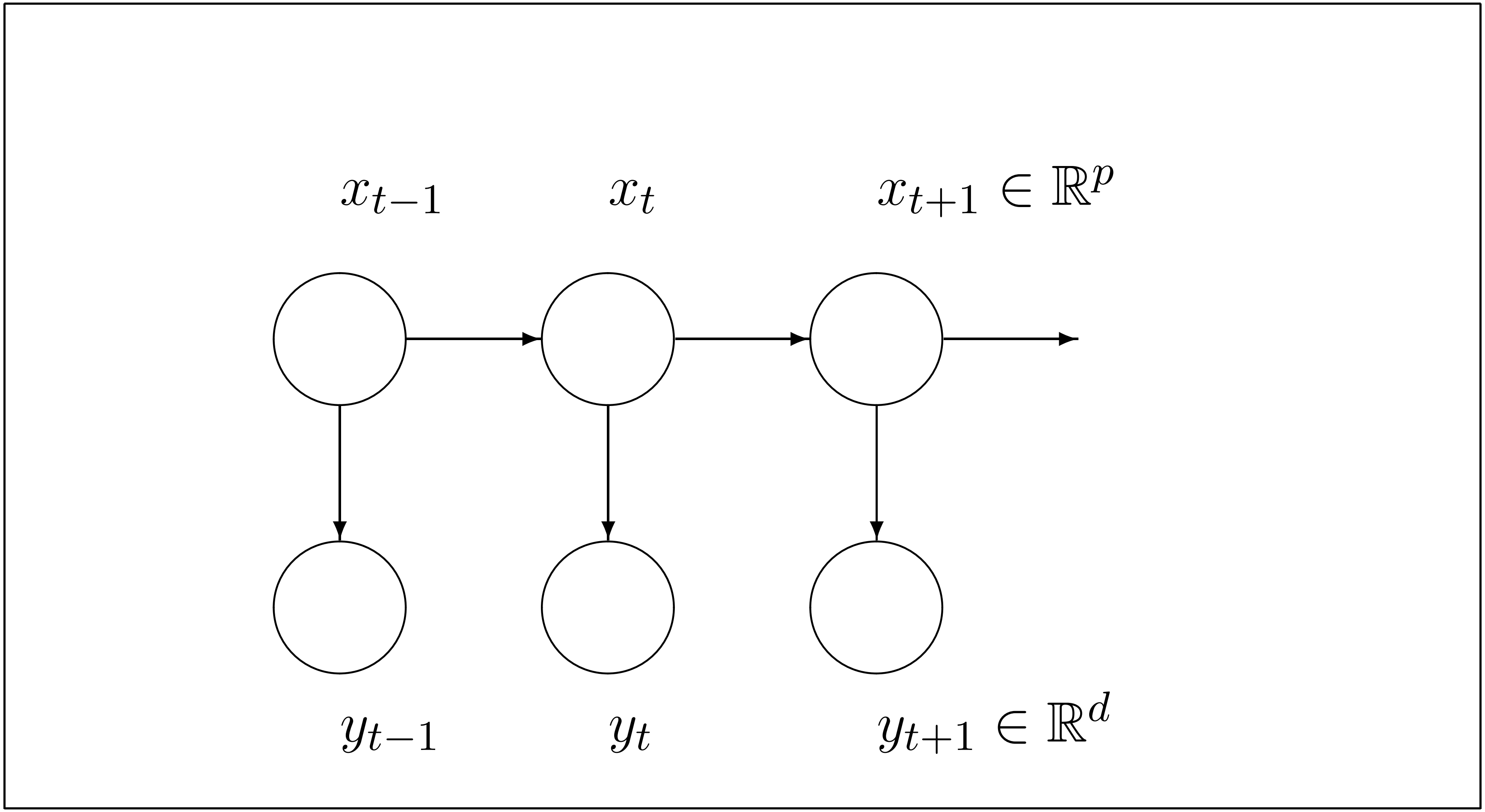
Key ideas of this lecture

- Review: EM in HMMs.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - Recall: All distributions are Gaussian, so parameters suffice.
  - Inference: Kalman filter and smoother.
  - Parameter estimation with EM.
  - Extended Kalman filter.
- Junction Tree Algorithm.

