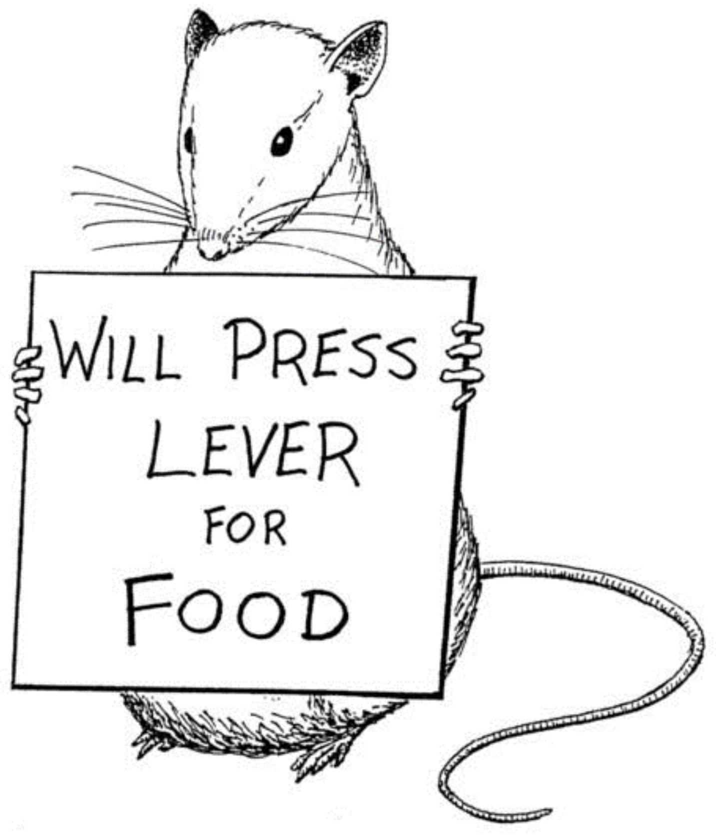
CS 277: Control and Reinforcement Learning

Winter 2021

Lecture 8: Stochastic Optimal Control

Roy Fox
Department of Computer Science
Bren School of Information and Computer Sciences
University of California, Irvine



Logistics

assignments

- Assignment 2 due this Friday

Today's lecture

Hamiltonian

LQR with process noise

Linear–Quadratic Estimator

Linear–Quadratic–Gaussian control

Optimal control: properties

- Linear control policy: $u_t = L_t x_t \qquad L_t \in \mathbb{R}^{m \times n}$
  - Feedback gain: $L_t = -(R + B^\intercal S_{t+1} B)^{-1} B^\intercal S_{t+1} A$
- Quadratic value (cost-to-go) function $\mathcal{J}_t(x_t)^* = \frac{1}{2} x_t^\intercal S_t x_t$
  - Cost Hessian $S_t = \nabla^2_{x_t} \mathcal{J}^*_t$ is the same for all x_t
- Ricatti equation for S_t can be solved recursively backward

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

  - Without knowing any actual states or controls (!) = at system design time
- Woodbury matrix identity shows $S_t = Q + A^\intercal (S^\dagger_{t+1} + B R^{-1} B^\intercal)^\dagger A \geq 0$

Infinite horizon

- Average cost: $\mathcal{J} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c(x_t, u_t)$

- For each finite T we solve with Bellman recursion, affected by end $\mathcal{J}_T = 0$

- In the limit, the solution must converge to time-independent

  - Discrete-time algebraic Ricatti equation (DARE):

$$S = Q + A^\intercal(S - SB(R + B^\intercal SB)^{-1}B^\intercal S)A$$

Non-homogeneous case

- More generally, LQR can have lower-order terms
  - $x_{t+1} = f_t(x_t, u_t) = A_t x_t + B_t u_t + c_t$
  - $c_t(x_t, u_t) = \frac{1}{2} x_t^\intercal Q_t x_t + \frac{1}{2} u_t^\intercal R_t u_t + u_t^\intercal N_t x_t + q_t^\intercal x_t + r_t^\intercal u_t + s_t$
- More flexible modeling, e.g. tracking a target trajectory $\frac{1}{2}(x_t - \tilde{x}_t)^\intercal Q (x_t - \tilde{x}_t)$
- Solved essentially the same way
  - Cost-to-go $\mathcal{J}$ will also have lower-order terms

