

CS 277: Control and Reinforcement Learning

Winter 2024

Lecture 8: Stochastic Optimal Control

Roy Fox

Department of Computer Science

School of Information and Computer Sciences

University of California, Irvine



Logistics

assignments

- Exercise 2 and Quiz 4 due **next Monday**

videos

- Video on trust-region methods on the course website
- Might help with Exercise 2 and Quiz 4

Today's lecture

LQR with process noise

Linear–Quadratic Estimator

Linear–Quadratic–Gaussian control

Reminder: Linear Quadratic Regulator (LQR)

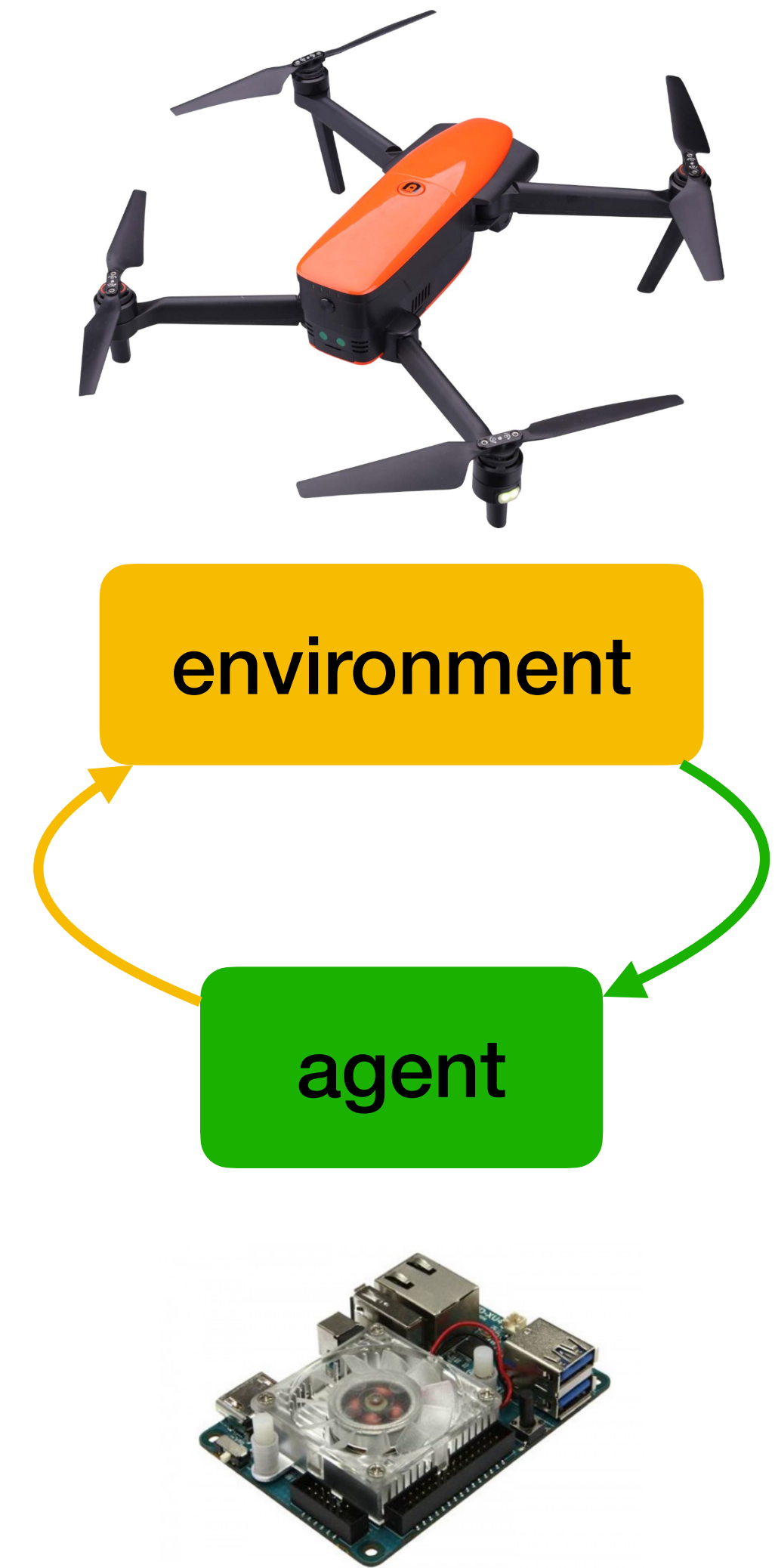
- Linear Quadratic Regulation (LQR) optimization problem:

- ▶ Given LTI dynamics + quadratic cost (A, B, Q, R)

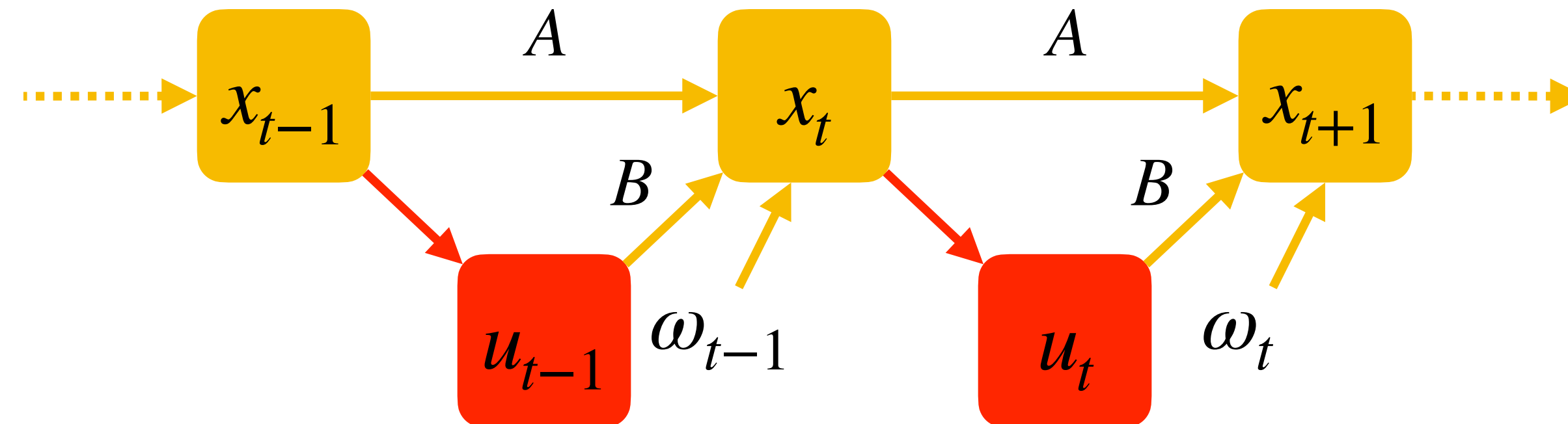
- ▶ Find the control function $u_t = \pi(x_t)$

- ▶ That minimizes $J^\pi = \sum_{t=0}^{T-1} c(x_t, u_t) = \frac{1}{2} \sum_{t=0}^{T-1} (x_t^\top Q x_t + u_t^\top R u_t)$

- ▶ Such that $x_{t+1} = Ax_t + Bu_t$ for all t



Stochastic control



- Simplest stochastic dynamics – **Gaussian**: $p(x_{t+1} | x_t, u_t) = \mathcal{N}(x_{t+1}; Ax_t + Bu_t, \Sigma_\omega)$

$$x_{t+1} = Ax_t + Bu_t + \omega_t \quad \omega_t \sim \mathcal{N}(0, \Sigma_\omega) \quad \Sigma_\omega \in \mathbb{R}^{n \times n}$$

- ▶ **Markov property**: all ω_t are i.i.d for all t
- Why is there **process noise**?
 - ▶ Part of the state we **don't model**; Gaussian = **maximum entropy** given Σ_ω
- In continuous time = **Langevin equation**; $Bu_t =$ **external force**

Stochastic optimal control

- Minimize **expected cost-to-go**

$$V_t^\pi(x_t) = \frac{1}{2}x_t^\top Qx_t + \frac{1}{2}u_t^\top Ru_t + \mathbb{E}[V_{t+1}^\pi(x_{t+1}) | x_t, u_t = \pi(x_t)]$$

- **Bellman equation:**

$$V_t(x_t) = \min_{u_t} \frac{1}{2}x_t^\top Qx_t + \frac{1}{2}u_t^\top Ru_t + \mathbb{E}_{(x_{t+1}|x_t, u_t) \sim \mathcal{N}(Ax_t + Bu_t, \Sigma_\omega)}[V_{t+1}(x_{t+1})]$$

