Optimal Control Theory and Linear Quadratic Regulators

Sham Kakade and Wen Sun

CS 6789: Foundations of Reinforcement Learning
Today

- Recap:
  - TRPO/PPO

- Today: LQRs
  - The model + planning + SDP formulations
  - LQRs are MDPs with special structure
Recap
TRPO: second order Taylor's expansion

\[
\max_{\pi_\theta} V^{\pi_\theta}(\rho) \quad \text{s.t., } KL(Pr_{\pi_\theta} || Pr_{\pi_0}) \leq \delta
\]

\[
\max_{\theta} \nabla V^{\pi_\theta_0}(\rho)^\top (\theta - \theta_0)
\quad \text{s.t. } (\theta - \theta_0)^\top F_{\theta_0}(\theta - \theta_0) \leq \delta
\]

We have a closed form solution:

\[
\theta = \theta_0 + \sqrt{\frac{\delta}{(\nabla V^{\pi_\theta_0})^\top F_{\theta_0}^{-1} \nabla V^{\pi_\theta_0} \cdot F_{\theta_0}^{-1} \nabla V^{\pi_\theta_0}}} 
\]

• Self-normalized step-size
  (Learning rate is adaptive)
• Solve with CG
PPO

• To find the next policy $\pi_{t+1}$, use objective:

$$\max_{\theta} \mathbb{E}_{s \sim d} \mathbb{E}_{a \sim \pi(\cdot | s)} A_{\pi_{t}}(s, a)$$

subject to

$$\sup_{s} \left\| \pi^0(\cdot | s) - \pi_t(\cdot | s) \right\|_{TV} \leq \delta,$$

This is like the CPI greedy policy chooser.
PPO

• To find the next policy $\pi_{t+1}$, use objective:

$$\max_{\theta} E_{s \sim d_{\pi_t}} E_{a \sim \pi^\theta(\cdot | s)} A^{\pi_t}(s, a)$$

subject to $\sup_s \| \pi^\theta(\cdot | s) - \pi_t(\cdot | s) \|_{TV} \leq \delta,$

This is like the CPI greedy policy chooser.

• We can do multiple gradient steps by rewriting the objective function using importance weighting:

$$\max_{\theta} E_{s \sim d_{\pi_t}} E_{a \sim \pi_t(\cdot | s)} \left[ \frac{\pi^\theta(a | s)}{\pi_t(\cdot | s)} A^{\pi_t}(s, a) \right]$$

practice: enforce constraint by just changing $\theta$ a “little” (say with a few gradient steps)
Today:

Natural Policy Gradient and Approximation
Robotics and Controls

Dexterous Robotic Hand Manipulation
OpenAI, 2019
The LQR Model
Optimal Control

- a dynamical system is described as

\[ x_{t+1} = f_t(x_t, u_t, w_t) \]

where \( f_t \) maps a state \( x_t \in \mathbb{R}^d \), a control (the action) \( u_t \in \mathbb{R}^k \), and a disturbance \( w_t \), to the next state \( x_{t+1} \in \mathbb{R}^d \), starting from an initial state \( x_0 \).
Optimal Control

- A dynamical system is described as
  \[ x_{t+1} = f_t(x_t, u_t, w_t) \]
  where \( f_t \) maps a state \( x_t \in \mathbb{R}^d \), a control (the action) \( u_t \in \mathbb{R}^k \), and a disturbance \( w_t \), to the next state \( x_{t+1} \in \mathbb{R}^d \), starting from an initial state \( x_0 \).

- The objective is to find the control policy \( \pi \) which minimizes the long term cost,
  \[
  \min_{\pi} E_{\pi} \left[ \sum_{t=0}^{H-1} c_t(x_t, u_t) \right]
  \]
  such that
  \[ x_{t+1} = f_t(x_t, u_t, w_t) \]
  where \( H \) is the time horizon (which can be finite or infinite) and where \( w_t \) is either statistical or constrained in some way.
Linearization Approach
Linearization Approach

- In practice, this is often solved by considering the linearized control (sub-)problem where the dynamics are approximated by

\[ x_{t+1} = A_t x_t + B_t u_t + w_t, \]

with the matrices \( A_t \) and \( B_t \) are derivatives of the dynamics \( f \) (around some trajectory) and where the costs are approximated by a quadratic function in \( x_t \) and \( u_t \).
Linearization Approach

- In practice, this is often solved by considering the linearized control (sub-)problem where the dynamics are approximated by

\[ x_{t+1} = A_t x_t + B_t u_t + w_t, \]

with the matrices \( A_t \) and \( B_t \) are derivatives of the dynamics \( f \) (around some trajectory) and where the costs are approximated by a quadratic function in \( x_t \) and \( u_t \).