Co-state

- Consider the cost-to-go $\mathcal{J}_t(x_t, u) = c(x_t, u_t) + \mathcal{J}_{t+1}(f(x_t, u_t), u)$

- To study its landscape over state space, consider its gradient

$$\nu_t = \nabla_{x_t} \mathcal{J}_t = \nabla_{x_t} c_t + \nabla_{x_{t+1}} \mathcal{J}_{t+1} \nabla_{x_t} f_t = \nabla_{x_t} c_t + \nu_{t+1} \nabla_{x_t} f_t$$

  - Co-state $\nu_t \in \mathbb{R}^n$ = direction of steepest increase in cost-to-go

  - Linear backward recursion, initialization: $\nu_T = 0$

  - $\nabla_{x_t} f_t$ = Jacobian of the dynamics

Lagrangian

- Constrained optimization: $\max_u g(u)$ s.t. $h(u) = 0$
  - Equivalent to Lagrangian (with Lagrange multiplier λ): $\max_u \min_\lambda g(u) + \lambda h(u)$
- Our optimization problem: $\min_u \mathcal{J}$ s.t. $x_{t+1} = f(x_t, u_t)$
  - Lagrangian: $\mathcal{L} = \sum_{t=0}^{T-1} c(x_t, u_t) + \nu_{t+1}(f(x_t, u_t) - x_{t+1})$
  - At the optimum: $\nabla_{x_t}\mathcal{L} = 0 \implies \nu_t = \nabla_{x_t} c_t + \nu_{t+1} \nabla_{x_t} f_t$ = the co-state
- Lagrange multipliers often have their own meaning

Hamiltonian

- Hamiltonian = first-order approximation of cost-to-go

$$\mathscr{H}_t = c(x_t, u_t) + \nu_{t+1} f(x_t, u_t)$$

- At the optimum, defines all x, u, and ν in one equation:

$$\nabla_{x_t} \mathscr{H}_t = \nabla_{x_t} c_t + \nu_{t+1} \nabla_{x_t} f_t = \nu_t$$

$$\nabla_{\nu_{t+1}} \mathscr{H}_t = f(x_t, u_t) = x_{t+1}$$

$$\nabla_{u_t} \mathscr{H}_t = \nabla_{u_t} (\mathscr{L} + \nu_{t+1} x_{t+1}) = 0$$

- Can solve these $(2n + m)T$ equations in $(2n + m)T$ variables
  - Generally, nonlinear with many local optima

Hamiltonian in LQR

- In LQR, the Hamiltonian is quadratic

$$\mathscr{H}_t = \frac{1}{2}x_t^\intercal Q x_t + \frac{1}{2}u_t^\intercal R u_t + \nu_{t+1}(Ax_t + Bu_t)$$

- This suggest forward–backward recursions for x, u, and ν:

$$x_{t+1} = \nabla_{\nu_{t+1}}\mathscr{H}_t = Ax_t + Bu_t$$

$$\nu_t = \nabla_{x_t}\mathscr{H}_t = \nu_{t+1}A + x_t^\intercal Q$$

$$\nabla_{u_t}\mathscr{H}_t = Ru_t + B^\intercal \nu_{t+1}^\intercal = 0$$

- These correspond to the previous approach with $\nu_t^\intercal = S_t x_t \quad u_t = L_t x_t$

Recap

- LQR = simplest dynamics: linear; simplest cost: quadratic
- Can characterize stability, reachability, stabilizability in terms of (A, B)
- Can use Ricatti equation to find cost-to-go Hessian
- Alternatively: Hamiltonian gives state forward / co-state backward recursions