- The cost-to-go is still quadratic, but with a **free term**
 - $x_t = 0$ is no longer **absorbing** $\Rightarrow V_t(0) \neq 0$

$$V_t(x_t) = \frac{1}{2}x_t^\top S_t x_t + V_t(0)$$

Solving the Bellman recursion

- Good to know: expectation of **quadratic** under **Gaussian** is $\mathbb{E}_{x \sim \mathcal{N}(\mu_x, \Sigma_x)}[x^\top S x] = \mu_x^\top S \mu_x + \text{tr}(S \Sigma_x)$

$$V_t(x_t) = \min_{u_t} \mathbb{E}_{(x_{t+1}|x_t, u_t) \sim \mathcal{N}(Ax_t + Bu_t, \Sigma_\omega)} \left[\frac{1}{2} x_t^\top Q x_t + \frac{1}{2} u_t^\top R u_t + \underbrace{\frac{1}{2} x_{t+1}^\top S_{t+1} x_{t+1}} + V_{t+1}(0) \right]$$

$$= \min_{u_t} \left(\frac{1}{2} x_t^\top Q x_t + \frac{1}{2} u_t^\top R u_t + \frac{1}{2} (Ax_t + Bu_t)^\top S_{t+1} (Ax_t + Bu_t) + \frac{1}{2} \text{tr}(S_{t+1} \Sigma_\omega) + V_{t+1}(0) \right)$$

new term, constant in u_t

- **Linear control**: $u_t^* = L_t x_t$ with same **feedback gain**: $L_t = - (R + B^\top S_{t+1} B)^{-1} B^\top S_{t+1} A$
- Same **Ricatti equation** for cost-to-go Hessian: $S_t = Q + A^\top (S_{t+1} - S_{t+1} B (R + B^\top S_{t+1} B)^{-1} B^\top S_{t+1}) A$

- **Cost-to-go**: $V_t(x_t) = \frac{1}{2} x_t^\top S_t x_t + \sum_{t'=t+1}^T \frac{1}{2} \text{tr}(S_{t'} \Sigma_\omega)$ ← **noise-cost term, due to process noise**

- ▶ **Infinite horizon case**: $\lim_{T \rightarrow \infty} \frac{1}{T} V_0(x_0) = \lim_{T \rightarrow \infty} \frac{1}{2T} \left(x_0^\top S x_0 + \sum_{t=1}^T \text{tr}(S \Sigma_\omega) \right) = \frac{1}{2} \text{tr}(S \Sigma_\omega)$ ← **state independent**

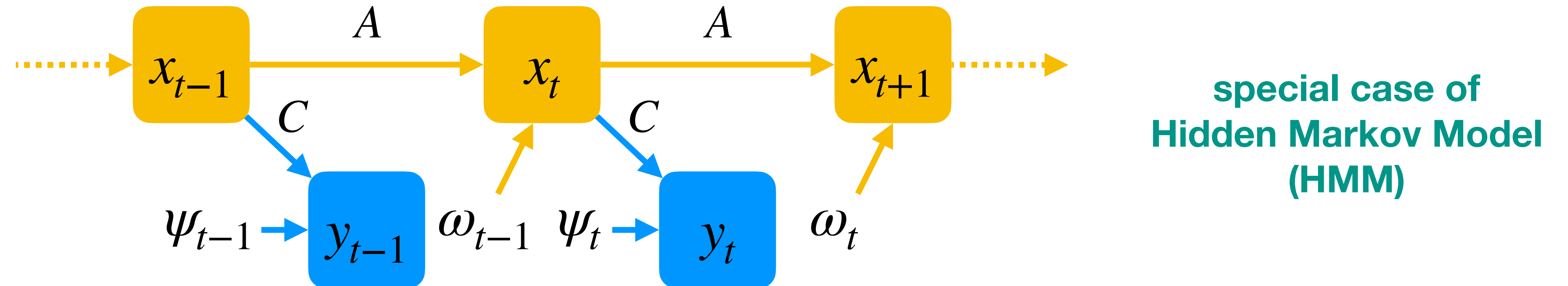
Today's lecture

LQR with process noise

Linear–Quadratic Estimator

Linear–Quadratic–Gaussian control

Partial observability



- What happens when we see just an **observation** $y_t \in \mathbb{R}^k$, not the full state x_t