State Space Models



Linear System: Definition



Linear System: Definition

State	$x_t \in \mathbb{R}^p$
Observation	$y_t \in \mathbb{R}^d$
Initial state	$x_0 \sim \mathcal{N}(0, P_0)$
Dynamics	$x_{t+1} = Ax_t + w_t, \qquad w_t \sim \mathcal{N}(0, Q)$
Observation	$y_t = Cx_t + v_t, \qquad v_t \sim \mathcal{N}(0, R).$

Linear Systems: Recall

1. All the distributions are Gaussian (joints, marginals, conditionals), so they can be described by their means and variances.
2. The conditional distribution of the next state, $x_{t+1}|x_t$, is
$$\mathcal{N}(Ax_t, Q).$$
3. The marginal distribution of x_t is $\mathcal{N}(0, P_t)$, where P_0 is given and, for $t \geq 0$,
$$P_{t+1} = AP_tA' + Q.$$

Inference in SSMs

Filtering: $p(x_t|y_0, \ldots, y_t)$.

Smoothing: $p(x_t|y_0, \ldots, y_T)$.

For inference, it suffices to calculate the appropriate conditional means and covariances.

Inference in SSMs: Notation

$$\hat{x}_{t|s} = \mathbb{E}(x_t | y_0, \ldots, y_s),$$

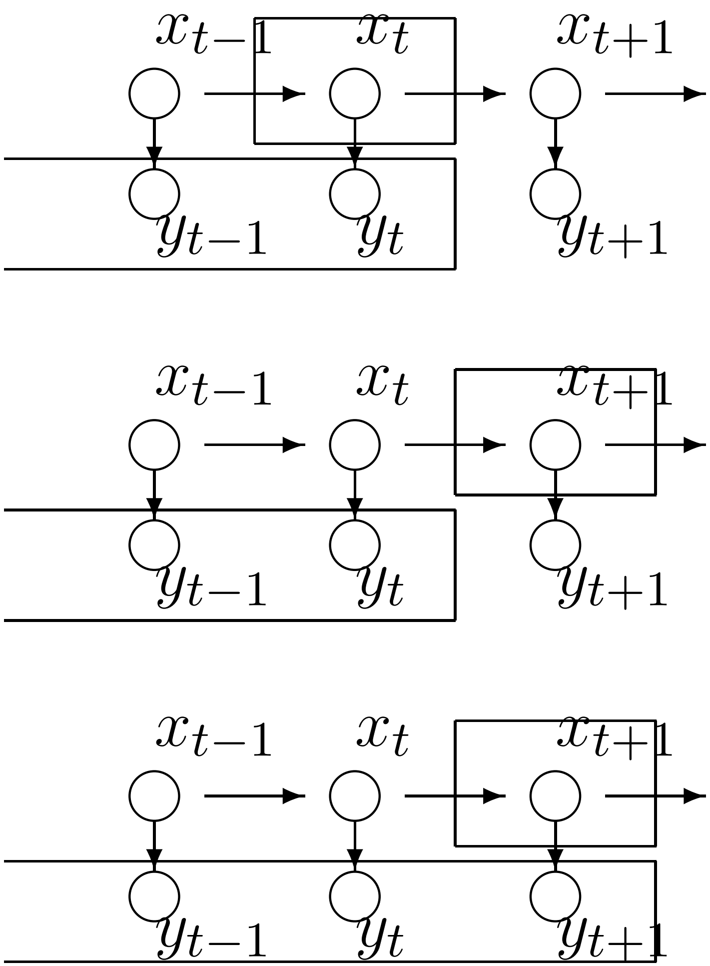
$$P_{t|s} = \mathbb{E}((x_t - \hat{x}_{t|s})(x_t - \hat{x}_{t|s})' | y_0, \ldots, y_s).$$

Filtering: $x_{t|t} \sim \mathcal{N}(\hat{x}_{t|t}, P_{t|t}),$

Smoothing: $x_{t|T} \sim \mathcal{N}(\hat{x}_{t|T}, P_{t|T}).$

The **Kalman Filter** is an inference algorithm for $\hat{x}_{t|t}$, $P_{t|t}$.
The **Kalman Smoother** is an inference algorithm for $\hat{x}_{t|T}$, $P_{t|T}$.

The Kalman Filter



$p(x_t|y_0, \ldots, y_t)$ $\quad \hat{x}_{t|t}, P_{t|t}$

time update

$p(x_{t+1}|y_0, \ldots, y_t)$ $\quad \hat{x}_{t+1|t}, P_{t+1|t}$

measurement update

$p(x_{t+1}|y_0, \ldots, y_{t+1})$ $\quad \hat{x}_{t+1|t+1}, P_{t+1|t+1}$

The Kalman Filter

$$\hat{x}_{t+1|t} = A\hat{x}_{t|t},$$
$$P_{t+1|t} = AP_{t|t}A' + Q.$$

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t}),$$
$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1}CP_{t+1|t}.$$