- This linearization is often accurate provided the noise is ‘small’ and the dynamics are ‘smooth’. (The details are important).
Linearization Approach

- In practice, this is often solved by considering the linearized control (sub-)problem where the dynamics are approximated by
  \[ x_{t+1} = A_t x_t + B_t u_t + w_t, \]
  with the matrices \( A_t \) and \( B_t \) are derivatives of the dynamics \( f \) (around some trajectory) and where the costs are approximated by a quadratic function in \( x_t \) and \( u_t \).

- This linearization is often accurate provided the noise is ‘small’ and the dynamics are ‘smooth’. (The details are important).

- This approach does not capture global information.
The Linear Quadratic Regulator (LQR) (finite horizon case)
The Linear Quadratic Regulator (LQR)  
(finite horizon case)  

• Let’s suppose this local approximation to a non-linear model is globally valid.  
(clearly false but this is an effective approach once when we ‘close’).
The Linear Quadratic Regulator (LQR)
(finite horizon case)

• Let’s suppose this local approximation to a non-linear model is globally valid.
  (clearly false but this is an effective approach once when we ‘close’).

• The finite horizon LQR problem is given by

\[
\begin{array}{l}
\text{minimize } E \left[ x_H^T Q x_H + \sum_{t=0}^{H-1} (x_t^T Q x_t + u_t^T R u_t) \right] \\
\text{such that } x_{t+1} = A_t x_t + B_t u_t + w_t, \quad x_0 \sim D, \ w_t \sim N(0, \sigma^2 I),
\end{array}
\]

where initial state \( x_0 \sim D \) is randomly distributed according \( D \);
the disturbance \( w_t \in \mathbb{R}^d \) is multi-variate normal, with covariance \( \sigma^2 I \);
\( A_t \in \mathbb{R}^{d\times d} \) and \( B_t \in \mathbb{R}^{d\times k} \) are referred to as system (or transition) matrices;
\( Q \in \mathbb{R}^{d\times d} \) and \( R \in \mathbb{R}^{k\times k} \) are psd matrices that parameterize the quadratic costs.
The Linear Quadratic Regulator (LQR)  
(finite horizon case) 

- Let’s suppose this local approximation to a non-linear model is globally valid. (clearly false but this is an effective approach once when we ‘close’).

- The finite horizon LQR problem is given by

\[
\text{minimize } \mathbb{E}\left[ x_H^T Q x_H + \sum_{t=0}^{H-1} (x_t^T Q x_t + u_t^T R u_t) \right]
\]

such that \( x_{t+1} = A_t x_t + B_t u_t + w_t, \ x_0 \sim D, \ w_t \sim \mathcal{N}(0, \sigma^2 I), \)

where initial state \( x_0 \sim D \) is randomly distributed according \( D \);
the disturbance \( w_t \in \mathbb{R}^d \) is multi-variate normal, with covariance \( \sigma^2 I \);
\( A_t \in \mathbb{R}^{d \times d} \) and \( B_t \in \mathbb{R}^{d \times k} \) are referred to as system (or transition) matrices; \( Q \in \mathbb{R}^{d \times d} \) and \( R \in \mathbb{R}^{k \times k} \) are psd matrices that parameterize the quadratic costs.

- Note that this model is a finite horizon MDP, where the \( S = R^d \) and \( A = R^k \).
The Linear Quadratic Regulator (LQR)  
(infinite horizon case)
The Linear Quadratic Regulator (LQR)  
(infinite horizon case)

- The infinite horizon LQR problem is given by

\[
\text{minimize} \quad \lim_{H \to \infty} \frac{1}{H} \mathbb{E} \left[ \sum_{t=0}^{H} (x_t^T Q x_t + u_t^T R u_t) \right]
\]

such that \( x_{t+1} = A x_t + B u_t + w_t, \quad x_0 \sim D, \quad w_t \sim N(0, \sigma^2 I) \).

where \( A \) and \( B \) are time homogenous.
The Linear Quadratic Regulator (LQR)
(infinite horizon case)

• The infinite horizon LQR problem is given by

\[
\text{minimize} \quad \lim_{H \to \infty} \frac{1}{H} E \left[ \sum_{t=0}^{H} (x_t^T Q x_t + u_t^T R u_t) \right]
\]

such that

\[
x_{t+1} = Ax_t + Bu_t + w_t, \quad x_0 \sim D, \quad w_t \sim N(0, \sigma^2 I).
\]

where \(A\) and \(B\) are time homogenous.