- ▶ Simplest **observability model** – **Linear-Gaussian**: $p(y_t | x_t) = \mathcal{N}(y_t; Cx_t, \Sigma_\psi)$

$$y_t = Cx_t + \psi_t \quad \psi_t \sim \mathcal{N}(0, \Sigma_\psi) \quad C \in \mathbb{R}^{k \times n}, \Sigma_\psi \in \mathbb{R}^{k \times k}$$

- ▶ **Markov property**: all ω_t and ψ_t are independent, for all t
- Why is there **observation noise**?
 - ▶ **Transient** process noise that doesn't affect future states; only in agent's sensors

Gaussian Processes

- **Jointly Gaussian** variables: $\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \Sigma_{(x,y)} = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_y \end{bmatrix} \right)$

- ▶ **Conditional** distribution: $(x | y) \sim \mathcal{N}(\mu_{x|y}, \Sigma_{x|y})$

$$\mu_{x|y} = \mathbb{E}[x | y] = \mu_x + \Sigma_{xy} \Sigma_y^{-1} (y - \mu_y)$$

$$\Sigma_{x|y} = \text{Cov}[x | y] = \Sigma_x - \Sigma_{xy} \Sigma_y^{-1} \Sigma_{yx} = \Sigma_{(x,y)} / \Sigma_y$$

Schur complement

- ▶ **Converse** also true: Gaussian y and $(x | y)$, linear $\mu_{x|y} \implies$ jointly Gaussian (x, y)

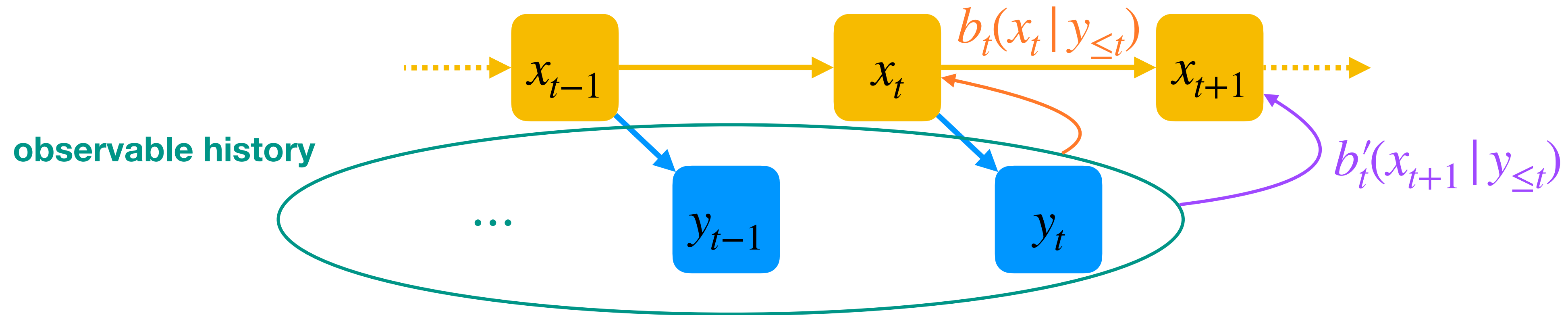
sufficient

- **Gaussian Process** (GP) $x_0, y_0, u_0, x_1, \dots$: all variables are (**pairwise**) jointly Gaussian

Linear–Quadratic Estimator (LQE)

- **Belief**: our distribution over state x_t given what we know
- Belief given past observations (**observable history**): $b_t(x_t | y_{\leq t})$
- b_t is **sufficient statistic** of $y_{\leq t}$ for x_t = nothing more $y_{\leq t}$ can tell us about x_t
 - In principle, we can update b_{t+1} only from b_t and y_{t+1} = **filtering**
 - Probabilistic Graphical Models terminology: **belief propagation**
- **Linear–Quadratic Estimator (LQE)**: belief for our Gaussian Process
 - Update equations = **Kalman filter**

Belief and prediction



- **Belief** = what the observable history says of **current state**: $b_t(x_t | y_{\leq t})$
- **Prediction** = what the observable history says of **next state**: $b'_t(x_{t+1} | y_{\leq t})$
- In this Gaussian Process, both belief and prediction are **Gaussian**
 - ▶ Can be represented by their means \hat{x}_t , \hat{x}'_{t+1} and covariances Σ_t , Σ'_{t+1}
 - ▶ Computed **recursively forward**