The Kalman Filter

Time update

$$
\begin{aligned}
x_{t+1} &= Ax_t + w_t. \\
\hat{x}_{t+1|t} &= \mathbb{E}(x_{t+1}|y_0, \ldots, y_t) \\
&= A\mathbb{E}(x_t|y_0, \ldots, y_t) \\
&= A\hat{x}_{t|t}. \\
P_{t+1|t} &= \mathbb{E}((x_{t+1} - \hat{x}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})'|y_0, \ldots, y_t) \\
&= \mathbb{E}(A(x_t - \hat{x}_{t|t}) + w_t)(A(x_t - \hat{x}_{t|t}) + w_t)'|y_0, \ldots, y_t) \\
&= AP_{t|t}A' + Q,
\end{aligned}
$$

since w_t and x_t are uncorrelated.

The Kalman Filter

Measurement update 1. Compute the parameters of the joint Gaussian distribution

$$p(x_{t+1}, y_{t+1} | y_0, \ldots, y_t)$$

We know the x_{t+1} part from the time update. For the y_{t+1} part,

$$\begin{aligned}\hat{y}_{t+1|t} &= \mathbb{E}(y_{t+1} | y_0, \ldots, y_t) \\ &= \mathbb{E}(Cx_{t+1} + v_{t+1} | y_0, \ldots, y_t) \\ &= C\hat{x}_{t+1|t}.\end{aligned}$$

The Kalman Filter

$$\begin{aligned}
&\mathbb{E}((y_{t+1} - \hat{y}_{t+1|t})(y_{t+1} - \hat{y}_{t+1|t})'|y) \\
&= \mathbb{E}(C(x_{t+1} - \hat{x}_{t+1|t}) + v_{t+1})(C(x_{t+1} - \hat{x}_{t+1|t}) + v_{t+1})'|y) \\
&= CP_{t+1|t}C' + R,
\end{aligned}$$

since v_{t+1} and x_{t+1} are uncorrelated.

The Kalman Filter

And for the cross terms,

$$
\begin{aligned}
&\mathbb{E}((y_{t+1} - \hat{y}_{t+1|t})(x_{t+1} - \hat{x}_{t+1|t})'|y) \\
&= \mathbb{E}(C(x_{t+1} - \hat{x}_{t+1|t}) + v_{t+1})(x_{t+1} - \hat{x}_{t+1|t})'|y) \\
&= CP_{t+1|t}.
\end{aligned}
$$

The Kalman Filter

Hence, the distribution $p(x_{t+1}, y_{t+1}|y_0, \ldots, y_t)$ is

$$\mathcal{N}\left(\begin{pmatrix} \hat{x}_{t+1|t} \\ C\hat{x}_{t+1|t} \end{pmatrix}, \begin{pmatrix} P_{t+1|t} & P_{t+1|t}C' \\ CP_{t+1|t} & CP_{t+1|t}C' + R \end{pmatrix}\right)$$

2. Hence, compute the parameters of the conditional Gaussian distribution

$$p(x_{t+1}|y_0, \ldots, y_t, y_{t+1})$$

This follows from the decomposition of a joint Gaussian into a marginal and a conditional:

The Kalman Filter

The conditional has mean

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1}(y_{t+1} - C\hat{x}_{t+1|t})$$

and the variance is the Schur complement,

$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1}CP_{t+1|t}.$$

The Kalman Filter: Interpretation

If we define the **Kalman gain matrix,**

$$K_{t+1} = P_{t+1|t}C'(CP_{t+1|t}C' + R)^{-1},$$

then the time and measurement updates give

$$\hat{x}_{t+1|t+1} = A\hat{x}_{t|t} + K_{t+1}(y_{t+1} - CA\hat{x}_{t|t}).$$

Notice that the last term is prediction error, since

$$\mathbb{E}(y_{t+1}|y_0, \ldots, y_t) = CA\hat{x}_{t|t}.$$

Thus, the state estimate evolves as

$$\hat{x}_{t+1|t+1} = A\hat{x}_{t|t} + K_{t+1}(y_{t+1} - CA\hat{x}_{t|t}).$$

cf. LMS: $$\theta_{t+1} = \theta_t + \rho x_t(y_{t+1} - x_t'\theta_t).$$

The Kalman Filter: Other Variants

Information filter: Kalman filter recursion in terms of natural parameters ($\Lambda = \Sigma^{-1}, \eta = \Sigma^{-1}\mu$).