• Studied often in theory, but less relevant in practice (?
(largely due to that time homogenous, globally linear models are rarely good approximations)
The Linear Quadratic Regulator (LQR)
(infinite horizon case)

- The infinite horizon LQR problem is given by

\[
\text{minimize } \lim_{H \to \infty} \frac{1}{H} \mathbb{E} \left[ \sum_{t=0}^{H} (x_t^TQx_t + u_t^TRu_t) \right]
\]

such that

\[ x_{t+1} = Ax_t + Bu_t + w_t, \quad x_0 \sim D, \quad w_t \sim N(0, \sigma^2 I). \]

where \( A \) and \( B \) are time homogenous.

- Studied often in theory, but less relevant in practice (\(?\))
  (largely due to that time homogenous, globally linear models are rarely good approximations)

- Discounted case never studied.
  (discounting doesn’t necessarily make costs finite)
The Linear Quadratic Regulator (LQR)
(infinite horizon case)

- The infinite horizon LQR problem is given by

\[
\text{minimize } \lim_{H \to \infty} \frac{1}{H} E \left[ \sum_{t=0}^{H} (x_t^T Q x_t + u_t^T R u_t) \right]
\]

such that \( x_{t+1} = A x_t + B u_t + w_t, \quad x_0 \sim D, \quad w_t \sim N(0, \sigma^2 I) \).

where \( A \) and \( B \) are time homogenous.

- Studied often in theory, but less relevant in practice (?) (largely due to that time homogenous, globally linear models are rarely good approximations)

- Discounted case never studied.
  (discounting doesn’t necessarily make costs finite)

- Note that we can have ‘unbounded’ average cost.
Bellman Optimality: Value Iteration and the Ricatti Equations

What do the values look like in an LQR??

$LQR \text{ vs. Lin MDP}$

$s \in \mathcal{S}$ arbitrary
\[ \phi(s, a) \in \mathbb{R}^d. \]
\[ s \sim P(s' | s, a) \]
\[ P(s' | s, a) = \mathbb{E}[s'] \cdot \phi(s, a) \]

$LQR$ vs. Lin MDP

\[ P(s' | s, a) = \mathbb{E}[s'] \cdot \phi(s, a) \]

$s \in \mathcal{S}$ arbitrary
\[ \phi(s, a) \in \mathbb{R}^d. \]

$LQR \text{ vs. Lin MDP}$

$s \in \mathcal{S}$ arbitrary
\[ \phi(s, a) \in \mathbb{R}^d. \]
\[ s \sim P(s' | s, a) \]
\[ P(s' | s, a) = \mathbb{E}[s'] \cdot \phi(s, a) \]

$LQR$ vs. Lin MDP

$s \in \mathcal{S}$ arbitrary
\[ \phi(s, a) \in \mathbb{R}^d. \]
\[ s \sim P(s' | s, a) \]
\[ P(s' | s, a) = \mathbb{E}[s'] \cdot \phi(s, a) \]

$LQR$ vs. Lin MDP

$s \in \mathcal{S}$ arbitrary
\[ \phi(s, a) \in \mathbb{R}^d. \]
\[ s \sim P(s' | s, a) \]
\[ P(s' | s, a) = \mathbb{E}[s'] \cdot \phi(s, a) \]
Same defs (but for costs)

- define the value function $V_h^\pi : R^d \rightarrow R$ as
  $$V_h^\pi(x) = E_x^\pi \left[ x_H^T Q x_H + \sum_{t=h}^{H-1} (x_t^T Q x_t + u_t^T R u_t) \right| \pi, x_h = x],$$

- and the state-action value $Q_h^\pi : R^d \times R^k \rightarrow R$ as:
  $$Q_h^\pi(x, u) = E_x^\pi \left[ x_H^T Q x_H + \sum_{t=h}^{H-1} (x_t^T Q x_t + u_t^T R u_t) \right| \pi, x_h = x, u_h = u].$$
Value Iteration and the Ricatti Equations
Value Iteration and the Ricatti Equations

Theorem: (for the finite horizon case, with time homogenous $A_t = A, B_t = B$)