Kalman filter

- Given **belief** $b_t(x_t | y_{\leq t}) = \mathcal{N}(\hat{x}_t, \Sigma_t)$, **predict** x_{t+1} :

$$\hat{x}'_{t+1} = \mathbb{E}[x_{t+1} | y_{\leq t}] = \mathbb{E}[Ax_t + \omega_t | y_{\leq t}] = A\hat{x}_t$$

$$\Sigma'_{t+1} = \text{Cov}[x_{t+1} | y_{\leq t}] = \text{Cov}[Ax_t + \omega_t | y_{\leq t}] = A\Sigma_t A^\top + \Sigma_\omega$$

- Given **prediction** $b'_t(x_t | y_{<t}) = \mathcal{N}(\hat{x}'_t, \Sigma'_t)$, update **belief** of x_t on seeing y_t :

$$\hat{x}_t = \mathbb{E}[x_t | y_{\leq t}] = \mu_{x_t | y_{<t}} + \Sigma_{x_t y_t | y_{<t}} \Sigma_{y_t | y_{<t}}^{-1} (y_t - \mu_{y_t | y_{<t}})$$

$y_t = Cx_t + \text{noise} \implies \Sigma_{x_t y_t | y_{<t}} = \Sigma_{x_t | y_{<t}} C^\top$

prediction error / innovation e_t

like conditioning x_t on y_t
and doing this given $y_{<t}$

$$= \hat{x}'_t + \Sigma'_t C^\top (C \Sigma'_t C^\top + \Sigma_\psi)^{-1} (y_t - C \hat{x}'_t)$$

$\Sigma_{y_t | y_{<t}} = C \Sigma_{x_t | y_{<t}} C^\top + \Sigma_\psi$

$$\Sigma_t = \text{Cov}[x_t | y_{\leq t}] = \Sigma_{x_t | y_{<t}} - \Sigma_{x_t y_t | y_{<t}} \Sigma_{y_t | y_{<t}}^{-1} \Sigma_{y_t x_t | y_{<t}}$$

$$= \Sigma'_t - \Sigma'_t C^\top (C \Sigma'_t C^\top + \Sigma_\psi)^{-1} C \Sigma'_t$$

Kalman filter

- Linear belief update: $\hat{x}_t = A\hat{x}_{t-1} + K_t e_t = (I - K_t C)A\hat{x}_{t-1} + K_t y_t$
 $e_t = y_t - C\hat{x}_t$
- Kalman gain: $K_t = \Sigma'_t C^\top (C \Sigma'_t C^\top + \Sigma_\psi)^{-1}$
- Covariance update — Ricatti equation:

$$\Sigma'_{t+1} = A(\Sigma'_t - \Sigma'_t C^\top (C \Sigma'_t C^\top + \Sigma_\psi)^{-1} C \Sigma'_t) A^\top + \Sigma_\omega$$

- ▶ Compare to prior (no observations): $\Sigma_{x_{t+1}} = A \Sigma_{x_t} A^\top + \Sigma_\omega$
- Observations reduce covariance
 - ▶ Actual observation not needed to say by how much

Control as inference

- View Bayesian inference as optimization: **minimizes MSE** $\mathbb{E}[\|x_t - \hat{x}_t\|^2]$
- Control** and **inference** are deeply connected:

$$\Sigma'_{t+1} = A(\Sigma'_t - \Sigma'_t C^\top (C \Sigma'_t C^\top + \Sigma_\psi)^{-1} C \Sigma'_t) A^\top + \Sigma_\omega$$

$$S_t = Q + A^\top (S_{t+1} - S_{t+1} B (R + B^\top S_{t+1} B)^{-1} B^\top S_{t+1}) A$$

- The shared form (Ricatti) suggests **duality**:

LQR	LQE
backward	forward
S_{T-t}	Σ'_t
A	A^\top
B	C^\top
Q	Σ_ω
R	Σ_ψ

- Information filter**: recursion on $(\Sigma'_t)^{-1}$, presents a more principled duality