Today's lecture

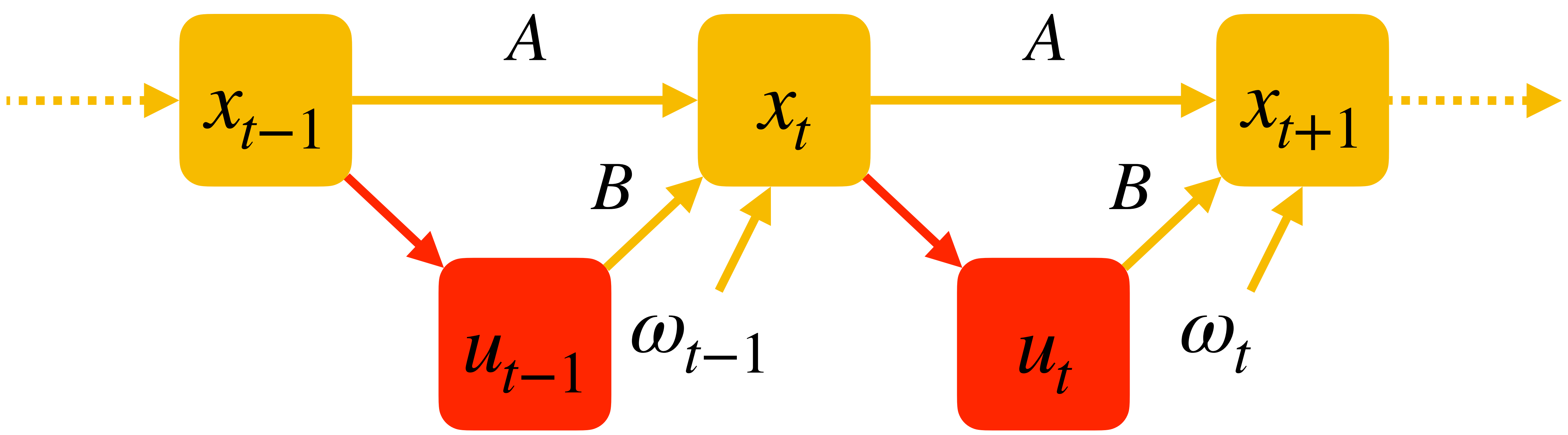
Hamiltonian

LQR with process noise

Linear–Quadratic Estimator

Linear–Quadratic–Gaussian control

Stochastic control



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

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

  - 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; maximum entropy if we only assume ω_t is not large

- In continuous time = Langevin equation; Bu_t = external force

Stochastic optimal control

- Minimize expected cost-to-go $\mathcal{J}_t(x_t, u_{\geq t}) = \sum_{t'=t}^{T-1} \mathbb{E}[c(x_{t'}, u_{t'}) \,|\, x_t, u_{\geq t}]$

$$= \mathbb{E}\left[\frac{1}{2}x_t^\intercal Q x_t + \frac{1}{2}u_t^\intercal R u_t + \mathcal{J}_{t+1}(x_{t+1}, u_{\geq t+1}) \,|\, x_t, u_{\geq t}\right]$$

- Bellman equation:

$$\mathcal{J}_t^*(x_t) = \min_{u_t} \mathbb{E}_{x_{t+1}|x_t,u_t \sim \mathcal{N}(Ax_t+Bu_t,\Sigma_\omega)}\left[\frac{1}{2}x_t^\intercal Q x_t + \frac{1}{2}u_t^\intercal R u_t + \mathcal{J}_{t+1}^*(x_{t+1})\right]$$

- Now the cost-to-go is quadratic, but with free term:

$$\mathcal{J}_t^*(x_t) = \frac{1}{2}x_t^\intercal S_t x_t + \mathcal{J}_t^*(0)$$

$x_t = 0$ **is no longer absorbing**

Solving the Bellman recursion

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

$$\mathcal{J}_t^*(x_t) = \min_{u_t} \mathbb{E}_{x_{t+1}|x_t,u_t\sim\mathcal{N}(Ax_t+Bu_t,\Sigma_\omega)}\left[\tfrac{1}{2}x_t^\intercal Qx_t+\tfrac{1}{2}u_t^\intercal Ru_t + \underbrace{\tfrac{1}{2}x_{t+1}^\intercal S_{t+1}x_{t+1}} + \mathcal{J}_{t+1}^*(0)\right]$$

$$= \min_{u_t}\left(\tfrac{1}{2}x_t^\intercal Qx_t+\tfrac{1}{2}u_t^\intercal Ru_t + \overbrace{\tfrac{1}{2}(Ax_t+Bu_t)^\intercal S_{t+1}(Ax_t+Bu_t)+\tfrac{1}{2}\mathrm{tr}(S_{t+1}\Sigma_\omega)} + \mathcal{J}_{t+1}^*(0)\right)$$

new term, constant

- Linear control: $u_t^* = L_t x_t$ with same feedback gain: $L_t = -(R+B^\intercal S_{t+1}B)^{-1}B^\intercal S_{t+1}A$

- Same Ricatti equation for cost-to-go Hessian: $S_t = Q + A^\intercal(S_{t+1} - S_{t+1}B(R+B^\intercal S_{t+1}B)^{-1}B^\intercal S_{t+1})A$