The optimal policy is a linear controller specified by:

$$\pi^*(x_t) = -K^* x_t$$

where

$$K^* = (B^T P_{t+1} B + R)^{-1} B^T P_{t+1} A$$
Value Iteration and the Ricatti Equations

Theorem: (for the finite horizon case, with time homogenous \( A_t = A, B_t = B \))

The optimal policy is a linear controller specified by:

\[
\pi^*(x_t) = -K^*_t x_t \quad \text{where} \quad K^*_t = (B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A
\]

where \( P_t \) can be computed iteratively, in a backwards manner, using the following algebraic Ricatti equations, where for \( t \in [H] \),

\[
P_t = A^\top P_{t+1} A + Q - A^\top P_{t+1} B (B^\top P_{t+1} B + R)^{-1} B^\top P_{t+1} A
\]

\[
= A^\top P_{t+1} A + Q - (K^*_{t+1})^\top (B^\top P_{t+1} B + R) K^*_{t+1}
\]

and where \( P_H = Q \).
Value Iteration and the Ricatti Equations

Theorem: (for the finite horizon case, with time homogenous $A_t = A, B_t = B$)

The optimal policy is a linear controller specified by:

$$\pi^*(x_t) = -K_t^*x_t$$

where $K_t^* = (B^TP_{t+1}B + R)^{-1}B^TP_{t+1}A$

where $P_t$ can be computed iteratively, in a backwards manner, using the following algebraic Ricatti equations, where for $t \in [H]$,

$$P_t = A^TP_{t+1}A + Q - A^TP_{t+1}B(B^TP_{t+1}B + R)^{-1}B^TP_{t+1}A$$

$$= A^TP_{t+1}A + Q - (K_{t+1}^*)^T(B^TP_{t+1}B + R)K_{t+1}^*$$

and where $P_H = Q$.

The above equation is simply the value iteration algorithm.
Value Iteration and the Ricatti Equations

Theorem: (for the finite horizon case, with time homogenous \( A_t = A, B_t = B \))

The optimal policy is a linear controller specified by:
\[
\pi^*(x_t) = -K_t^*x_t \text{ where } K_t^* = (B^TP_{t+1}B + R)^{-1}B^TP_{t+1}A
\]

where \( P_t \) can be computed iteratively, in a backwards manner, using the following algebraic Ricatti equations, where for \( t \in [H] \),
\[
P_t = A^TP_{t+1}A + Q - A^TP_{t+1}B(B^TP_{t+1}B + R)^{-1}B^TP_{t+1}A
\]
\[
= A^TP_{t+1}A + Q - (K_{t+1}^*)^T(B^TP_{t+1}B + R)K_{t+1}^*
\]

and where \( P_H = Q \).

The above equation is simply the value iteration algorithm.

Furthermore, for \( t \in [H] \), we have that:
\[
V_t^*(x) = x^TP_tx + \sigma^2\text{Trace}(P_{t+1})
\]
Proof: optimal control at $h = H - 1$

- Bellman equations $\Rightarrow$ there is an optimal policy which is deterministic and stationary.
Proof: optimal control at $h = H - 1$

- Bellman equations $\Rightarrow$ there is an optimal policy which is deterministic and stationary.
- Due to that $x_H = Ax + Bu + w_{H-1}$, we have:

$$Q_{H-1}(x, u) = E[(Ax + Bu + w_{H-1})^T Q(Ax + Bu + w_{H-1})] + x^T Q x + u^T R u$$

$$= (Ax + Bu)^T Q(Ax + Bu) + \sigma^2 \text{Trace}(Q) + x^T Q x + u^T R u$$
Proof: optimal control at $h = H - 1$

- Bellman equations $\Rightarrow$ there is an optimal policy which is deterministic and stationary.
- Due to that $x_H = Ax + Bu + w_{H-1}$, we have:
  \[ Q_{H-1}(x, u) = E[(Ax + Bu + w_{H-1})^T Q(Ax + Bu + w_{H-1})] + x^T Qx + u^T Ru \]
  \[ = (Ax + Bu)^T Q(Ax + Bu) + \sigma^2 \text{Trace}(Q) + x^T Qx + u^T Ru \]
- This is a quadratic function of $u$. Solving for the optimal control at $x$, gives:
  \[ \pi^*_H(x) = - (B^T QB + R)^{-1} B^T QAx = - K^*_H x, \]
  where the last step uses that $P_H := Q$.