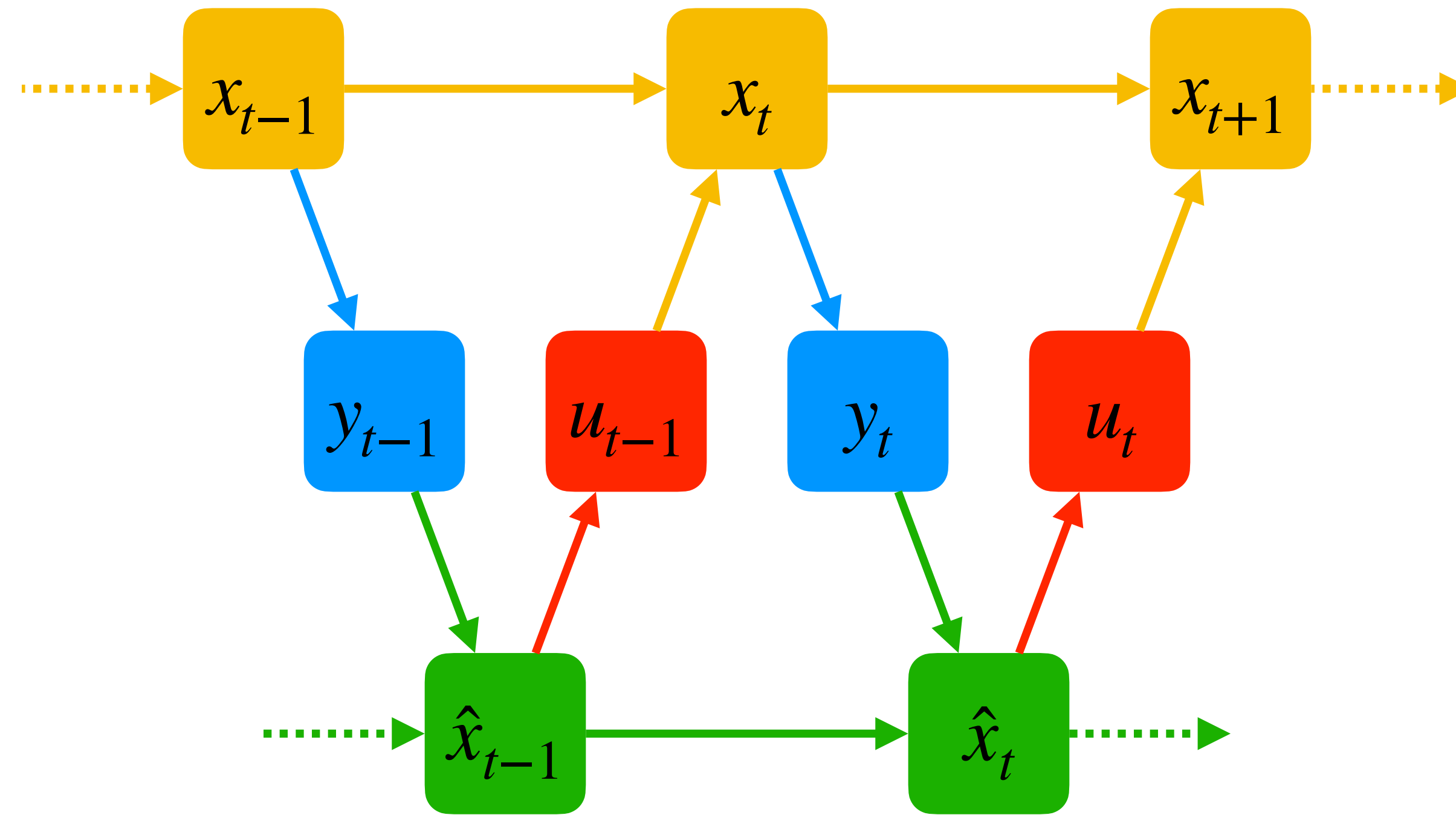
Today's lecture

LQR with process noise

Linear–Quadratic Estimator

Linear–Quadratic–Gaussian control

Linear–Quadratic–Gaussian (LQG) control



- Putting it all together:

$$x_{t+1} = Ax_t + Bu_t + \omega_t \quad \omega_t \sim \mathcal{N}(0, \Sigma_\omega) \quad \Sigma_\omega \in \mathbb{R}^{n \times n}$$

$$y_t = Cx_t + \psi_t \quad \psi_t \sim \mathcal{N}(0, \Sigma_\psi) \quad C \in \mathbb{R}^{k \times n}, \Sigma_\psi \in \mathbb{R}^{k \times k}$$