- Cost-to-go: $\mathcal{J}_t^*(x_t) = \frac{1}{2}x_t^\intercal S_t x_t + \sum_{t'=t+1}^{T}\frac{1}{2}\mathrm{tr}(S_{t'}\Sigma_\omega)$ noise–cost term, due to process noise

  - Infinite horizon case: $\lim_{T\to\infty}\frac{1}{T}\mathcal{J}_0^*(x_0) = \lim_{T\to\infty}\frac{1}{2T}\left(x_0^\intercal Sx_0 + \sum_{t=1}^{T}\mathrm{tr}(S\Sigma_\omega)\right) = \frac{1}{2}\mathrm{tr}(S\Sigma_\omega)$ state independent

Today's lecture

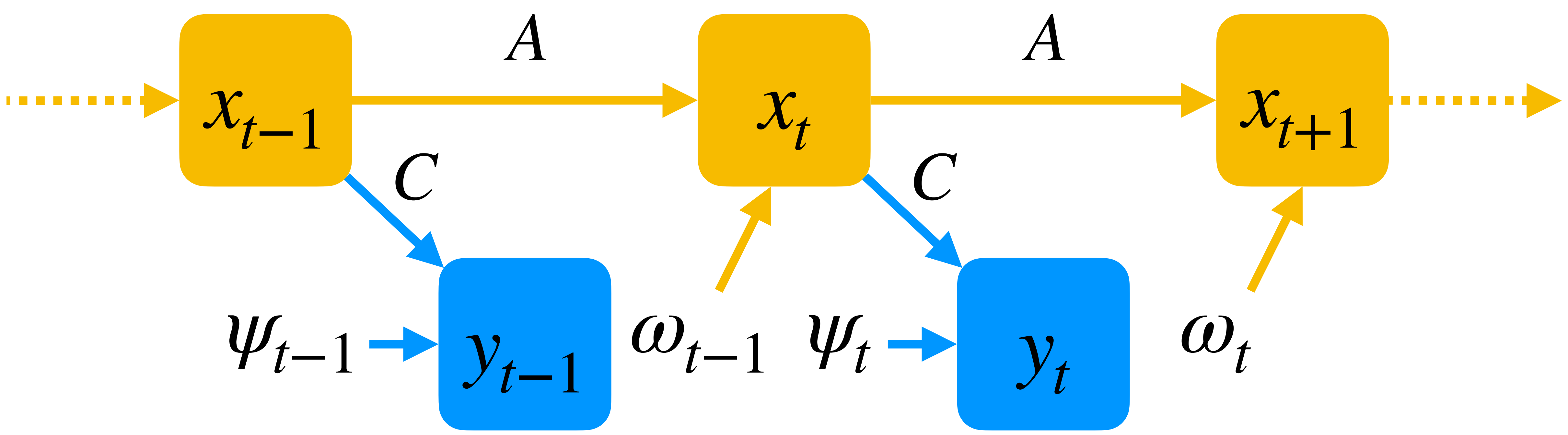
Hamiltonian

LQR with process noise

Linear–Quadratic Estimator

Linear–Quadratic–Gaussian control

Partial observability



special case of Hidden Markov Model (HMM)

- 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: $p(y_t \mid x_t) = \mathcal{N}(y_t; Cx_t, \Sigma_\psi)$

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

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

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

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

  - Converse also true: if y and $x \,|\, y$ are Gaussian $\Longrightarrow (x, y)$ jointly Gaussian