\[ = \arg\min_u Q_{h+1}(x, u) \]
Proof: optimal value at $h = H - 1$
Proof: optimal value at $h = H - 1$

- (shorthand $K_{H-1}^* = K$). using the optimal control at:

$$V_{H-1}^*(x) = Q_{H-1}(x, -K_{H-1}^*x) = Q_{H-1}(x, x) = Q_{H-1}(x, x) + x^T K^T R K x - \sigma^2 \text{Trace}(Q)$$
Proof: optimal value at \( h = H - 1 \)

- (shorthand \( K^*_H = K \)). using the optimal control at:
  \[
  V_{H-1}^*(x) = Q_{H-1}(x, -K^*_{H-1}x)
  = x^T(A - BK)^TQ(A - BK)x + x^TQx + x^TK^TRKx - \sigma^2\text{Trace}(Q)
  \]

- Continuing
  \[
  V_{H-1}^*(x) - \sigma^2\text{Trace}(Q) = x^T\left((A - BK)^TQ(A - BK) + Q + K^TRK\right)x
  = x^T\left(AQA + Q - 2K^TB^TQA + K^T(B^TQB + R)K\right)x
  = x^T\left(AQA + Q - 2K^T(B^TQB + R)K + K^T(B^TQB + R)K\right)x
  = x^T\left(AQA + Q - K^T(B^TQB + R)K\right)x
  = x^TP_{H-1}x.
  \]
  where the fourth step uses our expression for \( K = K^*_H \).
Proof: wrapping up...
Proof: wrapping up…

• This implies that:

\[ Q^*_H^{-2}(x, u) = E \left[ V^*_H(Ax + Bu + w_H^{-2}) \right] + x^T Q x + u^T R u \]

\[ = (Ax + Bu)^T P_{H-1} (Ax + Bu) + \sigma^2 \text{Trace}(P_{H-1}) + x^T Q x + u^T R u. \]
Proof: wrapping up...

- This implies that:

$$Q^*_{H-2}(x, u) = E[V^*_H(Ax + Bu + w_{H-2})] + x^\top Qx + u^\top Ru$$

$$= (Ax + Bu)^\top P_{H-1}(Ax + Bu) + \sigma^2 \text{Trace}(P_{H-1}) + x^\top Qx + u^\top Ru.$$  

- The remainder of the proof follows from a recursive argument, which can be verified along identical lines to the $t = H - 1$ case.
Infinite horizon case
Infinite horizon case

Theorem:
Suppose that the optimal average cost is finite.
Infinite horizon case

Theorem:
Suppose that the optimal average cost is finite.
Let $P$ be a solution to the following algebraic Riccati equation:
\[
P = A^TPA + Q - A^TPB(B^TPB + R)^{-1}B^TPA.
\]
(Note that $P$ is a positive definite matrix).
Infinite horizon case

Theorem:
Suppose that the optimal average cost is finite.
Let $P$ be a solution to the following algebraic Riccati equation:

$$P = A^T PA + Q - A^T PB (B^T PB + R)^{-1} B^T PA.$$ 

(Note that $P$ is a positive definite matrix).

We have that the optimal policy is:

$$\pi^*(x) = - K^* x$$

where the optimal control gain is:

$$K^* = - (B^T PB + R)^{-1} B^T PA.$$
Infinite horizon case

Theorem:
Suppose that the optimal average cost is finite.
Let $P$ be a solution to the following algebraic Riccati equation:

$$ P = A^T P A + Q - A^T P B (B^T P B + R)^{-1} B^T P A. $$

(Note that $P$ is a positive definite matrix).

We have that the optimal policy is:

$$ \pi^*(x) = - K^* x $$

where the optimal control gain is:

$$ K^* = - (B^T P B + R)^{-1} B^T P A $$

We have that $P$ is unique and that the optimal average cost is $\sigma^2 \text{Trace}(P)$. 
Semidefinite Programs to find $P$
The Primal SDP:
(for the infinite horizon LQR)

- The primal optimization problem is given as:
  
  \[
  \text{maximize} \quad \sigma^2 \text{Trace}(P) \\
  \text{subject to} \quad \begin{bmatrix} A^TPA + Q - I & A^TPB \\ B^TPA & B^TPB + R \end{bmatrix} \succeq 0, \quad P \succeq 0
  \]
  
  where the optimization variable is $P$. 
The Primal SDP:
(for the infinite horizon LQR)

• The primal optimization problem is given as:

\[
\begin{align*}
\text{maximize} & \quad \sigma^2 \text{Trace}(P) \\
\text{subject to} & \quad \begin{bmatrix}
A^T PA + Q - I & A^T PB \\
B^T PA & B^T PB + R
\end{bmatrix} \succeq 0, \quad P \succeq 0
\end{align*}
\]

where the optimization variable is \( P \).