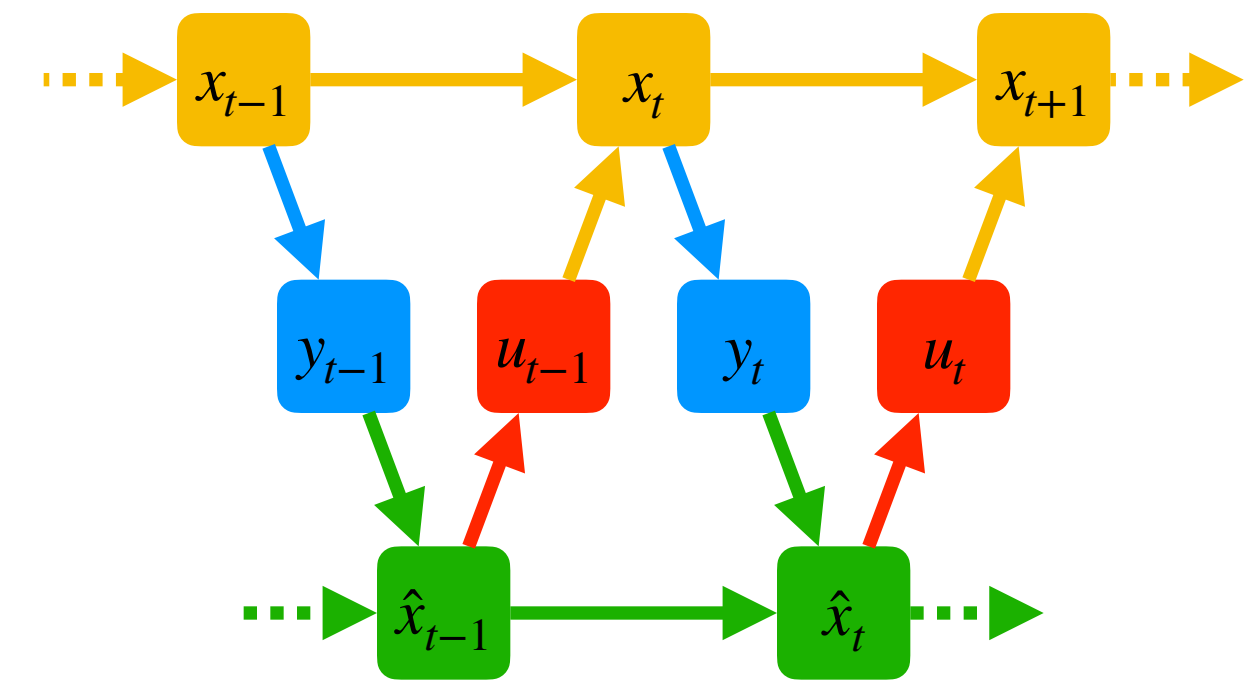
LQE with control

- How does control affect estimation?

- ▶ Shifts predicted next state $\hat{x}'_{t+1} = A\hat{x}_t + Bu_t$
- ▶ Bu_t known \Rightarrow no change in covariances \Rightarrow Ricatti equation still holds
- ▶ Same Kalman gain K_t

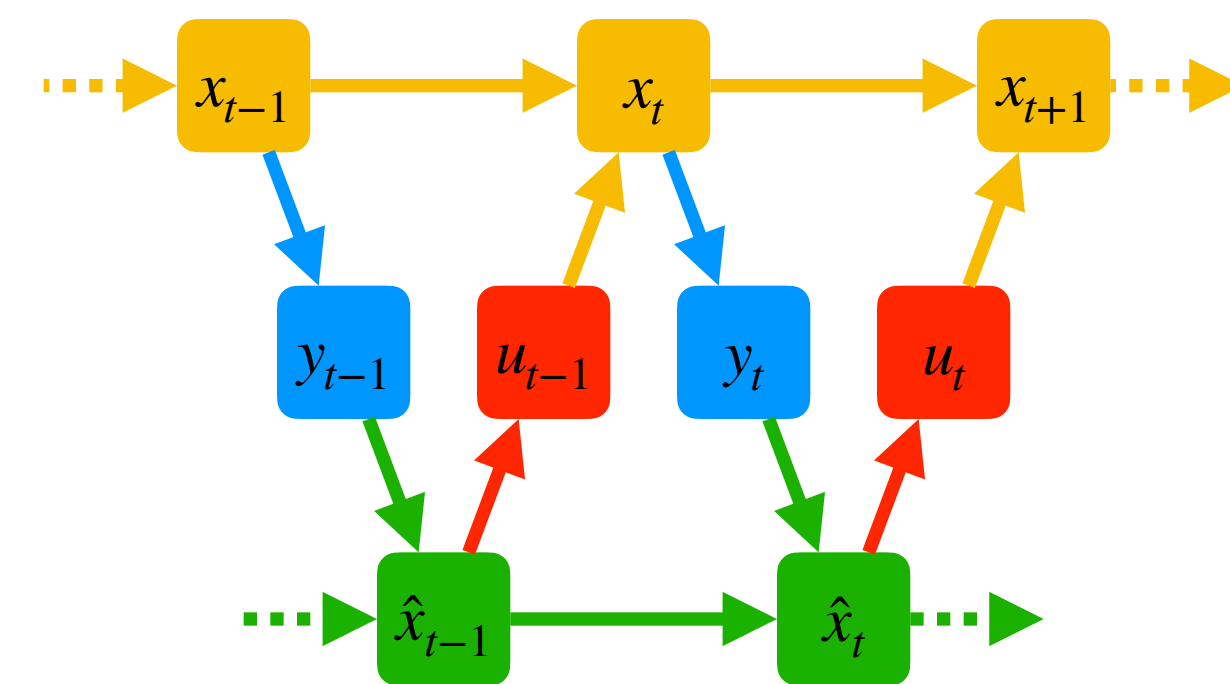
$$\hat{x}_t = A\hat{x}_{t-1} + K_t e_t = (I - K_t C)(A\hat{x}_{t-1} + Bu_{t-1}) + K_t y_t$$