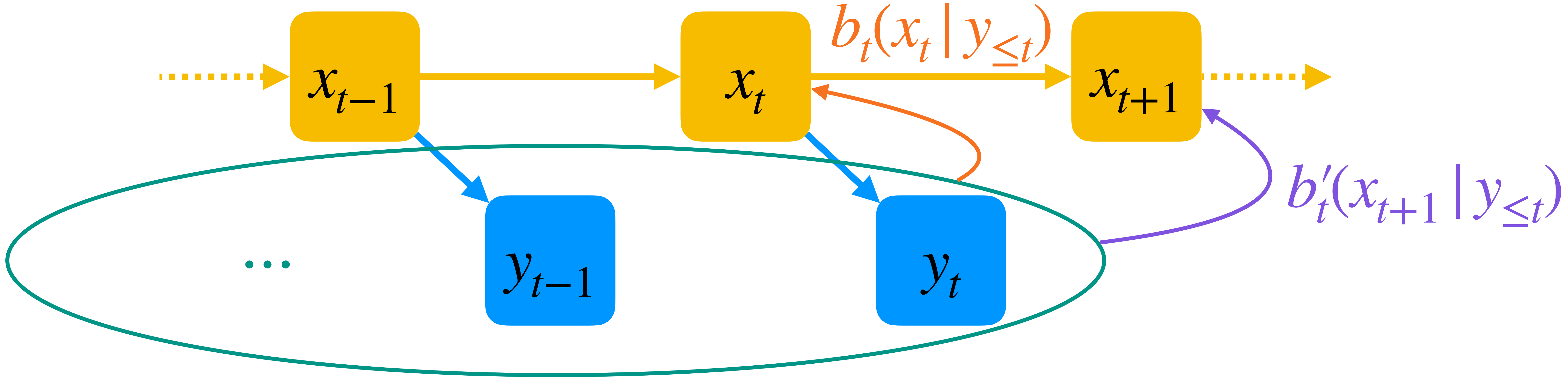
sufficient

- Gaussian Process (GP) $x_0, y_0, u_0, x_1, \ldots$: 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 observable history says of current state: $b_t(x_t \mid y_{\le t})$
- Prediction = what observable history says of next state: $b'(x_{t+1} \mid y_{\le t})$
- In this Gaussian Process, both 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 $b_t(x_t | y_{\le t}) = \mathcal{N}(\hat{x}_t, \Sigma_t)$, predict x_{t+1}:

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

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

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

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

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

prediction error / innovation e_t

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

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

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

$$\Sigma_t = \mathrm{Cov}[x_t | y_{\le t}] = \Sigma_{x_t | y_{<t}} - \Sigma_{x_t y_t | y_{<t}} \Sigma^{-1}_{y_t | y_{<t}} \Sigma_{y_t x_t | y_{<t}}$$

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

Kalman filter

$e_t = y_t - C\hat{x}'_t$

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

- Kalman gain: $K_t = \Sigma'_t C^\intercal (C\Sigma'_t C^\intercal + \Sigma_\psi)^{-1}$

- Covariance update — Ricatti equation:

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

  - Compare to prior (no observations): $\Sigma_{x_{t+1}} = A\Sigma_{x_t} A^\intercal + \Sigma_\omega$

  - Observations help, but actual observation not needed to say by how much

Control as inference

- View Bayesian inference as optimization: minimizes MSE $\mathbb{E}[(x_t - \hat{x}_t)]$

- Control and inference are deeply connected:

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

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

- The shared form (Ricatti) suggests duality:

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

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

Today's lecture

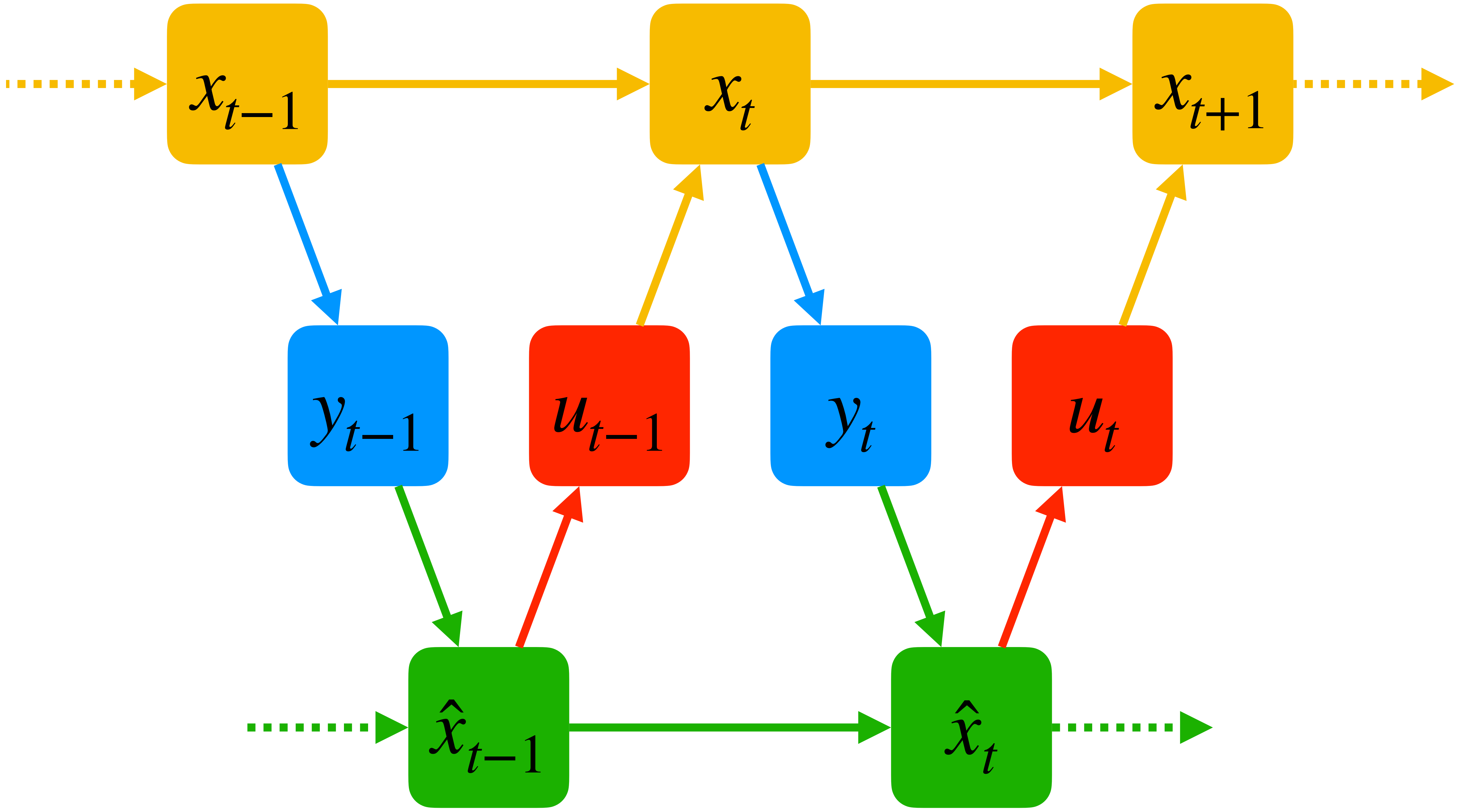
Hamiltonian

LQR with process noise

Linear–Quadratic Estimator

Linear–Quadratic–Gaussian control

Linear–Quadratic–Gaussian (LQG) control



- Putting it all together:

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

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

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 $\implies$ no change in covariances, 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:

$$\mathcal{J}_t(x_t, \hat{x}_t, u) = \mathbb{E}[c(x_t, u_t) + \mathcal{J}_{t+1}(x_{t+1}, \hat{x}_{t+1}, u) \,|\, x_t, \hat{x}_t]$$

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

works because $\hat{x}_{t+1}$ is sufficient for x_{t+1}, separating it from $\hat{x}_t$

$$\mathcal{J}_t(\hat{x}_t, u) = \mathbb{E}[\mathcal{J}_t(x_t, \hat{x}_t, u) \,|\, \hat{x}_t] = \mathbb{E}[c(x_t, u_t) + \mathcal{J}_{t+1}(x_{t+1}, \hat{x}_{t+1}, u) \,|\, \hat{x}_t] = \mathbb{E}[c(x_t, u_t) + \mathcal{J}_{t+1}(\hat{x}_{t+1}, u) \,|\, \hat{x}_t]$$

- 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

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

  - Compute feedback gain:

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

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

- LQE:
  - Compute belief covariance recursively forward

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

  - Compute Kalman gain:

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

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

Extensive cost-to-go term

- Optimal cost-to-go: $\mathcal{J}_t^*(x_t) = \frac{1}{2} x_t^\intercal S_t x_t + \mathcal{J}_t^*(0)$

- Extensive (linear in T) term:

$$\mathcal{J}_t^*(0) = \frac{1}{2} \sum_{t'=t}^{T} (\underbrace{\mathrm{tr}(Q\Sigma_{t'})} + \underbrace{\mathrm{tr}(S_{t'+1}(\Sigma'_{t'+1} - \Sigma_{t'+1}))})$$

immediate cost of uncertainty in x_t

cost-to-go of uncertainty added by 1-step prediction

- Infinite horizon: $\mathcal{J}^* = \frac{1}{2}\mathrm{tr}(Q\Sigma) + \frac{1}{2}\mathrm{tr}(S(\Sigma' - \Sigma))$

    - S and Σ' are solutions of algebraic Ricatti equation