• This SDP has a unique solution, \( P^* \), which implies:
  • \( P^* \) satisfies the Ricatti equations.
  • The optimal average cost of the infinite horizon LQR is \( \sigma^2 \text{Trace}(P^*) \)
  • The optimal policy use the gain matrix: \( K^* = - (B^T PB + R)^{-1} B^T PA \)
The Primal SDP:
(for the infinite horizon LQR)

- The primal optimization problem is given as:

  \[
  \begin{align*}
  \text{maximize} & \quad \sigma^2 \text{Trace}(P) \\
  \text{subject to} & \quad \begin{bmatrix} A^T PA + Q - I & A^T PB \\ B^T PA & B^T PB + R \end{bmatrix} \succeq 0, \quad P \succeq 0
  \end{align*}
  \]

  where the optimization variable is \( P \).

- This SDP has a unique solution, \( P^* \), which implies:
  - \( P^* \) satisfies the Ricatti equations.
  - The optimal average cost of the infinite horizon LQR is \( \sigma^2 \text{Trace}(P^*) \)
  - The optimal policy use the gain matrix: \( K^* = -(B^T PB + R)^{-1} B^T PA \)

- Proof idea: Following from the Ricatti equation, we have the relaxation that for all matrices \( K \), the matrix \( P \) must satisfy:

  \[
  P \succeq (A - BK)^T P (A - BK) + Q - K^T R K.
  \]
The Dual SDP:

- The dual optimization problem is:

\[
\begin{align*}
\text{minimize} & \quad \text{Trace} \left( \Sigma \cdot \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \right) \\
\text{subject to} & \quad \Sigma_{xx} = (A \ B) \Sigma (A \ B)^\top + \sigma^2 I, \quad \Sigma \geq 0
\end{align*}
\]

where the optimization variable is \( \Sigma \), a \((d + k) \times (d + k)\) matrix, with the block structure:

\[
\Sigma = \begin{bmatrix}
\Sigma_{xx} & \Sigma_{xu} \\
\Sigma_{ux} & \Sigma_{uu}
\end{bmatrix}
\]
The Dual SDP:

• The dual optimization problem is:

\[
\text{minimize} \quad \text{Trace} \left( \Sigma \cdot \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \right) \\
\text{subject to} \quad \Sigma_{xx} = (A \ B)\Sigma(A \ B)^\top + \sigma^2 I, \quad \Sigma \succeq 0
\]

where the optimization variable is \( \Sigma \), a \((d + k) \times (d + k)\) matrix, with the block structure:

\[
\Sigma = \begin{bmatrix}
\Sigma_{xx} & \Sigma_{xu} \\
\Sigma_{ux} & \Sigma_{uu}
\end{bmatrix}
\]

• The interpretation of \( \Sigma \) is that it is the covariance matrix of the stationary distribution. This analogous to state-action visitation distributions (the dual variables in the MDP LP).
The Dual SDP:

• The dual optimization problem is:

\[
\begin{align*}
\text{minimize} & \quad \text{Trace} \left( \Sigma \cdot \begin{bmatrix} Q & 0 \\ 0 & R \end{bmatrix} \right) \\
\text{subject to} & \quad \Sigma_{xx} = (A \ B)\Sigma(A \ B)\top + \sigma^2 I, \quad \Sigma \succeq 0
\end{align*}
\]

where the optimization variable is \( \Sigma \), a \((d + k) \times (d + k)\) matrix, with the block structure:

\[
\Sigma = \begin{bmatrix}
\Sigma_{xx} & \Sigma_{xu} \\
\Sigma_{ux} & \Sigma_{uu}
\end{bmatrix}
\]

• The interpretation of \( \Sigma \) is that it is the covariance matrix of the stationary distribution. This analogous to state-action visitation distributions (the dual variables in the MDP LP).

• This SDP has a unique solution, say \( \Sigma^* \). The optimal gain matrix is then given by:

\[
K^* = -\Sigma^*_{ux}(\Sigma^*_{xx})^{-1}
\]