- And... that's it, everything else the same



LQR with partial observability

- Bellman recursion must be expressed in terms of **what u_t can depend on** = \hat{x}_t



- Problem:** but value depends on the true state x_t
- Value recursion for **full state** (environment + agent):

$$V_t^\pi(x_t, \hat{x}_t) = c(x_t, u_t) + \mathbb{E}[V_{t+1}^\pi(x_{t+1}, \hat{x}_{t+1}) | x_t, \hat{x}_t]$$

- In terms of only \hat{x}_t :

\hat{x}_{t+1} is sufficient for x_{t+1}
 \Rightarrow separates it from \hat{x}_t

$$V_t^\pi(\hat{x}_t) = \mathbb{E}[V_t^\pi(x_t, \hat{x}_t) | \hat{x}_t] = \mathbb{E}[c(x_t, u_t) + V_{t+1}^\pi(x_{t+1}, \hat{x}_{t+1}) | \hat{x}_t] = \mathbb{E}[c(x_t, u_t) + V_{t+1}^\pi(\hat{x}_{t+1}) | \hat{x}_t]$$

- Certainty equivalent** control: $u_t = L_t \hat{x}_t$ with the same feedback gain L_t
- And... that's it, everything else the same

LQG separability

Given $(A, B, C, \Sigma_\omega, \Sigma_\psi, Q, R)$, solve LQG = LQR + LQE separately

- LQR:

- ▶ Compute value Hessian recursively backwards

$$S_t = Q + A^\top(S_{t+1} - S_{t+1}B(R + B^\top S_{t+1}B)^{-1}B^\top S_{t+1})A$$

- ▶ Compute feedback gain:

$$L_t = -(R + B^\top S_{t+1}B)^{-1}B^\top S_{t+1}A$$

- ▶ Control policy: $u_t = L_t \hat{x}_t$

- LQE:

- ▶ Compute belief covariance recursively forward

$$\Sigma'_{t+1} = A(\Sigma'_t - \Sigma'_t C^\top (C \Sigma'_t C^\top + \Sigma_\psi)^{-1} C \Sigma'_t) A^\top + \Sigma_\omega$$

- ▶ Compute Kalman gain:

$$K_t = \Sigma'_t C^\top (C \Sigma'_t C^\top + \Sigma_\psi)^{-1}$$

- ▶ Belief update: $\hat{x}_t = A \hat{x}_{t-1} + K_t e_t$

- ▶ with $e_t = y_t - C(A \hat{x}_{t-1} + B u_{t-1})$

Recap

- **Stochastic optimal control**: control with process noise (stochastic dynamics)
 - Same concepts of **controllability**, but can't stop at $x_t = 0$
- **LQE** = linear–Gaussian observability $y_t = Cx_t + \psi_t$
 - **Kalman filter** = forward recursion to find belief $b_t(x_t | y_{\leq t})$
- **LQG = LQR + LQE**: estimate and control at the same time
 - **Separability** = optimal to solve LQR and LQE separately (only in LQG!)
 - Only differences: **use \hat{x}_t** for control; **add Bu_t** to prediction