Kalman Smoother: Analogous to the α-β (forward-backward) recursion for inference in HMMs.
Recall that the αs are like $\hat{x}_{t|t}$, $P_{t|t}$. The βs calculate parameters of conditional distribution of x_t given $y_t, \ldots, y_T$. This is equivalent to running a Kalman filter backwards: find an equivalent time-reversed version of the linear system, and run a Kalman filter for it.

The Kalman Filter: Other Variants

Rauch-Tung-Streibel: Analogous to the α-γ recursion for inference in HMMs.
Recall that the γs express parameters of the conditional distribution of x_t given $y_0, \ldots, y_T$, using the already computed αs.
You can read the details.

Parameter Estimation with EM

Given observed data $y = (y_0, \ldots, y_T)$ and hidden states $x = (x_0, \ldots, x_T)$, we want to estimate the parameters $\theta = (P_0, A, C, Q, R)$:

$$x_0 \sim \mathcal{N}(0, P_0),$$
$$x_{t+1} \sim \mathcal{N}(Ax_t, Q),$$
$$y_t \sim \mathcal{N}(Cx_t, R).$$

Parameter Estimation with EM: ℓ_c

We can write the complete log likelihood as

$$\begin{aligned}
\ell_c(\theta; x, y) = -\frac{1}{2}\Bigg(&\ln(2\pi|P_0|) + x_0' P_0^{-1} x_0 \\
&+ \sum_{t=0}^{T-1}\Big(\ln(2\pi|Q|) + (x_{t+1} - Ax_t)' Q^{-1} (x_{t+1} - Ax_t)\Big) \\
&+ \sum_{t=0}^{T}\Big(\ln(2\pi|R|) + (y_t - Cx_t)' R^{-1} (y_t - Cx_t)\Big)\Bigg).
\end{aligned}$$

Parameter Estimation with EM: E step

$$
\begin{aligned}
&\mathbb{E}(\ell_c(\theta; x, y)|y) \\
&= \text{const} - \frac{1}{2}\Big(\ln|P_0| + \mathrm{tr}(P_0^{-1}\mathbb{E}(x_0 x_0'|y)) \\
&+ T\ln|Q| + \mathrm{tr}\left(Q^{-1}\sum_{t=0}^{T-1}\mathbb{E}\left((x_{t+1} - Ax_t)(x_{t+1} - Ax_t)'|y\right)\right) \\
&+ (T+1)\ln|R| + \mathrm{tr}\left(R^{-1}\sum_{t=0}^{T}\mathbb{E}\left((y_t - Cx_t)(y_t - Cx_t)'|y\right)\right)\Big).
\end{aligned}
$$

Parameter Estimation with EM: E step

Thus, the expected sufficient statistics are:

$$\mathbb{E}(x_t|y) = \hat{x}_{t|T}$$

$$\mathbb{E}(x_t x_t'|y) = \hat{x}_{t|T}\hat{x}_{t|T}' + P_{t|T}$$

$$\mathbb{E}(x_t x_{t+1}'|y) = \hat{x}_{t|T}\hat{x}_{t+1|T}' + \textsf{cov}(x_t, x_{t+1}|y).$$

(Can calculate the latter covariance from the output of, for example, the Rauch-Tung-Striebel algorithm.)

Parameter Estimation with EM: M step

Choose θ to minimize. Can rearrange and decompose to show that the optimal A, C are solutions to minimization problems of the following form (multiple output linear regression):

Claim: For a positive definite symmetric M and positive semidefinite symmetric W, the matrix A that minimizes

$$\text{tr}\left(W(A'MA - N'A - A'N)\right)$$

is $M^{-1}N$.

Parameter Estimation with EM: M step

For example, for C, you can check that the optimization is minimization of

$$\mathrm{tr}\left(W(A'MA - N'A - A'N)\right)$$

$$\begin{aligned} M &= \sum_{t=0}^{T} (\hat{x}_{t|T}\hat{x}'_{t|T} + P_{t|T}), \\ N &= \sum_{t=0}^{T} (\hat{x}_{t|T} y'_t), \\ W &= R^{-1}. \end{aligned}$$

Parameter Estimation with EM: M step

It is also clear that the optimal P_0, Q, R are solutions to minimization problems of the following form (maximum likelihood covariance estimation problems):

Claim: For positive definite symmetric P and S,

$$\ln|P| + \mathrm{tr}\left(P^{-1}S\right) \geq \ln|S| + \mathrm{tr}(S^{-1}S).$$

Proof:

$$\begin{aligned} \ln|P| + \mathrm{tr}\left(P^{-1}S\right) &= -\ln|P^{-1}S| + \mathrm{tr}\left(P^{-1}S\right) + \ln|S| \\ &= \sum_i \lambda_i - \ln\lambda_i + \ln|S| \\ &\geq \sum_i 1 + \ln|S|. \end{aligned}$$

Linear Systems: EM. Summary.

E step: Calculate the expected suff. stats:

$$\mathbb{E}(x_t|y) = \hat{x}_{t|T}$$
$$\mathbb{E}(x_t x_t'|y) = \hat{x}_{t|T}\hat{x}_{t|T}' + P_{t|T}$$
$$\mathbb{E}(x_t x_{t+1}'|y) = \hat{x}_{t|T}\hat{x}_{t+1|T}' + \text{cov}(x_t, x_{t+1}|y).$$

And use these to compute the various terms in $\mathbb{E}(\ell_c(\theta; x, y)|y)$.

M step: Maximize $\mathbb{E}[\ell_c|y]$:

- A, C are solutions to multiple output linear regression problems.
- P_0, Q and R are time averages of conditional covariances.

Extended Kalman Filter

- Suppose that the state and observation models follow some (typically known) nonlinear functions:

State	$x_t \in \mathbb{R}^p$
Observation	$y_t \in \mathbb{R}^d$
Initial state	$x_0 \sim \mathcal{N}(0, P_0)$
Dynamics	$x_{t+1} = f(x_t) + w_t, \qquad w_t \sim \mathcal{N}(0, Q)$
Observation	$y_t = g(x_t) + v_t, \qquad v_t \sim \mathcal{N}(0, R).$

Extended Kalman Filter

- If f and g are smooth (close to linear), then we can approximate them as linear functions about the current expected state

$$x_{t+1} \approx f(\hat{x}_{t|t}) + F(x_t - \hat{x}_{t|t}) + w_t,$$
$$y_{t+1} \approx g(\hat{x}_{t+1|t}) + G(x_{t+1} - \hat{x}_{t+1|t}) + v_{t+1}.$$

where the matrices F and G are the Jacobians of f and g that appear in the linearization.

$$F = \left.\frac{\partial f}{\partial x}\right|_{\hat{x}_{t|t}},$$
$$G = \left.\frac{\partial g}{\partial x}\right|_{\hat{x}_{t+1|t}}.$$

Extended Kalman Filter

- If the linear approximation is accurate in a region where most of the mass is contained, we can approximate the conditional distributions as Gaussian, and use a modification of the Kalman filter:

$$\hat{x}_{t+1|t} = f(\hat{x}_{t|t}),$$
$$P_{t+1|t} = F P_{t|t} F' + Q.$$

$$\hat{x}_{t+1|t+1} = \hat{x}_{t+1|t} + P_{t+1|t} G' (G P_{t+1|t} G' + R)^{-1} \left(y_{t+1} - h(\hat{x}_{t+1|t}) \right),$$
$$P_{t+1|t+1} = P_{t+1|t} - P_{t+1|t} G' (G P_{t+1|t} G' + R)^{-1} G P_{t+1|t}.$$

(The matrices F and G replace A and C.)

Key ideas of this lecture

- Review: EM in HMMs.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - Recall: All distributions are Gaussian, so parameters suffice.
  - Inference: Kalman filter and smoother.
  - Parameter estimation with EM.
  - Extended Kalman filter.
- Junction Tree Algorithm.

Junction Tree Algorithm

- Inference: Given
  - Graph $G = (V, E)$,
  - Evidence x_E, for $E \subseteq V$,
  - Set $F \subseteq V$,

  compute $p(x_F | x_E)$.

Junction Tree Algorithm

- Elimination:
  - Single set F.
  - Any G.
- Sum-product:
  - **All singleton sets** F simultaneously.
  - G a **tree**.
- **Junction tree:**
  - **All cliques** F simultaneously.
  - Any G.

Junction Tree Algorithm

- Combines elimination algorithm with caching of sum-product.
- Messages (marginalized potentials) passed between **cliques**, in a junction tree.

Junction Tree Algorithm

1. (For directed graphical models:) Moralize.
   So all potentials—local conditionals—are defined on **cliques.**
2. Triangulate.
   e.g., via elimination algorithm
3. Construct a junction tree.
4. Define potentials on maximal cliques.
5. Introduce evidence.
6. Propagate probabilities.

Key ideas of this lecture

- Review: EM in HMMs.
- State Space Models.
  - Linear dynamical systems, gaussian disturbances.
  - Recall: All distributions are Gaussian, so parameters suffice.
  - Inference: Kalman filter and smoother.
  - Parameter estimation with EM.
  - Extended Kalman filter.
- Junction Tree Algorithm.